THE DEVELOPMENT OF NUMERICAL CLASSIFICATION AND ORDINATION

P. GREIG-SMITH

School of Plant Biology, University College of North Wales, Bangor LL57 2UW, United Kingdom

The invitation to open this session set me thinking about the development of numerical procedures of classification and ordination. Their technical development has been reviewed by various authors from various viewpoints, e.g. Cormack (1971), Orlóci (1975, 1978), Dale (1975), Goodall (1970), Greig-Smith (1954, 1964, 1980), Whittaker (1967, 1973). I do not intend to discuss this in more than broad terms, but there is another aspect that has received less attention. This concerns the influences and constraints which have affected the development of numerical methods and their acceptance by phytosociologists. Their acceptance is particularly important; numerical methods are tools and unless they are used in the investigation of real ecological problems we are wasting our time in developing them.

What I have to say represents a personal view, but will, I hope, be of some interest. My excuse for attempting a broad survey of this kind is that I have been closely involved with numerical classification and ordination throughout their development. This has given me the opportunity to look back at the misconceptions and the failures to recognise what now seems obvious, that occur in all scientific development, but which are not so often talked about.

Numerical analysis of plant communities as we now understand it originated approximately 30 years ago, though interest in certain community attributes e.g. species/area relationships and in numerical approaches to the distribution of individual species had developed earlier and was sufficient to justify review articles in 1936 and 1948 (Ashby 1936, 1948). At this time classification was a long established approach, though there was controversy about the most appropriate system to use and the importance of classification to a broader understanding of vegetation. Although Ramensky had developed a technique of ordination (see Sobolev & Utekhin 1973) this was little known outside Russia and it came as a new approach to most phytosociologists. Ordination was thus linked with numerical approaches from the start and this, together with its association with the continuum concept of vegetation, influenced its reception.

The attitude of most ecologists to mathematics at that time can fairly be described as one of suspicion. This is curious because there was at the same time considerable respect for quantification, so that much time was sometimes devoted to obtaining quantitative data in the field, data largely wasted because no further analysis was made. It is interesting that Tansley could write in 1923 'in proportion to the advance of a branch of science its methods become more quantitative. This is true of biology in general and of ecology in particular as of other branches of science.' With characteristic percipience he went on to warn against gathering quantitative data for their own sake, but this warning was not infrequently ignored. The first edition of Fisher's *Statistical Methods for Research Workers*, which was to have such a profound impact on biology in general, appeared in 1925. One wonders whether this may paradoxically have delayed the development of numerical methods of phytosociology. For most biologists the kind of statistical analysis developed by Fisher, with its emphasis on fit to hypothesis, probability and tests of significance, became the only kind of mathematics that was relevant; it did not prove helpful in dealing with plant communities.

The early advocates of ordination techniques were all supporters of the interpretation of variation in vegetational composition as a continuum e.g. Ramensky, Curtis, Goodall, Whittaker. Acceptance of the continuum view was undoubtedly a powerful stimulus to the development of techniques of ordination, but it is now generally accepted that the choice between classification and ordination depends on the objective of data analysis and the structure of the data set being examined, rather than on preconceptions about the nature of vegetation. Presentation of ordination techniques in the context of continuum undoubtedly led to their being ignored by many ecologists. Misunderstanding was increased by the chance that the first practical technique (Curtis & McIntosh 1951) was illustrated by a set of data in which the first and only axis extracted was a successional one and the method was regarded as aiming at elucidating succession rather than of more general application.

ſhe initial techniques were mostly crude and some could scarcely be regarded as numerical. There was little to attract the interest of mathematicians. Later, as techniques developed, mathematicians were to look at them critically and helpfully, as in Cormack's (1971) review of classification, but in the early stages numerical methods were regarded as irrelevant by most ecologists and as unworthy of notice by mathematicians.

Because of the influence of ordination techniques on classificatory techniques, it is convenient to consider ordination first. Dale (1975) has pointed out that there were three major sources of methods of ordination. One, direct gradient analysis, was dependent on recognition of environmental gradients and sought to relate vegetational data to them (Whittaker, 1952, 1956). The other two both sought patterns in the vegetational data and only after these had been identified was their relation to environmental gradients established. The first of these accepted, though not always explicitly, that a set of data could be considered in relation to as many independent axes as there were species present and argued that the dimensionality of the set could be reduced without serious loss of information if species occurrences are correlated (Goodall 1954, Bray & Curtis 1957). The second, path-seeking or more recently (Noy-Meir 1974) 'catenation', aimed to order stands so that species had a unimodal distribution along the axis (Curtis & McIntosh 1951). Reduction in dimensionality emphasises the overall relations between stands, catenation emphasises the relations between a stand and those most similar to it. Though the distinction between these approaches is evident in retrospect, it was not, I think, generally realised at the time. Even those most interested regarded Bray & Curtis' method at the time as a promising extension of that of Curtis & McIntosh to more than one axis, rather than a fundamentally different approach.

The first approaches were in terms of a single axis only. It is true that Whittaker (1956, 1960), by examining moisture gradients at different altitudes, produced what was in effect a two-dimensional ordination and, by comparing moisture and altitudinal effects on two contrasting soil types, was even able to include two different levels on a third axis, but the technique was essentially one of examination of single environmental gradients. There are two considerations of interest. These very simple ordinations were readily understood, even by those with little or no feel for a numerical approach, provided they were prepared to accept the context of continuum in which they were presented. On the other hand, the emphasis on single recognisable environmental gradients has had an influence on later developments which has not been wholly beneficial.

At this stage the limitations of available techniques were that indirect gradient analysis allowed for the derivation of a single axis only and was difficult to apply unless a single gradient was overriding, as in a successional situation, and direct gradient analysis depended on prior recognition of the most important environmental gradients; it was no accident that Whittaker's method was developed in a region of marked topographic variation. It was not surprising that attention was then concentrated on indirect methods that allowed more flexible derivation of several axes.

There were two independent approaches to multiaxis ordination, both dimension-reducing. Goodall (1954) applied a form of principal component analysis and Bray & Curtis (1957) used an essentially geometric approach to reduce the dimensionality of a matrix of interstand dissimilarities or distances. There is an interesting contrast in the subsequent developments from these two pioneer papers. Principal component analysis was not followed up as a practical tool until considerably later, though Dagnelie (1960) discussed the use of both it and factor analysis. Bray & Curtis' technique was relatively soon being used and appears as an important tool in Curtis' *The Vegetation of Wisconsin* (Curtis 1959). Although principal components analysis is more difficult to understand than the very simple Bray & Curtis technique, the delay in following up Goodall's suggestion resulted primarily from the computational load involved. Not until electronic computers of sufficient speed and capacity had been developed did it become feasible to use principal component analysis for any considerable body of data. This is not the only case where the application of analytical techniques has been delayed by the unavailability of adequate computing facilities, rapid though the development of the latter has been. By contrast, Bray & Curtis' technique is feasible 'by hand' i.e. with only a simple desk calculator; Ashton (1964) analysed a very considerable body of data from a species-rich tropical forest by hand. That it was developed in the very active group led by the late J.T. Curtis at the University of Wisconsin also contributed to its rapid adoption.

The simple ordination of Bray & Curtis was criticised on various grounds and modifications were proposed (e.g. Orlóci 1966, 1974, Austin & Orlóci 1966, Swan, Dix &

Wehrhahn 1969). Swan, Dix & Wehrhahn, noting that the use of the two stands farthest apart from one another in the unreduced species space as the endpoints of the first axis might not give the most efficient ordination, proposed to try all possible pairs of stands as endpoints and accept that pair giving the most efficient analysis, i.e. that retaining the maximum total interstand distance along the axis. To do this is computationally demanding and calls for consideration whether, given that the computational load is comparable, the technique has advantages on other grounds. This demonstrates a general risk, often ignored, that modification of a simple technique may eliminate one of its major advantages, its speed and simplicity.

Principal component and factor analyses have provoked more discussion and, I think, have been more misunderstood than any other numerical technique. There has been misunderstanding of the relation between principal component analysis and factor analysis (e.g. Greig-Smith 1964). They are deceptively similar in form, but involve different assumptions. Principal components analysis is essentially a different presentation of the data without any reduction in dimensionality; the initial variates, normally species, are transformed to an equal number of orthogonal variates, or components. Its value lies in the derivation of components in order of decreasing 'variance accounted for' and we may choose to ignore all but the first few components and still retain a large proportion of the information in the data. Factor analysis involves an assumption about the number of independent factors needed to account for the observed correlations between the occurrences of species. As Dale (1975) has commented, factor analysis appears attractive as an ordination technique, but there are formidable difficulties in practice (see also Williams 1976) and it has not been widely used.

Principal components analysis goes back to a paper of Karl Pearson in 1901 with the austere title 'On lines and planes of closest fit to a system of points in space'. Factor analysis was developed and both it and principal components analysis were initially mainly used in psychology, in attempts to define a limited number of independent factors of human ability from the results of psychological tests. This early association with psychology constrained the development of the use of principal component analysis in vegetational ordination. Principal component analysis involves two distinct stages: an initial transformation of the data and the extraction of the eigen values and eigen vectors of the matrix of cross-products between the transformed data. The transformation may involve either or both centering and some form of standardisation. Psychologists, for reasons which need not concern us, had necessarily to centre their data and to standardise by standard deviate i.e. to operate on a matrix of correlation coefficients. When principal component analysis was introduced into vegetational analysis, it was accepted uncritically that the correlation coefficient was the appropriate similarity measure to use. In many data sets all species are measured on the same scale and covariance was therefore sometimes used instead, but it remained for Noy-Meir (1973a, Noy-Meir, Walker & Williams 1975) to clarify the situation and to point out the advantages of not centering the data if they are markedly heterogeneous, as extensive field data often are, and to emphasise that different standardisations represent different weightings. Data standardisation is essentially a matter for biological, not mathematical, decision and depends on the answers to such questions as 'Are rare and common species to be given equal weight?' 'Are differences in standing crop to be ignored?'

It was early recognised that principal component analysis could operate on either cross-products between species or cross-products between stands, 'R' and 'Q' techniques, to give different ordinations, often both interpretable in ecological terms, and there was argument about which was more appropriate. This again resulted from the failure to recognise the two stage nature of principal component analysis. A correlation coefficient between species implies centering by species, a correlation coefficient between stands centering by stands and it is not surprising that they give different ordinations. If the same centering is used, R and Q analyses give the same ordination. Independent recognition of this by Gower (1966) and Orlóci (1967) allowed the useful economy in computing of choosing an R or Q analysis according to whether fewer species or stands are involved.

Misunderstanding of the use of principal components analysis in vegetational analyses was not confined to ecologists. Statisticians, the mathematicians most concerned, viewed principal components analysis as usefully applicable only if each of the variables is normally distributed, which is certainly not true of most real vegetational data. This discouraging judgement, which has probably deterred many ecologists, was presumably based on the mistaken idea that knowledge of the number of 'significant' components is of prime importance. Certainly, non-normal distribution prevents the valid application of significance tests, but these are irrelevant when the objective is data exploration rather than the testing of hypotheses (see below).

Principal components analysis has severe limitations, which were soon recognised, as an ordination technique. The underlying model assumes linearity of response to each component and additivity of response to different components. These are clearly unrealistic assumptions in relation to the control of species performance by the environment and non-linearity especially has been much discussed (see Austin 1976). There is abundant evidence from experimental work on the response of species to the levels of environmental factors and from direct gradient analyses that response curves are not only not linear, but they are not even monotonic except over narrow ranges. Typically they are unimodal, but may be bimodal as a result of competitive effects (Ellenberg 1953). The result is that if a single gradient with species showing bell-shaped response curves along it is ordinated by principal component analysis, the gradient is not recovered by a single axis but requires two or more dimensions to display it and may be infolded, making interpretation difficult (Swan 1970, Noy-Meir & Austin 1970).

Attempts over the last few years to develop more satisfactory techniques of ordination raise several interesting questions. Do the effects of non-linearity of response curves on the resulting ordination matter? If the objective is to examine individual species response curves, they clearly do. If, however, the objective is to explore the data in order to erect hypotheses about the control of composition of the vegetation by environment, the answer is less certain. A considerable amount of non-linearity in the pattern of a gradient of composition on the ordination will still permit recognition of correlation with values of environmental factors, the basis of hypothesis generation. To assess this we must turn to cases of the use of ordination as a tool in a real situation. Hall & Swaine (1976) examined a very extensive set of data from forests in Ghana by reciprocal averaging ordination (Hill 1973) and found it a fruitful approach. Reciprocal averaging, which can be regarded as a particular form of non-centered principal components analysis, is admittedly less vulnerable to the effects of non-linearity but still shows them. Greig-Smith, Austin & Whitmore (1967) used a conventional centered principal components analysis on data from rain forest in the Solomon Islands and found it profitable.

Procedures have been suggested for ordinating stands in such a way that the individual species values give the best fit to smooth response curves (Gauch, Chase & Whittaker 1974, Ihm & van Groenewoud 1975). These raise the problem of the appropriate form of response curve to use. It has commonly been assumed that response curves are Gaussian in form. Even if they are symmetrical, and Austin (1976) has argued convincingly that this assumption is unjustified, there seems to be a fundamental misunderstanding here. The Gaussian curve reflects the influence of effectively random deviations, due to numerous minor influencing factors, on the probability of observing a particular value of a variable in any one observation. This seems irrelevant to the response of a species to an ordered environmental gradient, though it may give an approximation to a symmetrical response curve, the exact form of which we do not know. Are attempts to ordinate data by fitting to response curves chasing a 'will o' the wisp' ('Irrlicht')?

It is true that if simulated data constructed from a series of Gaussian curves are analysed in this way, an efficient retrieval of the gradient is achieved, but this begs the question how the technique will perform with real data and opens up the whole problem of assessing the efficiency of ordination methods. The earlier approach was to calculate the percentage of variation in the original data accounted for by the ordination, but this is in terms of variation fed into the analysis; the choice of similarity or distance measure determines what information is used. The alternative of testing methods on simulated data has little relevance until we know how to simulate realistic data.

It is,worth emphasising that a dichotomy has developed in the objective of ordination, a dichotomy between data exploration as a basis for generating hypotheses about the relation between composition of the vegetation and its environment on the one hand, and elucidation of the relationship of individual species to environmental gradients on the other. The former appears to be dominant in most practical applications of ordination but can, so far, only be assessed empirically by the degree to which it is found to be helpful. The latter has figured prominently in recent methodological studies, perhaps as a result of the chance that many of the earlier procedures were concentrated on single axes, but has played much less part in practical application. I wonder how useful this methodological concentration on the narrower objective will prove to have been; I suspect that problems of species response are better tackled more directly.

Unlike numerical methods of ordination, numerical methods of classification developed against a background of a range of well-established non-numerical methods. It is interesting to consider the interaction between established views of vegetational classification and the development of numerical methods.

Three objectives can be identified in the classification of vegetation, though they are often not explicitly stated and more than one objective may be covered by one procedure. 1) Classification has one very practical function, as a basis of inventory and mapping, either as an objective in itself or as a basis of management. This is present in all the traditional systems, and at its most empirical represents a convenient partition of a range of variation which may or may not be continuous. 2) Classification may aim to identify 'real' entities with clear discontinuities between them, the antithesis of the concept of vegetation as a continuum. It is not always clear whether there is an element of this objective in a particular classificatory system or not. 3) Classification may be a tool in the exploration of correlations between vegetation and environment.

In addition to the general aversion to numerical methods, already referred to, other considerations contributed to a reluctance to accept numerical classification.

Though in practice most non-numerical systems based their classification on detailed recording of a limited number of stands, they aimed to produce a generally valid system into which further stands could be placed, i.e. a system comparable to a taxonomic treatment. The earliest numerical procedures, in contrast, were presented in the context of the examination of the relationships of a particular set of stands, with the implication that a different set from the same range of vegetation could give a different classification; the emphasis was almost entirely on the third objective.

Numerical and non-numerical approaches have in common the aim of producing final groups which are as homogeneous in composition as possible, but there were deeply entrenched convictions about the kind of species that would provide the most efficient criteria for doing so, e.g. dominant species, constant species, species of a particular life form. The distinctive contribution of numerical methods is to allow the data themselves to indicate the most efficient criteria; this came as a novel and unfamiliar idea.

As with the early development of ordination, the history of numerical classification shows ideas running ahead of computational facilities. The strategy of producing a classification may be either divisive or agglomerative. The whole set of data may be successively divided into subsets on an appropriate criterion to produce a hierarchy (divisive strategy) or individual stands may be grouped on an appropriate criterion and the resultant groups in turn grouped until all stands are finally fused into a single group, building a hierarchy from the bottom (agglomerative strategy). Both approaches are used in non-numerical systems, e.g. classification by dominant species is essentially divisive, the Braun-Blanquet system is agglomerative. Further, a strategy may be *monothetic*, based on a single criterion at each stage, i.e. the presence or absence of a single species, or *polythetic*, using many species as the criteria at each stage, i.e. assessment of overall similarity between stands. Again, both strategies are found in non-numerical systems; classification by dominants is monothetic, the Braun-Blanquet system is polythetic.

The first numerical method to be used at all widely was divisive (Williams & Lambert 1959, 1960, following on a suggestion of Goodall 1953) although Sørensen (1948) had earlier proposed an agglomerative method. It is perhaps no more than chance that the former was produced by workers trained in the Anglo-American tradition, but Sørensen came from the Scandinavian agglomerative tradition.

In principle, divisive-monothetic classification is straightforward. The data are divided on the presence or absence of each species in turn and that division is accepted which gives the minimum residual heterogeneity, measured in some appropriate way, within the two resulting subgroups. To do this, however, initially involved an unacceptable amount of computation. This led Williams & Lambert (1959) to suggest that division on that species which had the greatest amount of association with other species would tend to give the greatest reduction in heterogeneity. The resulting *association-analysis* was widely used, but with the increasing speed and capacity of computers, it became feasible to try division on each species in turn and a variety of methods resulted, differing only in the measure of heterogeneity used.

Monothetic procedures have the disadvantage that they ignore much of the information in the data. A divisive-polythetic strategy is not possible non-numerically in most circumstances, but the numerical approach is again straightforward in principle (Edwards & Cavalli-Sforza 1965); all possible divisions of the data into two are examined and that one is accepted which gives the maximum reduction in heterogeneity. With increasing number of stands, the number of possible divisions ($2^{n-1} - 1$) increases rapidly and the method is still not possible. The impossibility of this direct approach in practice led to various forms of 'directed search' which aimed to eliminate the less efficient divisions without having to test them (Macnaughton-Smith et al. 1964, Gower 1967, Lambert 1972). An alternative approach which has also produced a variety of methods is to start with the first axis of an ordination of the data and accept the most efficient split of that axis as the criterion for subdivision (Lambert 1972, Lambert et al. 1973, Noy-Meir 1973b, Hill. Bunce & Shaw 1975).

Agglomerative classification presents two problems, the choice of similarity or distance measure, and the strategy of fusion. Both have been the subject of misunderstanding, at least by potential users. Just as the preliminary data transformation in principal component analysis was confused with the analysis itself, the distinctive part played by data transformation in agglomerative classification has been misunderstood. Some measures involve no transformation, unless this is done as a separate preliminary operation e.g. Euclidean distance, others a readily recognised standardisation e.g. the correlation coefficient, and others, and this is where misunderstanding has been most evident, a standardisation differing for each comparison e.g. Sørensen's coefficient, standardised by the sum of the two stands being compared. Different standardisations give markedly different hierarchies (Austin & Greig-Smith 1968) because different aspects of species representation are emphasised; a conscious decision on standardisation is necessary, but this has often not been realised.

Choice of fusion strategy determines the way the distance between a group and a single stand, or between two groups, is measured. To take two contrasting strategies only, a stand may be regarded as having a distance from a group equal to its distance from the nearest member of that group (nearest-neighbour or single-link sorting). Alternatively, the stand may be regarded as having a distance from the group equal to its distance from the member of the group furthest away from it (furthest neighbour or complete-linkage sorting). There are a number of other possible strategies but only nearest-neighbour sorting is free from ambiguity if there is more than one case of the shortest observed difference at any stage. Unfortunately, nearest-neighbour sorting produces very strongly 'chained' hierarchies; once a group is formed further stands tend to be added to it rather than form new groups. Chained hierarchies are almost useless ecologically, either for producing a general purpose classification or for examining correlation with environment. There has been an interesting controversy over the importance of ambiguity. Sibson (1971) has argued that classification must be unambiguous and hence only nearest-neighbour sorting should be used, a view vigorously opposed by Williams et al. (1971), who take the more pragmatic view that a classification must be useful.

There is, I think, more to this disagreement that the contrast between the views of theorists and those who deal with real data. The numerical classification of vegetation has many apparent similarities with numerical taxonomy and each has influenced the development of the other, but there are important differences in assumptions and objectives. Taxonomy, in most cases, deals with what are believed to be real entities, having discontinuities between them, however difficult these may be to identify. Any ambiguity in procedures is therefore disturbing. Paradoxically, ambiguity is not a real problem in numerical taxonomy because operations start some way up a hierarchy – with 'orthodox taxonomic units' rather than individuals – and some procedures have been used successfully that are theoretically capable of giving rise to ambiguities.

Though the relation between taxonomy and phylogeny is a matter for argument, most taxonomists do appear to accept that taxonomic arrangement reflects phylogenetic relationship. *Degree* of similarity between groups is then more than a tool in constructing a classification; it is of interest in itself. This is not true of a vegetation classification, where we are interested only in erecting useful categories (for inventory, mapping, etc.) or in elucidating correlation between vegetation and environment as a means of generating hypotheses about the factors determining the composition of vegetation.

Interaction with numerical taxonomy has perhaps also influenced the relative attention paid to divisive and agglomerative strategies. At least until the opening up of divisive-polythetic classification, agglomerative strategies were more likely to give efficient classifications. Because it starts with orthodox taxonomic units, taxonomy normally deals with a relatively small number of individuals in any one analysis and the computational load of agglomerative strategy is not an obstacle. Ecologically the situation is different; in any real data set there are liable to be a large number of individuals and not until a relatively long way up the hierarchy are results likely to be of interest. Divisive strategy, which can be stopped when the appropriate level is reached, is attractive. Much effort has been put into developing agglomerative techniques which then proved unattractive to users because they involved so much unrewarding computation with large data sets.

As with ordination techniques, assessment of the efficiency of techniques of numerical classification presents problems. Blackith & Rayment (1971) have put it well '. . . there are no objective criteria against which the classifications can be judged. There is, therefore, a tendency for multivariate techniques to be condemned when they disagree with conventional methods, and regarded as superfluous when they agree.' Again, we can only judge by results, not by whether they reproduce our preconceptions, but by whether they are useful in practice or are fruitful of hypotheses. As experience accumulates, we are likely to be able to make a more informed guess as to which techniques are likely to be satisfactory in a given situation.

I have ranged rather erratically over the development of numerical methods. What is their future? They have three principal advantages. They can disclose features which are not revealed by non-numerical methods because relationships are too complex to analyse subjectively. They are particularly useful in little-known or very complex vegetation, such as tropical rain forest. They allow more efficient use of a scarce resource, the skill to interpret the complexity of vegetation in the field. Much, though not all, of the sorting of information that numerical methods achieve can be done by someone with the necessary aptitudes and experience, but such people are better employed in the ultimate interpretation.

There are certain dangers. Are we perhaps too concerned with refinements of methodology? I reiterate my belief that numerical methods are only worth developing if they are to be used on real data in attempts to answer real questions. There are limitations to real data, limitations not only of accuracy of quantitative measures, but also of the reliability of human observation. Hall & Okali (1978) have made revealing observations on the degree to which data from secondary forest in Nigeria are affected by season and by the experience of observers. It is clear that we must expect a considerable degree of inaccuracy in field data from all but very simple vegetation. Are such data adequate input for very refined methods of analysis?

Related to practical use too is the danger of what may be termed the 'black box syndrome'. With increasingly complex methods, and the increasing availability of computer programmes for these methods, it becomes all too easy for the user to take a programme and use it without understanding what it does. Association-analysis, one of the earliest classificatory methods, has been widely used. It is revealing to examine the user literature and see how frequently the method has been misunderstood in important respects although it is a very straightforward one and was clearly explained when it was introduced (Williams & Lambert 1959, 1960). With more complicated methods and the sophistication of modern computers, the risk of misunderstanding and consequent misuse is greater.

Summary

The paper reviews the constraints and influences which have affected the development of numerical classification and ordination of vegetation.

Initial development of ordination techniques and their reception by ecologists was hindered by the mistaken idea that ordination involved acceptance of variation in vegetation as a continuum, as well as by a general suspicion of mathematical approaches.

Three distinct approaches to ordination, largely unrecognised at the time, are apparent in earlier work: direct gradient analysis, reduction in dimensionality and path-seeking (catenation) (Dale 1975).

Modifications of simple initial techniques made them more efficient at the cost of increased computation. Acceptance of heavier computation as computers increased in capacity and speed turned attention to principal component analysis and the superficially similar factor analysis. These have been widely misunderstood largely because they were initially applied in the same way as in the analysis of psychological data, in which different constraints and objectives apply. The initial failure to recognise that principal component analysis involves a preliminary data transformation, the form of which depends on answers to biological, not mathematical, questions, was particularly unfortunate.

Principal component analysis has limitations as a technique of ordination resulting from its assumptions of linearity and additivity of plant responses. Attempts to devise more effective techniques raise questions about the practical importance of non-linearity if the objective is data-exploration rather than elucidating the nature of species-response curves and about the adequacy of using simulated data as test data when we do not know how to simulate realistic data.

Data-exploration has been more prominent in practical uses of ordination but many methodological developments have concentrated rather on species-response curves.

Numerical classification also met obstacles to its acceptance additional to a general aversion to numerical techniques. The first numerical techniques were presented in the context of the relationships of a particular set of data, rather than of a generally valid system, which was the more familiar concept in non-numerical classification.

Both numerical and non-numerical classification aim to produce as homogeneous groups as possible. The distinctive contribution of numerical methods is to allow the data to indicate the most efficient criteria of classification; this was an unfamiliar idea.

The strategy of classification may be either divisive or agglomerative and either monothetic or polythetic. Choice of strategy in earlier work was not only constrained by computational limitation but may also have been influ-

enced by an author's previous experience of non-numerical classification. As with ordination, the distinction between preliminary data transformation and subsequent analysis was at first not appreciated.

Numerical classification has been influenced by parallel numerical developments in formal taxonomy. Because objectives and assumptions are not always the same, this influence has not been altogether helpful.

The limitations of real data suggest that developments of technique are at risk of becoming too concerned with refinements of methodology. Increasingly complex methods and increasing availability of programmes for such methods carry the risk that they may be used without adequate understanding of what they do.

References

Ashby,E.1936. Statistical ecology. Bot. Rev. 2: 221–35.

Ashby, E. 1948. Statistical ecology. II. A reassessment. Bot. Rev. 14: 222–34.

Ashton, P.S. 1964. Ecological studies in the mixed dipterocarp forest of Brunei State. Oxf. For. Mem. 25.

Austin, M.P. 1976. On non-linear species response models in ordination. Vegetatio 33: 33–41.

Austin, M.P. & P. Greig-Smith. 1968. The application of quantitative methods to vegetation survey. II. Some methodological problems of data from rain forest. J. Ecol. 56: 827–44.

Austin, M.P. & L. Orloci. 1966. Geometric models in ecology. II. An evaluation of some ordination techniques. J. Ecol. 54: 217–27.

Bray, J.R. & J.T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. Ecol. Monogr. 27: 325–49.

Blackith, R.E. & R.A. Reyment. 1971. Multivariate Morphometrics. Academic Press, London and New York.

Cormack, R.M. 1971. A review of classification. Jl R. statist. Soc. A. 134: 321–67.

Curtis, J.T. 1959. The Vegetation of Wisconsin. Univ. of Wisconsin Press, Madison.

Curtis, J.T. & R.P. McIntosh. 1951. An upland forest continuum in the prairie-forest border region of Wisconsin. Ecology 32: 476–96.

Dale, M.B. 1975. On objectives of methods of ordination. Vegetatio 30: 15–32.

Dagnelie, P. 1960. Contribution à l'étude des communautés végétales par l'analyse factorielle. Bull. Serv. Carte phytogéogr. Sér. B 5: 7–71, 93–105.

Edwards, A.W.F. & L.L. Cavalli-Sforza. 1965. A method for cluster analysis. Biometrics 21: 39–63.

Ellenberg, H. 1953. Physiologisches und ökologisches Verhalten derselben Pflanzenarten. Ber. dt. bot. Ges. 65: 350–61.

Fisher, R.A. 1925. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.

Gauch, H.G., G.B. Chase. & R.H. Whittaker. 1974. Ordination of vegetation samples by Gaussian species distribution. Ecology 55: 1382–90.

Goodall, D.W. 1953. Objective methods for the classification of vegetation. I. The use of positive interspecific correlation. Aust. J. Bot. 1: 39–63.

Goodall, D.W. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Aust. J. Bot. 2: 304–24.

Goodall, D.W. 1970. Statistical plant ecology. Ann. Rev. Ecol. Syst. 1: 99–124.

Gower, J.C. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325–38.

Gower, J.C. 1967. Multivariate analysis and multidimensional geometry. Statistician 17: 13–28.

Greig-Smith, P. 1957. Quantitative Plant Ecology. Butterworth, London.

Greig-Smith, P. 1964. Quantitative Plant Ecology, 2nd edn. Butterworth, London.

Greig-Smith, P. 1980. Quantitative Plant Ecology, 3rd edn. In preparation.

Greig-Smith, P., M.P. Austin & T.C. Whitmore. 1967. The application of quantitative methods to vegetation survey. I. Association-analysis and principal component ordination of rain forest. J. Ecol. 55: 483–503.

Hall, John B. & D.U.U. Okali. 1978. Observer-bias in a floristic survey of complex tropical vegetation. J. Ecol. 66: 241–9.

Hall, J.B. & M.D. Swaine. 1976. Classification and ecology of closed-canopy forest in Ghana. J. Ecol. 64: 913–51.

Hill, M.O. 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61: 237–49.

Hill, M.O., R.G.H. Bunce & M.W. Shaw. 1975. Indicator species analysis, a divisive polythetic method of classification, and its application to a survey of native pinewoods in Scotland. J. Ecol. 63: 597–613.

Ihm, P. & H. van Groenewoud. 1975. A multivariate ordering of vegetation data based on Gaussian type gradient response curves. J. Ecol. 63: 767–77.

Lambert, J.M. 1972. Theoretical models for large-scale vegetation survey. Mathematical Models in Ecology (ed. by J.N.R. Jeffers), pp. 87–109. Blackwell, Oxford.

Lambert, J.M., S.E. Meacock, J. Barrs & P.F.M. Smartt. 1973. AXOR and MONIT: two new polythetic-divisive strategies for hierarchical classification. Taxon 22: 173–6.

Macnaughton-Smith, P., W.T. Williams. M.B. Dale. & L.G. Mockett. 1964. Dissimilarity analysis: a new technique of hierarchical subdivision. Nature, Lond. 202: 1034–5.

Noy-Meir, I. 1973a. Data transformations in ecological ordinations. I. Some advantages of non-centering. J. Ecol. 61: 329–41.

Noy-Meir, I. 1973b. Divisive polythetic classification of vegetation data by optimized division on ordination components. J. Ecol. 61: 753–60.

Noy-Meir, I. 1974. Catenation: quantitative methods for the definition of coenoclines. Vegetatio 29: 89–99.

Noy-Meir, I. & M.P. Austin. 1970. Principal component ordination and simulated vegetational data. Ecology 51: 551–2.

Noy-Meir, I., D. Walker & W.T. Williams. 1975. Data transfortions in ecological ordination. II. On the meaning of data standardization. J. Ecol. 63: 779–800.

Orlóci, L. 1966. Geometric models in ecology. I. The theory and application of some ordination methods. J. Ecol. 54: 193–215.

Orlóci, L. 1967. Data centering: a review and evaluation with reference to component analysis. Syst. Zool. 16: 208–12.

Orlóci, L. 1974. Revisions for the Bray & Curtis ordination. Can. J. Bot. 52: 1773–6.

Orlóci, L. 1975. Multivariate Analysis in Vegetation Research. W. Junk, The Hague.

Orlóci, L. 1978. Multivariate Analysis in Vegetation Research, 2nd edn. W. Junk, The Hague.

Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Phil. Mag. 6: 559–72.

Sibson, R. 1971. Some observations on a paper by Lance & Williams. Comput. J. 14: 156–7.

Sobolev, L.N. & V.D. Utekhin. 1973. Russian (Ramensky) approaches to community systematization. Ordination and Classification of Communities (Handbook of Vegetation Science Vol. 5), p 75–103. W. Junk, The Hague.

Sϕrensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content. Biol. Skr. 5(4): 1–35.

Swan, J.M.A. 1970. An examination of some ordination problems by use of simulated vegetation data. Ecology 51: 89–102.

Swan, J.M.A., R.L.Dix. & C.F. Wehrhahn. 1969. An ordination technique based on the best possible stand-defined axes and its application to vegetational analysis. Ecology 50: 206–12.

Tansley, A.G. 1923. Practical Plant Ecology. George Allen and Unwin, London.

Whittaker, R.H. 1952. A study of summer foliage insect communities in the Great Smoky Mountains. Ecol. Monogr. 22: 1–44.

Whittaker, R.H. 1956. Vegetation of the Great Smoky Mountains. Ecol. Monogr. 26: 1–80.

Whittaker, R.H. 1960. Vegetation of the Siskiyou Mountains, Oregon and California. Ecol. Monogr. 30: 279–338.

Whittaker, R.H. 1967. Gradient analysis of vegetation. Biol. Rev. 42: 207–64.

Whittaker, R.H. (ed.) 1973. Ordination and Classification of Communities (Handbook of Vegetation Science, Vol. 5). W. Junk, The Hague.

Williams, W.T. (ed.), 1976. Pattern Analysis in Agricultural Science. CSIRO and Elsevier, Melbourne and Amsterdam.

Williams, W.T. & J.M. Lambert. 1959. Multivariate methods in plant ecology. I. Association-analysis in plant communities. J. Ecol. 47: 83–101.

Williams, W.T. & J.M. Lambert, 1960. Multivariate methods in plant ecology. II. The use of an electronic digital computer for association-analysis. J. Ecol. 48: 689–710.

Williams, W.T., G.N. Lance, M.B. Dale. & H.T. Clifford. 1971. Controversy concerning the criteria for taxonometric strategies. Comput. J. 14: 162–5.

Accepted 20 December 1979