Mapping the ecogeographic distribution of biodiversity

15

L. Guarino[1]
Worked example: G. Maggs[2] and L. Guarino
[1]IPGRI, c/o International Laboratory for Research on Animal Diseases, PO Box 30709, Nairobi, Kenya: [2]National Botanical Research Institute, Private Bag 13184, Windhoek, Namibia.

Diversity is not evenly distributed within the geographic range of plant gene pools. This has been recognized since Vavilov described centres of origin and of diversity for a number of crops and has been amply demonstrated in numerous recent studies of variation in both crops and wild species in morphological, physiological, biochemical and molecular traits, though patterns at these different levels do not always coincide (Chapter 6). Some areas are richer in taxa within a given gene pool than others, a given species may be more genetically diverse in some areas than in others, ecotypes and clines have developed within species, variation in a species may be concentrated within or among populations, some alleles and characters are found in some populations and not others. Some of this heterogeneity in the distribution of diversity is explained by selection (natural and artificial) and gene flow and reflects ecogeographic adaptation and the preferences of farmers. Some is due to chance factors such as founder effect and genetic drift. Not only is diversity not uniformly distributed in ecogeographic space, it is also not uniformly distributed taxonomically, for example among the primary, secondary and tertiary gene pools. These contain widespread as well as narrowly endemic, and variable as well as relatively homogeneous, taxa.

It is up to the collector to use knowledge of the actual or presumed distribution of species and of genetic diversity within gene pools, and of how effectively this has been covered by previous collecting, to plan expeditions which will be efficient in the amount of new and useful diversity that they obtain. In particular, to plan a cost-effective mission, the collector must identify areas:

- where particular material of interest is known or suspected to occur (e.g. taxa missing from germplasm collections, material with particular adaptations or characters);

- which have high species or genetic diversity;
- which are significantly different from each other agriculturally, ecogeographically, floristically (i.e. in terms of representation of different taxa) or genetically;
- which have been undercollected;
- which are under threat of genetic erosion.

Other chapters are concerned with obtaining and organizing ecogeographic data. This chapter deals with some map-based techniques that may be used to explore the available data to identify the kinds of areas listed above. Chapter 5 discusses how such an analysis can be used to modify a basic collecting strategy, in particular how much extra weighting should be given to different areas and what the dangers of such bias are likely to be. Analysis of ecogeographic data will clearly be much simpler if a computer is used; some analyses, indeed, would be almost impossible without one. However, an attempt is made here to document manual, relatively low-technology solutions whenever possible. A geographic information system (GIS) is by no means essential to carry out many of the analyses described. The use of GIS technology is described in Chapter 16.

Preliminary data handling

Relevant data on germplasm accessions, herbarium specimens and other records of plant populations will include passport, characterization and evaluation descriptors. Passport data include both population descriptors and collecting site descriptors. There will also be data from sites located in the collecting region but where no collecting took place, e.g. climatic data from meteorological stations. Preliminary manipulation and analysis of some of these data prior to mapping may be useful. Each 'object' (plant population, collecting site or other kind of site) will often have data for several descriptors associated with it. Passport data will include information on a number of different aspects of the plant population and collecting site ecogeography, including climate, soil and vegetation. Germplasm accessions may have been characterized for various morphological descriptors or genetic loci. It may be possible to measure a number of morphological characters on herbarium specimens. Meteorological stations will have data on several temperature and rainfall parameters. So-called multivariate statistical techniques will therefore be particularly relevant for data analysis. These aim to simplify such multidimensional data in various ways.

Objects may be classified using the multivariate technique of *hierarchical clustering*. This will define groups of objects, called clusters, such that the overall similarity between two objects within the same cluster (taking all descriptors into account simultaneously) will be greater than that between two objects in different clusters. Clustering may be agglomerative (in which objects are joined into progressively larger clusters) or divisive (in which the group of all the objects is progressively divided). In general, the agglomerative method will bring out the best groupings and the divisive method will highlight discontinuities. The former has been most commonly used in numerical taxonomy, the latter in phytosociology and will not be discussed further here (but see below for references).

The first step in agglomerative clustering is measuring the similarity or dissimilarity (distance) between two objects. The method depends on the type of data (Dunn and Everitt, 1982):

- If all descriptors are recorded on a nominal scale (i.e. with discrete states the order of which is arbitrary, as may be the case with leaf shape, for example), the simple matching coefficient is usually used. This is defined as the number of characters that match, divided by the total number of characters.
- In the case of binary descriptors, negative matches may not be considered meaningful, and the Jaccard coefficient may be more useful. This is defined as the number of positive matches, divided by the number of mismatches.
- If all descriptors are continuous (or discrete and multistate, but with the order of states meaningful), Euclidean distance is commonly used as a measure of dissimilarity. Based on Pythagoras' theorem, it is calculated by adding up the squares of the differences between the two objects for each descriptor, and taking the square root of the sum.
- If some descriptors are nominal and some continuous, Gower's general similarity coefficient may be used (Gower, 1971).

There are other similarity and distance measures (Clifford and Stephenson, 1975; Dunn and Everitt, 1982), and software packages that perform multivariate analysis often allow a wide choice. Nei (1975; 1987) gives definitions of genetic identity (*I*) and distance (*D*) among populations based on allele frequency data, as might be produced by biochemical or molecular studies (see also Hoelzel and Dover, 1991; Hoelzel and Bancroft, 1992).

Having calculated similarities among all pairs of objects, the most similar objects are united into a cluster and similarities recalculated between this cluster, taken as a new object, and all other objects. The process is then repeated until all objects are in one cluster. To produce geographically compact clusters, a so-called contiguity constraint can be imposed on the clustering, such that two objects which would have been grouped into a cluster on the basis of their overall similarity are only actually allowed to do so if they are also geographically closer than a predetermined distance (Charmet, 1993). There are several different ways of choosing from which point within two clusters to measure the similarity between them (the sorting strategy), for example nearest neighbour, furthest link and centroid. Different combinations of similarity measure and sorting strategy will give slightly different results and the final choice among them is usually subjective. Pankhurst (1991) discusses different ways of objectively comparing and/or combining different clustering methods. Again, clustering programmes normally allow a choice of methods.

A second type of multivariate statistical technique that is often useful in the analysis of ecogeographic data is *ordination* or *multidimensional scaling*, of which the best-known example is principal components analysis (PCA). It will not explicitly produce groupings of objects, as clustering does, but will bring out the overall similarity among them by defining new variables, independent among each other, which are linear combinations of the original variables, but which differ in that the correlations among them, and therefore the redundancies, are taken into account. In this way, a long list of descriptors may be reduced to just a few new 'synoptic' variables (the first few so-called principal component axes), which despite being fewer nevertheless conserve most of the variability present in the original descriptors. The analysis will reveal how much of the original overall variability is represented by each principal component; the first will account for the largest single amount, the second for the second largest, and so on.

PCA is suitable only for quantitative data (but see Hill and Smith, 1976), as it is based on a matrix of the correlations among variables. However, a similarity matrix of the same kind as is used in clustering, which may be calculated whether descriptors are qualitative or quantitative, can also be used to plot the position of the objects in two- or three-dimensional space, in this case by using principal coordinates analysis (PCO). A minimum spanning tree of the similarity matrix superimposed on such a plot will reveal how well the original similarities are preserved (Dunn and Everitt, 1982).

Another widely used ordination technique is canonical variate analysis (CVA). This is similar to PCA but differs in that objects must be assigned to groups a priori and the analysis produces new axes (canonical variates, the equivalent of principal components in PCA) such that maximum discrimination is obtained not among objects, as in PCA, but among groups when tested against the variation within groups. Each a priori group could be a different taxon, a set of landrace accessions from different countries or a set of accessions of a wild species collected on different soil types, for example. Each group could also be a population, with observations on a number of different individuals per population being recorded.

Other relevant ordination techniques are non-metric multidimensional scaling (Kruskal, 1964; Clifford and Stephenson, 1975) and detrended correspondence analysis (Hill and Gauch, 1980; Gauch, 1982). The latter has been much used in community ecology but also has application in ecogeography (Myklestad and Birks, 1993) and, indeed, numerical taxonomy.

A brief introduction to multivariate statistical techniques as they have been used to describe and analyse plant variation is provided by Stuessy (1990). A more detailed but still relatively brief account is provided by Dunn and Everitt (1982). Pankhurst (1991) is also useful. Sneath and Sokal (1973; updated by Sokal, 1986), Blackith and Reyment (1971) and Clifford and Stephenson (1975) are standard works. The manuals of software packages are also good sources of information on the relative merits, capabilities and competencies of different methods. There are programs available which carry out just one kind of analysis, for example DECORANA (Hill, 1979a), TWINSPAN (Hill, 1979b) and CANOCO (Braak, 1988). Other program packages carry out a variety of multivariate analyses; examples include CLUSTAN IV, SYN-TAX III and NTSYS (Pankhurst, 1991). More general statistical packages with multivariate options include GENSTAT, SAS, UNISTAT and STATGRAPHICS. It is not possible to carry out some multivariate statistical procedures if there are gaps in the data set. Though some packages do provide routines for estimating missing data values, this can be a major problem in the use of multivariate techniques.

Data on the variation in morphological characters within populations, already mentioned in connection with CVA, are not often available in herbarium-based studies. Similarly, variation within germplasm accessions is only occasionally measured for morphological characterization descriptors (frequently the character is simply noted as being variable), though more often for evaluation descriptors (but see Hintum, 1993). However, as mentioned earlier, various biochemical and molecular techniques can be used to calculate genetic similarities among populations. Different measures of genetic diversity can also be calculated from such data both for populations and for groups of populations (Nei, 1975, 1987; Brown and Weir, 1983; Hoelzel and Dover, 1991; Hoelzel and Bancroft, 1992). Some examples are given in Box 15.1. One-locus gene diversity (h) may be averaged over all the loci studied to obtain a measure of mean gene diversity in the population (H_T). This is perhaps the most widely used genetic diversity parameter.

Mapping ecogeographic data

Point maps

Having carried out any appropriate preliminary analysis of the variation among plant populations and/or collecting sites, the next step will usually be the plotting of distribution data and the visual interpretation of the resulting maps. Chapter 14 discusses the different kinds of distribution maps that can be prepared and suggests that point maps showing exact localities are usually the most appropriate in the context of planning germplasm collecting missions. Though they are certainly the simplest starting-point, and thus the most logical, they are capable of considerable elaboration, enabling the display of more information

Box 15.1
Some parameters used to measure genetic diversity

Percentage of polymorphic loci: P

Mean number of alleles per locus: n_e

Within-population gene diversity at a locus: $h = 1 - \Sigma p_i^2$
where p_i is the frequency of the ith allele or haplotype

than just geographic location. This additional information on each plant population could include:

1. the type(s) of plant record(s) associated with it;
2. collecting site and population data;
3. the taxonomic category to which the population belongs;
4. the state or value of characterization or evaluation descriptors, taken one at the time;
5. membership of classes defined by combinations of characterization and evaluation descriptors;
6. the value of 'synoptic' characterization and evaluation descriptors;
7. within-population variation in characterization and evaluation descriptors.

Type of plant record

A first and basic elaboration of simple point distribution maps is to show the location of germplasm samples and of herbarium specimens (or literature records) in different ways, usually using two different symbols. It is most effective to represent the former by a filled symbol and the latter by an empty one, perhaps plotting them on separate transparencies, which can then be overlaid on the base map. Areas where the target taxon is known to occur but from which no germplasm is available often stand out very clearly in such an exercise.

Collecting site and population data

Plant records may also be distinguished according to various characteristics of the site and population to which they refer, i.e. passport data. Site data such as soil type or altitude, for example, could be shown by using symbols of different shapes, colours, shades or sizes (Fig. 15.1a, b). Such parameters as the size of the population and the threat of genetic erosion it faces could be shown in similar ways. In some cases, it may be useful to distinguish plant records on the basis of date, for example to show changes in distribution over years or differences in flowering time in different areas.

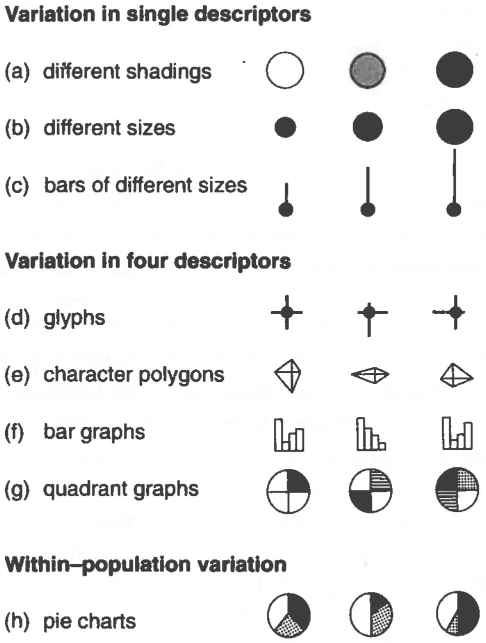

Fig. 15.1. Some symbols which may be used on dot maps to show variation in populations. See text for details.

Membership of different taxonomic categories

Different 'taxa', taken in a wide sense to include not just formal categories like species in a genus but also cultivar groups or landraces in a crop (Chapter 7), may also be displayed on the same map by using different symbols. Such a technique is useful in showing areas of high taxonomic diversity and areas where rare or endemic taxa may be found. Symbols must be chosen which can be superimposed and still remain identifiable, so that co-occurrence can be shown, and which can be filled to represent germplasm accessions or left unfilled to represent other kinds of records. If different symbols are being used to show site or population data, different taxonomic categories could be shown by lines emerging from the symbol at different angles (glyphs, see below).

Characterization or evaluation descriptors taken one at the time

Data on physiological, morphological or genetic descriptors can be added to a point map to investigate the spatial distribution of diversity. If a single descriptor measured on a discrete scale is being considered, its different states may be shown by different symbols or by the same symbol filled in with different shadings, patterns or colours (Fig. 15.1). The symbol may reflect the descriptor very closely. Variation in leaf shape among populations, for example, may be shown on a map by reproducing small representations of the actual leaf shape at each site of observation, and flower colour by differently coloured dots. If the descriptor is measured on a continuous scale, variation may be shown by the same symbol at different sizes or with different-sized bars emerging from it (Fig. 1b, c).

Membership of classes defined by combinations of descriptors

Combinations of plant descriptors may be displayed by lines (glyphs) emerging at different angles to the central dot (or different symbols, if site or population data are also being displayed), so that the length of a line at 45° to the vertical might represent mean petiole length in the population and that of a line at 90° mean leaf length, for example, and so on (up to a maximum of eight characters in this case). These are called radial bar graphs or rose diagrams (Richardson, 1985). A recent example of their use is Francisco-Ortega *et al.* (1993). An example involving four characters is given in Fig. 15.1d. The ends of adjacent glyphs may be joined up and the enclosed 'character polygon' filled in for ease of comparison (Fig. 15.1e). An example of their use is given by Jones (1972). Variation in a number of characters may also be shown in side-by-side bar graphs, as in Fig. 15.1f and Morley (1971).

Another method that has been adopted to display the distribution of more than one character on maps is to use a suitably large circle to mark localities and divide this up into two or more equal slices, each representing a descriptor (Fig. 15.1g). The descriptor state can be shown by filling in the appropriate quadrant or leaving it unfilled (in the case of binary descriptors), or by the use of different shadings, patterns or colours in the case of multistate descriptors (e.g. Seyani, 1988). Both these methods can also be used to show the co-occurrence of taxa. A plant population's membership of different clusters may be shown on point maps, again by using different symbols, as if it were a descriptor with discrete states.

'Synoptic' descriptors

The score of each plant population along each of the first few principal components (usually the first two to four) can be mapped as any other continuous descriptor, though each principal component in fact represents a combination of several of the original descriptors, each differently weighted (the exact weights will be specified in the analysis output). Plant populations with similar principal component scores will tend to be similar for the original descriptors contributing most to that principal component.

Within-population variation

The pattern of distribution of within-population variation for continuous variables may be displayed by mapping a measure of dispersion such as standard deviation in the same way as any other continuous variable, for example with differently sized dots. One could thus have a point map of mean leaf length of populations, say, along with a point map of the standard deviation of leaf length within each population. Values of the genetic parameters in Box 15.1 may also be mapped in the same way as any other continuous descriptor. For descriptors with discrete states, including such biochemical data as the occurrence of different alleles at a locus, pie charts giving the frequency of each state in the population may be used (Fig. 15.1h).

Contour maps

The distribution of classes of plant populations (be the classes defined taxonomically, by a single character, by the joint occurrence of characters, by ordination or by clustering) can be related to environmental factors by searching the map visually for associations. For example, mean leaf length in a species, as shown by the length of a particular glyph, may seem to be associated with higher altitudes or higher rainfall, as shown by the size of the central dot. Contour, or isoline, maps may facilitate the identification of such patterns.

To produce contour maps, a line is drawn joining all the points where the variable of interest takes a particular value and the process is repeated for a set of (usually) equally spaced values (Richardson, 1985; Burrough, 1986). When the variable is number of taxa, the contour lines are sometimes called isoflor lines (Chapter 14). One could also draw lines of equal genetic diversity; a suitable name might be isodive lines. Lines linking populations with similar values for morphological or other kinds of plant characters are called isophenes. For climate data, lines of equal temperature are called isotherms, lines of equal rainfall isohyets. The particular advantage of contour maps 'lies in the fact that they may be easily correlated with environmental and/or geological variables, and intercorrelations between these surfaces (whether they be morphological characters, chemical characters, abstract factors, etc.) are readily visualized by the reader' (Adams, 1970). What contour maps do is render a continuous variable discrete. Associations between discrete variables are often much easier to notice than relationships between continuous variables.

It is not impossible to draw contour lines by eye. Indeed, up to relatively recently that was the only way to draw them. The development of computer mapping has meant that the subjectivity that was the main drawback of such an approach has been overcome (Adams, 1970). Whether done by hand or by computer, the first step in the generation of contours is the formation of a regular mesh of points for which values must be estimated on the basis of interpolation from the points where measurements were actually made. Mesh point values are then ranked consecutively so that appropriate isoline values can be chosen. Finally, the isoline values are located by interpolation between mesh point values and the resulting jagged lines smoothed (Richardson, 1985). The areas between contour lines are often filled in using shadings or patterns of increasing intensity. There are different methods of interpolation (e.g. Adams, 1970; Burrough, 1986), and software mapping packages will specify which one is being used and may offer a choice.

Using different base maps

If point distribution maps (however elaborate) or contour maps are drawn on a clear medium such as very thin paper or acetate transparencies, they can be used in conjunction with different base maps, for example not just topographic maps, but also ones showing agroclimatic zones, soils and vegetation types. This allows one to look for associations of distribution, characterization and evaluation data with different environmental features without having to plot distributions anew each time. This overlaying facility is of course one of the most useful features of a GIS. The trouble with the low-technology solution is that maps come in a variety of scales and projections. A photocopier with a reducing facility is very useful in overcoming the former problem. A pantograph or an overhead projector may also be used to change the scale of maps. The problem of projections is much more complicated and the only really viable methods are computer-based. Most GIS packages have the capability of making maps comparable in scale and projection.

Further analysis

Visual analysis of point and contour maps can be taken quite far. The human brain is reasonably good at distinguishing patterns, certainly better than most computers. However, various analyses can be carried out on mapped ecogeographic data to test any hypotheses on the factors controlling the pattern of distribution of diversity generated by visual inspection. Some of these have already been alluded to.

The association between a continuous characterization and/or evaluation descriptor (which could be scored along a principal component) on the one side (the dependent variable) and an environmental site descriptor (e.g. altitude) on the other (the independent variable) may be tested using regression analysis. Multiple regression may be used when it is wanted to explain variation in a dependent variable with reference to more than one independent variable. The ordination techniques of canonical correlation analysis and canonical correspondence analysis (Braak, 1986) may be used to explore the structure of a data set with multiple dependent variables as well as multiple independent variables. When the data are classified (as in contour maps), association may be tested using χ^2 or other non-parametric methods.

Mantel's test may be used to compare similarity matrices (Sokal, 1979; Heywood, 1991). Thus, a classification of plant population localities based on collecting site data may be compared with one of the populations themselves based on characterization and/or evaluation data. Autocorrelation statistics such as Moran's I may be used to quantify the morphological or genetic similarity between pairs of plant populations (or groups of plant populations) as a function of the geographic distance between them (Sokal and Oden, 1978; Heywood, 1991).

To combine more than one characterization or evaluation descriptor into a single contour map, different approaches are possible. One is to map contours of principal component scores. Another is to use *differential systematics*. Proposed by Womble (1951) and also described by Adams (1970) and Kirkpatrick (1974), this allows the simultaneous consideration of multiple contour maps. What is involved is the calculation of a slope (differential) perpendicular to the contour at each mesh point for each descriptor. This is a facility available in many GIS packages. The absolute values of the slopes for all descriptors are then summed and divided by the number of descriptors. Contours can then be drawn of this so-called 'systematic function'. The ridges on the contour map reveal areas where maximum change over distance is occurring and thus separate plant populations which are markedly different in the expression of the characters included in the analysis. Gentle clines do not show up as forcefully as more abrupt patterns, but areas where the systematic function is high are the ones where most variation is concentrated in the least space. In a similar way, genetic or morphological distances may be divided by geographic distances and the resulting matrix used to detect 'genetic boundaries' where several gene frequencies or characters vary abruptly (Monmonier, 1973; Pigliucci and Barbujani, 1991).

Areographic methods

How is one to pick out on a map the kinds of potentially high-priority areas listed at the beginning of the chapter? Some ways of dealing with the problem have already been mentioned, for example the use of isoflor maps and differential systematics for areas of high diversity. Nabhan (1990) provides a useful introduction to another type of approach, one based on so-called areographic methods.

Defining subregions

The first and fundamental step in areographic analysis is the definition of subregions within the general target or study region. It is by making comparisons among these with regard to parameters such as diversity and extent of previous collecting that it becomes possible to set priorities within a target region. Subregions could be demarcated by administrative boundaries, and numerous examples exist of the analysis of the distribution of variation in a crop among countries, for example, usually using CVA (e.g. Spagnoletti Zeuli and Qualset, 1987; Erskine *et al.*, 1989; Morden *et al.*, 1989; Porceddu and Damania, 1992). For full comparability, however, subregions should be 'equivalent', i.e. similar in size and shape (Nabhan, 1990). In practice, this often means some sort of grid array of squares, for example defined by lines of latitude and longitude. The number and size of subregions defined within the target area will depend crucially on the density of plant records and on the heterogeneity of the environment, if this is known or can be estimated. Subregions should be large enough for the number of plant records in each to allow meaningful comparisons of diversity among subregions, but not so large – and hence their number so small – that resolution suffers.

Choosing subregions for sampling: preliminary surveys

In an initial survey, when little is known about either target region and taxon, a grid can simply be superimposed on a map of the target region, and subregions (i.e. grid squares) chosen to be visited according to various methods, subject of course to accessibility and other logistical constraints. Subregions could be chosen at *random*, by assigning each a number and picking a proportion out of a hat or by flipping a coin or throwing a die for each. In *systematic* sampling, in contrast, subregions would be visited in a regular pattern. This would be equivalent to collecting every so many kilometres along a chosen route. Since there is often strong correlation between neighbouring measurements in the natural environment, systematic sampling may be a better method to capture variation than random sampling, though care must be taken that the sampling pattern does not coincide with pattern in the environment. Also, a regular grid of measurements carries more spatial information than an equal number of random points, which makes the drawing of contours easier, as already noted. Certain kinds of statistical analysis, however, assume that the data have been collected from a random arrangement of sampling units.

When more is known of the environment in the target region or of the target taxon, a different strategy may be more appropriate. Thus, subregions may be picked out at random (or systematically) separately and independently in areas of the target region that are known a priori to be different in some relevant feature. They could have different kinds of soils or show marked differences in annual rainfall or be inhabited by different subspecies. An agroclimatic classification may also be used to define homogeneous areas, either one of the standard ones (Chapter 9) or a more specific one, for example the cluster analysis or PCA results of data from meteorological stations in the target region. In Chapter 5, the suggestion is made that the weighting that each environmental subunit receives in the sampling should be in proportion to the commonness of the target taxon. Such a *stratified random* approach reduces the possibility of missing out an obviously important aspect of variation purely by chance and allows more flexible allocation of resources. *Gradset* or gradient-directed sampling is another example, involving the sampling of grid squares (chosen randomly or systematically) on replicate transects along some perceived environmental gradient or cline. The disadvantage is that any variation perpendicular to the gradient may be undersampled, depending on the relative allocation of resources between and within parallel transects.

Austin and Adomeit (1991) have carried out a cost–benefit analysis of different vegetation sampling methods by simulation. The cost was measured in terms of time and effort in the field, the benefit in terms of the number of species recorded. It could equally well have been genotypes within a species. Their conclusions are worth quoting:

- Simple statistical sampling designs (i.e. methods like random and systematic sampling) are unlikely to be cost-effective at detecting the range of community variation in a regional landscape.
- Environmentally stratified surveys can be much more effective, particularly if the stratification is based on the most important variables for species distribution. Maximal use of existing information for stratification will reduce costs and improve effectiveness.
- Logistical constraints indicate that gradset sampling stratified by topographic units is the most practical and effective strategy.

In crop collecting, the stratification may involve a combination of environmental and cultural or socioeconomic parameters. Thus, if an area of broadly homogeneous mean annual rainfall, or falling in a single agroclimatic zone, is inhabited by two or more distinct ethnic groups, for example, this should be taken into account to define areas that are not only ecogeographically homogeneous, but also relatively homogeneous in terms of their human population.

These considerations are applied here in choosing where to have collecting sites within a region, but they are no less relevant within collecting sites. Chapter 5 discusses sampling of individuals within a population, and also recommends a stratified random approach at this level in cases where a subpopulation structure exists.

Choosing subregions for sampling: more advanced surveys

If sufficient information is available on a study region, it is useful to build up a matrix giving the number of plant records belonging to each class in each subregion. As already pointed out, the term class could mean here a formal taxonomic unit such as species, but could also refer to categories derived from other kinds of classification: groups of plant populations distinguished by having particular states for a given descriptor, groups of plant populations distinguished by the joint occurrence of particular states of more than one descriptor, clusters derived from multivariate characterization and/or evaluation data or, indeed, traditionally recognized landraces. (It must be remembered, however, that the analysis for classes within the target gene pool defined in a particular way will not necessarily be valid for other classes.) If coverage is patchy or generally low, rather than the actual number of plant records, simple presence/absence of the class in the subregion could be recorded. A third possibility is to divide up each subregion, again in a regular grid, and record the number of sub-subregions occupied by the class within each subregion. Separate matrices could be built up for each level in the taxonomic hierarchy. For example, diversity within a family could be investigated through a matrix of the number of records for each species in each subregion together with further matrices giving the number of species or records in each genus and subregion.

Picking plant records at random one after the other from within a subregion and plotting cumulative number of classes against cumulative number of plant records can give an idea of whether a maximum has been reached, that is, whether collecting a further population is likely to result in an additional class being represented in the sample (Peeters, 1988). The total number of classes found in a subregion can be regressed against the total number of plant records recorded from that subregion to decide which subregions are relatively undercollected. According to Nabhan (1990), in any subregion falling above the upper 95% confidence limit of the regression line of classes on plant records, the number of classes recorded is higher than would be expected on the basis of the total number of collections made in the past, that is, the thoroughness of sampling. Further collecting in such a subregion is therefore more likely to turn up classes not previously collected there than further collecting in subregions falling within the 95% confidence limits or, even more so, falling below the lower confidence limit.

Rare, widespread variants can also be useful in assessing the completeness of collecting in a given area. Thus, for example, if a rare allele (occurring in less that 5% of accessions overall, say) occurs in only one accession in one subregion but in several accessions in others, the former may be relatively undercollected (Weeden *et al.*, 1988).

In addition to undercollectedness, the data in the class vs. subregion matrices can be used to calculate the class diversity of subregions. This may be measured in a number of different ways that bear different and not always obvious interpretations. Magurran (1988) provides a comprehensive survey of the calculation, interpretation and relative merits of several diversity indices. The definitions of some of the more widely used ones are given in Box 15.2.

Number of classes (richness) is the simplest one to use and understand. Another simple richness measure is Margalef's diversity index D_{Mg}. As has been mentioned, areas with equal numbers of taxa may be shown using isoflor maps. This is a useful method for picking out areas of high species richness, but three points should be kept in mind.

First, two areas may have equal numbers of species, but the species may be closely related in one case and taxonomically widely separated in the other. Isoflor maps of both species and sections within the genus might need to be compared to get an accurate impression of the distribution of diversity. Some ways of taking into account the similarities among classes in calculating class diversity are reviewed by Cousins (1991) and Krajewski (1994). Vane-Wright *et al.* (1991) and Williams *et al.* (1991, 1993) have produced a computer program (WORLDMAP) which can be used to identify areas of high biodiversity defined according to four different weightings of simple species richness derived from different measures of taxonomic relatedness. High species richness does not necessarily coincide with high genetic diversity within individual species and different measures of genetic diversity (based on morphological, biochemical or molecular characters) may not give corresponding

Box 15.2
Definitions of class diversity indices

Margalef diversity index: $D_{\mathrm{Mg}} = (S - 1)/\ln N$

Shannon–Weaver diversity index: $H = -\Sigma p_i(\ln p_i)$

Simpson's index: $D = \sum \left(\frac{n_i(n_i - 1)}{N(N - 1)}\right)$

Log series diversity α: $S = \alpha \ln(1 + N/\alpha)$

where p_i is the proportion of objects in the ith class, n_i is the number of objects in the ith class, N is the total number of objects and S is the total number of classes

results (Chapter 6). It is important to remember what level of diversity is being mapped and analysed. The results may not be applicable at other levels.

Secondly, areas with equal numbers of species (or, indeed, landraces) are not necessarily floristically similar, i.e. the species involved may be completely different. This is discussed further below.

Thirdly, the number of species present in an area says nothing of the relative abundance of each. The distribution of plant records within classes, or of classes within higher-level classes, is an important aspect of diversity. Ten plant records may be distributed evenly among five species, or one species may be represented by six plant records and the other four by one each. Intuitively, the former case should score the higher diversity.

The two most commonly used measures of diversity which take into account the proportional abundances of different classes are the Shannon–Weaver diversity index H and Simpson's index D. The latter gives relatively more weight to common species; otherwise there is little to choose between them on theoretical grounds (Goldsmith *et al.*, 1986). The gene diversity parameter calculated from allele frequency data (Box 15.1) is the genetic equivalent of Simpson's index; the Shannon–Weaver index can also be applied to such data (Brown and Weir, 1983). Hutchenson (1970) has provided a test for comparing diversity scores computed using the Shannon–Weaver method.

These diversity measures are also called heterogeneity indices, as they take both evenness of distribution and richness into account, and non-parametric indices, to distinguish them from measures which assume an underlying abundance distribution (Magurran, 1988). The most widely used of these is α, which assumes that the number of classes of different sizes will form a log series, though it can still be used if such a distribution is not in fact the best description of the underlying class abundance pattern. Southwood (1978) has suggested that α is the best candidate for a universal diversity statistic in ecology. It and Simpson's index are the least sensitive to sample size (Magurran, 1988).

Diversity within a subregion can be measured for individual continuous descriptors as the standard deviation of the individual plant population values in a subregion, which may then be compared using statistical tests of the homogeneity of variances. Overall diversity for several descriptors together may be measured as:

- the standard deviation of the principal component scores of the plant populations in a subregion;
- the mean of the standard deviations for individual continuous characters;
- the mean of the diversity index scores for individual categorical characters.

Biochemical or molecular data can be used to calculate genetic diversity parameters for subregions, factoring out within-population variation; thus, total gene diversity among populations (H_T) within a subregion may be partitioned as within-population diversity (H_S) and among-populations diversity (G_{ST}) (Nei, 1975, 1987; see also Hoelzel and Dover, 1991; Hoelzel and Bancroft, 1992).

It is useful to produce maps in which subregions are marked in different ways according to their environmental or biological diversity. Maps that show the distribution of qualitative or quantitative information in a geographic region by segregating that information as it is found within subregions are called choropleth maps (Richardson, 1985; Burrough, 1986). Grid squares can be shaded or coloured in different ways or filled circles of different sizes can be placed at their centres. Chapter 5 discusses on what basis collecting effort should be divided up among subregions of different diversities.

Having identified subregions that are undercollected or particularly diverse (on whatever definition), the next step is to investigate the similarities among subregions. Clearly, it may only be necessary to visit both of two equally diverse subregions if different classes are contributing to the high diversity scores of the two areas. The WORLDMAP software has a facility for the exclusion of particular areas from the analysis. If the area with highest diversity is so excluded, and the diversity measurement recalculated for the remaining areas, and so on, a set of areas can be chosen which will optimally cover overall diversity (see also Rebello and Sigfried, 1992). Another useful approach is calculating a similarity matrix for subregions and applying a clustering method. Membership of different clusters of subregions can be shown by using different shadings, patterns or colours.

Similarities among subregions may be calculated not only on the basis of characterization and/or evaluation data but also for various environmental factors. If the presence/absence (or extent) of particular soils or vegetation types in each subregion is recorded, for example, clustering can be carried out to reveal subregions that are environmentally similar. Such data could come from passport data or from soil and vegetation maps, meteorological stations, etc. In the absence of data on the similarity among subregions in terms of the classes of plant populations found there, it will be most efficient to target for priority collecting subregions that are maximally dissimilar on environmental grounds. Indeed, as already mentioned, these clusters are a good starting-point for the definition of the distinct but internally homogeneous areas within which stratified sampling could take place.

A final element that will have to be considered in deciding where to collect – besides completeness of collections, diversity within subregions and similarity among subregions – is risk of genetic erosion in a subregion. Population characteristics such as size and age structure will be important in determining this, but also a variety of external factors which it may be possible to quantify for a particular area. One way of doing this is the genetic erosion prediction system of Goodrich (1987) (Chapter 4).

How can one bring all these data together to compile a list of subregions ranked in order of priority for collecting? An approach that may be useful is to prepare a table listing subregions and record in separate columns the score of each subregion for several variables: number of classes, number of endemic classes, diversity (however measured), undercollectedness and presence/absence of classes missing from germplasm collections or otherwise of particular significance. The scores for each of these variables should be transformed to a discrete scale (e.g. 0–4) or simply given a rank. The system is flexible in that, if particular variables are thought to be more or less important than others, they can be given more or less weight by appropriate changes in the scale. See Chapter 5 and Nabhan (1990) for an assessment of the benefits and dangers of prioritizing on these different criteria. Adding up the transformed variable scores for each subregion would give an idea of rank in overall collecting priority.

Worked example

The Cucurbitaceae are a predominantly tropical family of about 118 extant genera and 825 species, the overwhelming majority of which are distributed in three main areas: Africa and Madagascar, Central and South America, and Southeast Asia and Malesia (Bates *et al.*, 1990). There are nine cultivated cucurbit species of major importance in four genera and six minor crops (Whitaker and Bemis, 1976). In addition, some wild species are exploited as food, fodder or water sources by local peoples, and there is interest in their possible domestication. An example of the latter is *Acanthosicyos horridus*, a shrub of the Namib Desert of the southwestern African coast. Southern Africa is also important as the native home of wild forms of the watermelon, *Citrullus lanatus*, which are extremely diverse in the Kalahari region. Both these areas are represented in Namibia, which makes this country an important one for cucurbit diversity. Some 34 cucurbit species in 11 genera are so far recorded from Namibia, which has been fairly well collected from the point of view of herbarium material but where germplasm collecting of cucurbits has only very recently started in earnest.

This section presents an analysis of the distribution of cucurbit diversity at the species level in Namibia which could be used as the basis for the planning of germplasm collecting in the country. It is based on the areographic principles outlined in the preceding parts of this chapter and in particular by Nabhan (1990).

Materials and methods

Definition of subregions

The study region, i.e. the territory of Namibia, was divided into a number of arbitrarily defined subregions of equal size. In this case, 17 subregions of 2° latitude by 2° longitude were defined, plus two further subregions of approximately the same size but irregular outline (Fig. 15.2). These subregions will be referred to as SR1–19.

Data collection

The raw data for the analysis were the locality information recorded on herbarium sheets held at the National Botanical Institute, Pretoria (PRE), and the National Herbarium, Windhoek (WIND). As collecting continues, the distribution patterns of some species will need to be amended, and therefore also some of the results. The herbarium data presented here were collated in 1991. All the $\frac{1}{4}$° squares where each species was collected were recorded. These data were worked up into a matrix of 34 species by 19 subregions (Matrix A), giving in each cell the number of $\frac{1}{4}$° squares within each subregion where each species was present (Table 15.1). A further matrix was derived from Matrix A of 11 genera by 19 subregions (Matrix B), giving in each cell the number of species within each genus recorded from each subregion.

Data analysis

Several analyses were carried out on the data in Matrices A and B.

1. From Matrix A, diversity values were calculated for each subregion using the Shannon–Weaver index (H_{spp}). From Matrix B were calculated the numbers of species and genera per subregion and a second measure of diversity, again using the Shannon–Weaver index (H_{gen}). Zeros were taken as 0.001, which is standard practice in these cases.
2. Number of species per subregion was plotted against number of

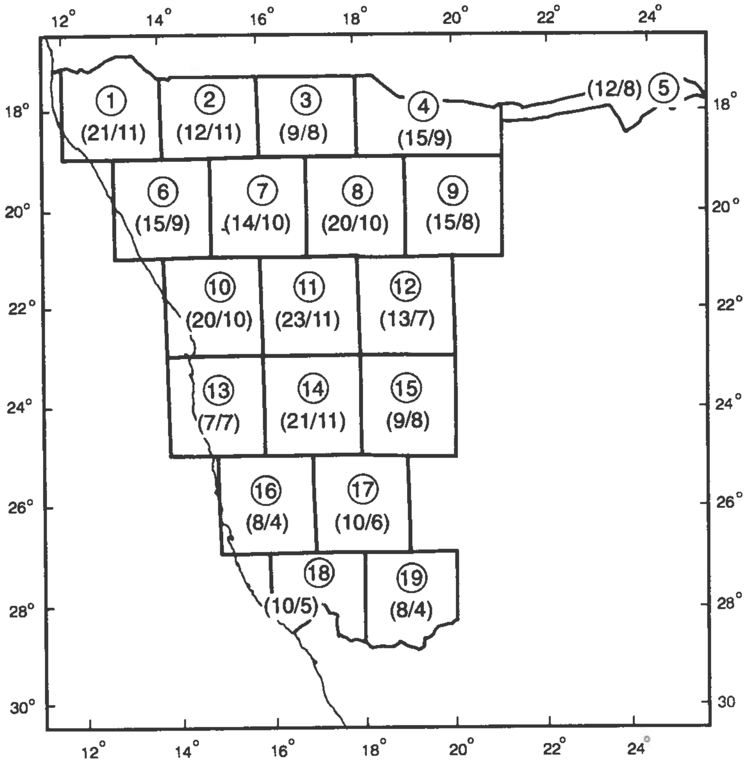

Fig. 15.2. Map of Namibia showing the 19 subregions used in the analysis (numbers enclosed by circles). In brackets are given the number of species/genera recorded from each subregion.

records per subregion (the column totals of Matrix B and Matrix A respectively). A regression line and 95% confidence limits were calculated.

3. Single linkage (nearest-neighbour) cluster analysis was used on a data matrix of the presence/absence of each species in each subregion.

Results

Distribution of diversity

Figure 15.2 gives numbers of species and genera for each subregion. The number of genera per subregion ranged from four (out of 11, or 36%) in SR16 and SR19 in the south to 11 (i.e. all Namibian genera) in SR1-2 in the northwest and SR11 and SR14 in the centre of the country. Number of species per subregion ranged from seven (out of 34, or 21%)

Table 15.1. Number of $\frac{1}{4}$° squares within each subregion (SR1–SR19) where each cucurbit species is present.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Tot.
Acanthosicyos horridus	2	–	–	–	–	–	–	–	–	5	1	–	5	1	–	3	–	1	–	18
Acanthosicyos naudinianus	–	2	3	3	3	–	2	3	3	–	5	1	–	–	2	–	1	–	–	28
Citrullus lanatus	1	2	–	1	–	1	2	1	2	5	7	1	1	1	3	1	1	2	1	33
Citrullus ecirrhosus	2	–	–	–	–	1	–	–	–	6	–	–	–	1	–	5	2	–	–	17
Coccinia sessilifolia	–	–	–	–	–	–	3	1	2	2	8	1	–	2	1	–	–	–	–	20
Coccinia rehmannii	3	1	–	1	–	1	2	2	2	4	6	3	1	6	2	–	1	–	–	35
Coccinia adoensis	–	–	–	1	3	–	–	1	2	–	–	–	–	–	–	–	–	–	–	7
Corallocarpus bainesii	–	–	1	1	3	–	2	–	4	–	3	1	–	1	–	–	–	–	–	16
Corallocarpus triangularis	–	–	–	–	–	–	–	–	–	–	–	1	–	–	1	–	–	–	–	2
Corallocarpus dissectus	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	1	2	–	3
Corallocarpus welwitschii	6	3	2	–	1	2	–	3	2	5	5	1	1	2	–	2	2	1	1	39
Corallocarpus schinzii	–	–	–	–	–	1	–	–	–	1	1	–	–	1	–	–	–	–	1	5
Cucumella cinerea	2	1	–	–	–	1	2	1	–	5	3	–	–	–	–	–	–	–	–	15
Cucumella aspera	6	–	–	–	–	3	–	–	–	4	–	–	1	2	–	–	–	–	–	16
Cucumis metuliferus	–	–	–	1	2	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
Cucumis mecusei	3	–	–	1	–	4	1	2	–	3	9	–	–	5	–	4	2	2	1	37

Cucumis sagittatus	6	–	–	–	–	6	1	–	–	4	3	–	1	8	–	3	3	3	8	46
Cucumis rigidus	–	–	–	–	–	–	–	–	–	1	–	–	–	–	–	2	–	4	2	9
Cucumis africanus	2	–	–	–	–	3	7	3	–	7	13	2	–	10	2	3	2	2	2	58
Cucumis kalahariensis	–	–	–	1	1	1	–	2	4	–	–	3	–	–	–	–	–	–	–	12
Cucumis humifructus	1	–	–	1	–	–	–	2	2	–	–	–	–	–	–	–	–	–	–	6
Cucumis anguria	1	2	1	3	–	–	2	2	2	5	6	–	–	1	–	–	–	1	–	26
Dactyliandra welwitschii	10	1	1	–	–	3	8	–	–	6	1	–	1	4	1	–	–	–	–	36
Kedrostis foetidissima	1	–	–	3	2	–	–	1	1	1	4	–	–	–	–	–	–	–	–	13
Kedrostis hirtella	1	1	–	1	1	1	–	4	2	–	2	–	–	–	–	–	–	–	–	13
Kedrostis africana	–	–	–	–	–	–	–	–	–	–	1	–	–	1	1	–	–	–	–	3
Kedrostis capensis	–	–	–	–	–	–	–	–	–	1	–	–	–	1	–	–	–	6	1	9
Momordica welwitschii	7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	7
Momordica boivinii	–	–	–	–	–	–	–	2	–	–	1	–	–	–	–	–	–	–	–	3
Momordica balsamina	1	3	1	2	6	–	–	1	–	–	5	5	–	3	1	–	1	–	–	29
Momordica humilis	4	1	1	–	1	3	7	1	–	7	3	1	–	2	–	–	–	–	–	31
Trochomeria macrocarpa	1	4	1	1	1	–	1	3	2	2	4	1	–	1	–	–	–	–	–	22
Trochomeria debilis	2	–	–	–	–	1	–	1	1	1	3	1	–	2	–	–	–	–	–	12
Zehneria marlothii	3	1	–	2	1	–	2	5	3	–	3	–	–	1	–	–	–	–	–	21

in SR13 to 23 (68%) in the nearby SR11. The strongest concentration of subregions with a large number of species recorded is in the centre of the country (SR8, SR10, SR11 and SR14). SR1 and SR8 had both ≥3 *Momordica* species and ≥5 *Cucumis* species. SR4 and SR10 also had ≥5 *Cucumis* species. The range of number of genera and number of species per subregion was divided into three sections with approximately equal numbers of subregions. A value of 1 was assigned to subregions at the bottom of the range and a value of 3 to subregions in the top category. Table 15.2 shows these data. Subregions with ≥5 *Cucumis* species are given a score of 1 in a separate column in the table, other regions scoring 0.

The ranges of H_{spp} and of H_{gen} values were each divided into four ranges in such a way that each range contained approximately equal numbers of subregions and values of 1 to 4 were assigned to subregions accordingly (Table 15.2). For H_{spp}, the subregions at the top of the range were SR1, SR8–11 and SR14. This corresponds to the central portion of the country and the northwest. In these subregions, a given number of records will represent more species than in other subregions. Subregions in the southern portion of the country generally have low H_{spp} values. An essentially similar pattern was evident for H_{gen}, with the highest values occurring in SR1–2 in the northwest and SR11 and SR14 in the centre and the lowest in the south. A higher value for this diversity index means that a given number of species will represent more genera.

Undercollected areas

Subregions that fall above the 95% confidence limits of the regression line of species per subregion against records per subregion may be considered to be relatively undercollected. SR4, SR6, SR8, SR9, SR12 and SR14 may be thus described, subregions mainly concentrated in the northeast and centre. They are given a score of 1 in a separate column of Table 15.2, all other regions scoring 0.

Subregional similarities

Clustering of the presence/absence data revealed the existence of two main floristic zones. One comprises SR2–5 and SR7–9 in the central and eastern parts of the north, the other includes SR13 and SR16–19 in the south. Membership of each subregion for one of six clusters (A–F) is shown in Fig. 15.3 and Table 15.2.

Distribution of important species

In addition to the distribution of diversity, the distribution of particular taxa is often of importance. An area of high diversity for the family as a whole might not contain the most interesting species, for example

Table 15.2. Various cucurbit diversity parameters for each subregion (SR1–SR19). See text for details.

SR	Number of genera	Number of species	Number of *Cucumis* species	H_{spp}	H_{gen}	'Collectedness'	Presence of *Citrullus lanatus*	Cluster membership	Total
1	3	3	1	4	4	0	1	A	16
2	3	2	0	3	4	0	2	B	14
3	2	1	0	1	3	0	0	B	7
4	2	2	1	3	2	1	1	B	12
5	2	2	0	2	3	0	0	B	9
6	2	2	0	3	3	1	1	C	12
7	3	2	0	3	3	0	2	B	13
8	3	3	1	4	3	1	1	B	16
9	2	2	0	4	2	1	2	B	13
10	2	3	1	4	3	0	2	C	15
11	2	3	0	4	4	0	2	D	15
12	1	2	0	3	3	1	1	E	11
13	1	1	0	1	2	0	1	F	6
14	3	3	0	4	4	1	1	C	16
15	2	1	0	2	3	0	2	E	10
16	1	1	0	2	1	0	1	F	7
17	1	1	0	2	1	0	1	F	6
18	1	1	1	2	1	0	2	F	8
19	1	1	0	1	1	0	1	F	5

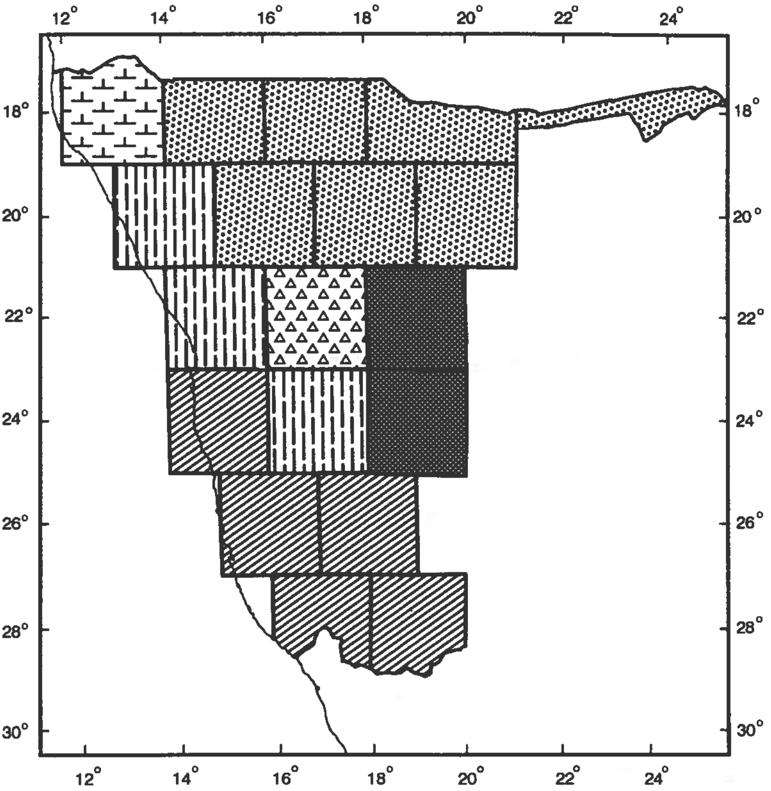

Fig. 15.3. Map of Namibia showing to which of six floristic clusters each subregion belongs.

those in the primary or secondary gene pool of a crop. Table 15.2 therefore also records the occurrence of wild *Citrullus lanatus* in each subregion. A single record in the subregion is given a score of 1, two or more records a score of 2.

Discussion

To rank subregions in terms of the importance of their contribution to overall cucurbit species diversity in Namibia, the scores of each subregion for the seven parameters shown in Table 15.2 were added together. The nature of the categorization of the data gives weights in the ratios of 4 : 4 : 3 : 3 : 2 : 1 : 1 to, respectively, the scores for the two diversity measures, genera and species numbers, occurrence of *Citrullus lanatus*, number of *Cucumis* species and 'undercollectedness'.

Clearly, these weightings can easily be changed to reflect different collecting aims. The five subregions scoring highest are SR1, SR8, SR10–11 and SR14 (Table 15.2). Bringing in the cluster analysis results reveals that SR10 and SR14 are in the same cluster. SR14 would be the higher collecting priority if *C. lanatus* occurrence were to be ignored. With the weight attached to this particular variable here, however, SR10 has the higher total score. This makes the point that different collecting aims and objectives will result in different weights being attached to a given criterion, and possibly therefore different priorities being accorded to a given subregion.

The clustering results also reveal that clusters E and F are not represented in this initial selection of priority subregions. Based on their total scores, subregions SR12 and SR18 can be identified as the higher priorities within clusters E and F respectively.

If it was only possible to visit a limited number of subregions, a possible final selection for first-priority collecting (in numerical order) would thus be as follows: SR1, SR8, SR11, SR12, SR14 and SR18. Germplasm collecting within this limited selection of subregions would, according to the analysis presented here, result in the capture of species diversity within the family adequately representative of the study area as a whole, with something of a bias towards subregions with good representation of the genus *Cucumis* and the species *C. lanatus*. Alternatively, the total scores in Table 15.2 could be used to weight the amount of collecting time or effort in each subregion or, indeed, the number of samples collected in each subregion.

Conclusion

The worked example in this section should be seen as presenting a generalized model capable of being applied to a wide variety of different situations. Thus, though species were considered, it could equally well have been formal or informal infraspecific taxonomic categories within a crop (Chapter 7), traditionally recognized landraces or even genotypes. The aim of the analysis was to devise a strategy for the maximal sampling of species diversity within a botanical family. A separate analysis would be needed if the target were diversity within *C. lanatus*, say, in which case a measure of genetic diversity could be used.

To complement the information on the plants themselves, some measure of environmental diversity within the different subsections could have been calculated, for example by noting the extent of different vegetation types or agroecological zones in each. In the same way, though similarities among subregions were calculated here floristically, on the basis of species representation, they could be calculated as genetic or environmental similarities in other cases. Groups of environmentally similar subregions will form the basis of a stratified sampling strategy in situations when little is known of genetic or taxonomic variation within the target group.

Acknowledgements

Mark Perry and Tom Hazekamp (IPGRI) made valuable comments on early drafts of this paper.

References

Adams, R.P. (1970) Contour mapping and differential systematics of geographic variation. *Systematic Zoology* 19:385–390.

Austin, M.P. and E.M. Adomeit (1991) Sampling strategies costed by simulation. In: Margules, C.R. and M.P. Austin (eds) *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. CSIRO, Australia.

Bates, D.M., R.W. Robinson and C. Jeffrey (eds) (1990) *Biology and Utilization of the Cucurbitaceae*. Cornell University Press, Ithaca.

Blackith, R.E. and R.A. Reyment (1971) *Multivariate Morphometrics*. Academic Press, London.

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

Braak, C.J.F. ter (1988) CANOCO - an extension of DECORANA to analyze species-environment relationships. *Vegetatio* 75:159–160.

Brown, A.H.D. and B.S. Weir (1983) Measuring genetic variability in plant populations. In: Tansley, S.D. and T.J. Orton (eds) *Isozymes in Plant Genetics and Breeding*. Part A. pp. 219–239. Elsevier Scientific Publishing, Amsterdam.

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

Charmet, G. (1993) Geographically constrained clustering: a tool for establishing a core from a large collection of wild populations in forage grasses. In: IBPGR *Report of the Fourth Meeting of the ECP/GR Forages Working Group*. IBPGR, Rome.

Clifford, H.T. and W. Stephenson (1975) *An Introduction to Numerical Classification*. Academic Press, New York.

Cousins, S.H. (1991) Species diversity measurement: choosing the right index. *Trends in Ecology and Evolution* 6:190–192.

Dunn, G. and B.S. Everitt (1982) *An Introduction to Mathematical Taxonomy*. Cambridge University Press, Cambridge.

Erskine, W., Adham, Y. and Holly, L. (1989) Geographic distribution of variation in quantitative traits in a world lentil collection. *Euphytica* 43:97–104.

Francisco-Ortega, J., M.T. Jackson, A. Santos-Guerra and B.V. Forde-Lloyd (1993) Morphological variation in the *Chamaecytisus proliferus* (L.f.) Link complex (Fabaceae: Genisteae) in the Canary Islands. *Botanical Journal of the Linnean Society* 112:187–202.

Gauch, H.G. (1982) *Multivariate Analysis in Community Ecology*. Cambridge University Press, Cambridge.

Goldsmith, F.B., C.M. Harrison and A.J. Morton (1986) Description and analysis of vegetation. In: Moore, P.D. and S.B. Chapman (eds) *Methods in Plant Ecology*. pp. 437–524. Blackwell Scientific Publications, Oxford.

Goodrich, W.J. (1987) Monitoring genetic erosion: detection and assessment. Internal report. IBPGR, Rome.

Gower, J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics* 27:857–872.

Heywood, J.S. (1991) Spatial analysis of genetic variation in plant populations. *Annual Review of Ecology and Systematics* 22:335–355.

Hill, M.O. (1979a) *DECORANA – A FORTRAN Program for Detrended Correspondence Analysis and Reciprocal Averaging.* Cornell University, Ithaca.

Hill, M.O. (1979b) *TWINSPAN – A FORTRAN Program for Arranging Multivariate Data in an Ordered Two-way Table by Classification of Individuals and Attributes.* Cornell University, Ithaca.

Hill, M.O. and H.G. Gauch (1980) Detrended Correspondence Analysis: an improved ordination technique. *Vegetatio* 42:47–58.

Hill, M.O. and A.J.E. Smith (1976) Principal components analysis of taxonomic data with multi-state discrete characters. *Taxon* 25:249–255.

Hintum, Th.J.L. van (1993) A computer compatible system for scoring heterogeneous populations. *Genetic Resources and Crop Evolution* 40:133–136.

Hoelzel, A.R. and D.R. Bancroft (1992) Statistical analysis of variation. In: Hoelzel, A.R. (ed.) *Molecular Genetic Analysis of Populations.* pp. 297–305. IRL Press, Oxford.

Hoelzel, A.R. and G.A. Dover (1991) *Molecular Genetic Ecology.* Oxford University Press, Oxford.

Hutchenson, K. (1970) A test for comparing diversities based on the Shannon formula. *Journal of Theoretical Biology* 29:151–154.

Jones, S.B. (1972) A systematic study of the fasciculate group of *Vernonia. Brittonia* 24:28–45.

Kirkpatrick, J.B. (1974) The use of differential systematics in geographic research. *Area* 6:52–53.

Krajewski, C. (1994) Phylogenetic measures of biodiversity: a comparison and critique. *Biological Conservation* 69:33–39.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29:28–42.

Magurran, A.E. (1988) *Ecological Diversity and its Measurement.* Chapman and Hall, London.

Monmonier, M. (1973) Maximum-difference barriers: an alternative numerical regionalization method. *Geographical Analysis* 3:245–261.

Morden, C.W., J.F. Doebley and K.F. Schertz (1989) Allozyme variation in Old World races of *Sorghum bicolor* (Poaceae). *American Journal of Botany* 76:247–255.

Morley, T. (1971) Geographic variation in a widespread neotropical species, *Mouriri myrtilloides* (Melastomataceae). *Brittonia* 23:413–424.

Myklestad, Å. and H.J.B. Birks (1993) A numerical analysis of the distribution of patterns of *Salix* L. species in Europe. *Journal of Biogeography* 20:1–32.

Nabhan, G.P. (1990) *Wild* Phaseolus *Ecogeography in the Sierra Madre Occidental, Mexico.* Systematic and Ecogeographic Studies on Crop Genepools 5. IBPGR, Rome.

Nei, M. (1975) *Molecular Population Genetics and Evolution.* North-Holland Publishing Co., Amsterdam.

Nei, M. (1987) *Molecular Evolutionary Genetics.* Columbia University Press, New York.

Pankhurst (1991) *Practical Taxonomic Computing.* Cambridge University Press, Cambridge.

Peeters, J.P. (1988) The emergence of new centres of diversity: evidence from barley. *Theoretical and Applied Genetics* 76:737–752.

Pigliucci, M. and G. Barbujani (1991) Geographical pattern of gene frequencies in Italian populations of *Ornithogalum montanum* (Liliaceae). *Genetical Research* 58:95–104.

Porceddu, E. and A.B. Damania (1992) *Sampling Strategies for Conserving Variability of Genetic Resources in Seed Crops*. Technical Manual No. 17. ICARDA, Aleppo.

Rebello, A.G. and W.R. Sigfried (1992) Where should nature reserves be located in the Cape Floristic Region, South Africa? Models for the spatial configuration of a reserve network aimed at maximizing the protection of diversity. *Conservation Biology* 6:243–252.

Richardson, G.T. (1985) *Illustrations*. Humana Press, Clifton.

Seyani, J.H. (1988) The taxonomy of *Dombeÿa burgessiae* complex (Sterculiaceae) in Africa. Monograph. *Systematic Botany, Missouri Botanical Garden* 25.

Sneath, P.H.A. and R.R. Sokal (1973) *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W.H. Freeman, San Francisco.

Sokal, R.R. (1979) Testing statistical significance of geographical variation patterns. *Systematic Zoology* 28:227–232.

Sokal, R.R. (1986) Phenetic taxonomy: theory and methods. *Annual Review of Ecology and Systematics* 17:423–442.

Sokal, R.R. and N.L. Oden (1978) Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society* 10:199–228.

Southwood, T.R.E. (1978) *Ecological Methods*. Chapman and Hall, London.

Spagnoletti Zeuli, P.L. and C.O. Qualset (1987) Geographical diversity for quantitative spike characters in a world collection of durum wheat. *Crop Science* 27:235–241.

Stuessy, T.F. (1990) *Plant Taxonomy*. Columbia University Press, New York.

Vane-Wright, R.I., C.J. Humphries and P.H. Williams (1991) What to protect? Systematics and the agony of choice. *Biological Conservation* 55:235–254.

Weeden, N.F., B. Wolko, A.C. Emmo and J. Burr (1988) Measurement of genetic diversity in pea accessions collected near the center of origin of domesticated pea. Internal report. IBPGR, Rome.

Whitaker, T.W. and W.P. Bemis (1976) Cucurbits. In: Simmonds, N.W. (ed.) *Evolution of Crop Plants*. pp. 64–69. Longman, London.

Williams, P.H., C.J. Humphries and R.I. Vane-Wright (1991) Measuring biodiversity: taxonomic relatedness for conservation priorities. *Australian Systematic Botany* 4:665–679.

Williams, P.H., R.I. Vane-Wright and C.J. Humphries (1993) Measuring biodiversity for choosing conservation areas. In: LaSalle, J. (ed.) *Hymenoptera and Biodiversity*. CAB International, Wallingford.

Womble, W.H. (1951) Differential systematics. *Science* 114:315–322.