

4

Interpreting Configurations

'Do you know what this is?'
'No' said Piglet
'It's an A'
'Oh', said Piglet
'Not O—A' said Eeyore severely

A. A. MILNE (The House at Pooh Corner)

The last chapter demonstrated how it is possible to construct a picture of a set of data in the form of a configuration of points, the distances between which represent the dissimilarity between the objects on which the data were collected. Such a representation is easier to assimilate than a matrix of coefficients, but we do not simply want to provide a picture but also, hopefully, to discover hitherto unremarked or unnoticed characteristics of the data. In other words we wish to interpret the configuration.

4.1 What Information is Significant and Stable?

Our interpretation is essentially a two-stage process. First, we look for significant patterns in the configuration, i.e. detect structure, and secondly we ascribe a meaning or interpretation to those patterns or structures. In so doing, it is important that the patterns make use of features of the configuration that are not simply arbitrary artefacts of the scaling procedure. The first questions, then, are: what information is significant and which aspects are stable in a configuration obtained from the basic MDS model? This restriction is important since output from other models sometimes preserves different significant information. The most basic significant information in the solution is the set of relative distances between the points. In the program these distances were calculated from a configuration which has a number of arbitrary, conventional, characteristics. In particular, the configuration was rotated to principal axes (see 3.5.5), and its actual size is also arbitrary, since it is the relative and not the absolute distances which are pertinent. This means that we may legitimately change the axes and the co-ordinate values so long as the relative distances remain unchanged.

The changes or transformations which may legitimately be made to a configuration in the case of the basic model are referred to as a similarity transformation: these consist of rotation, reflection, rescaling and translation (the three 'rs', with a foreign language for good measure?). These turn out to be of vital importance in Chapter 7 when we discuss the comparison of configurations. A detailed definition of similarity transformations is contained in Appendix A7.1 but an intuitive description can be given as follows.

Rotation of axes

MDS solutions using the Euclidean distance model are invariant under rotation—

i.e. the significant distance information is in no way changed by rigid rotation of the axes. Any set of axes which are orthogonal (at 90° to each other) will do as well as the axes given in the final configuration from an MDS program.

Reflection of axes

Consider a configuration reproduced on a photographic slide. It does not affect the distances in any way if the slide is looked at from the front or from the back, or indeed whether the slide is the right way up. Technically a reflection occurs when the positive or negative signs of the co-ordinate values on a given dimension are systematically changed.

Rescaling

The co-ordinates of the space may be *uniformly* rescaled, that is stretched or shrunk without destroying the significant information in the solution. This is simply another way of saying that relative distances are ratio-level quantities and may be multiplied by a constant.

Translation of origin

The origin, or zero point on all the axes, may be located at will within the space without changing the distances. (Note however that this does not apply for factor, or vector, solutions).

Clearly we do not want to interpret or make use of information in the configuration which may be affected by any of these transformations, and in the same way we do not want to make use of aspects of the configuration which are not stable.

We noted in Chapter 3 that the position of a particular point in the configuration was not uniquely fixed by the procedure, but rather fixed within a portion of the space known as its isotonic region, within which the point might be positioned without affecting stress. Obviously, the more data there are, the smaller such isotonic regions become. It remains the case, however, that whereas the overall (global) structure of a configuration is stable and reliable, local information (information about close or adjacent points) is not stable because of this freedom to move each point within its isotonic region.

We shall see that it is rarely possible to interpret the whole of a configuration. Rather, we will be discovering structure within parts of it. We want to be careful not to place too much stress (pun intended) on these local instabilities. Sometimes we can detect which points are most unstable in their location—either from studying the point contribution to stress (see 3.6.2.1), and/or by making several runs from different initial configurations and comparing the final configurations to see which points tend to change location. But in any event, analysis which relies upon small differences of location is not recommended, since it will almost certainly capitalise upon non-unique and unstable characteristics of the solution.

4.2 Internal and External Interpretation

Arrangements of points in a space do not normally exhibit any self-evident structure; we have to bring additional information to the task of interpretation. Two aspects will concern us particularly: pattern and meaning. Patterning refers to

the way in which points are located and related quite independently of what they may 'mean'. This may be evident by inspection or discovered by analysis. It is not usually difficult, for instance, to identify sets of points forming a straight line, or a circle, or a parabola, or even a set of discrete clusters—but it is more difficult to pick out general directions, or overlapping clumps, or even a 2-dimensional plane in a 3-dimensional space.

The 'meaning' of a configuration is a more complex matter. Once labels are attached to the points we bring all sorts of other information into play—what we know about the objects, what connotations they have for us, what subjects said about them and so on. As these meanings are put together we begin to recognise more subtle relationships. In fortunate circumstances, hitherto unsuspected characteristics of the data may then become apparent.

There is no procedure that will automatically detect structure in a configuration. The procedures described below will only assist the user to set about the task of identification in a fairly systematic manner, but there is no guarantee that the types of structure identified will be the most significant ones or relevant in any particular analysis.

It is worthwhile at this point to distinguish between *internal* and *external* methods of interpretation. In internal analysis only the original data are used in interpretation, whereas additional information is employed in the external case.

If we are to use only the original data in our interpretation we have two broad alternatives. Aspects of the original data may be represented in a graph-theoretic way, as line-segments within the configuration. This is a useful method by which simple structures may be identified. Alternatively, we may submit the same data to a clustering analysis and use this to interpret the scaling solution.

On the other hand, interpretation is often made easier by using information about the stimuli obtained independently of (externally to) the scaling itself. For instance:

in scaling subjects' judgments of similarity between nations one might use economic and political information on the nations concerned (cf. Wish 1972);

in scaling judgments of psychological stimuli (subjective loudness, brightness, colour saturation and hue), one would use information on the physical or objective variables involved (Carroll and Wish 1974);

in scaling judgments of personality-trait words, one might use known semantic properties and/or subjects' own ratings of the general properties of the words (Rosenberg and Sedlak 1972).

A closely related source of information for interpreting the meaning of a configuration, at least when the original data are obtained from human subjects, is what they themselves say when generating the original data. This is a much under-used resource, since a fair proportion of studies encourage subjects to verbalise as they produce data. Subjects' comments can be inspected and analysed in terms of their general semantic or particular substantive content and then related to features of the scaling representation.

4.3 Internal Methods of Interpretation

In this section we consider three aspects of configurations which have received

particular attention: the *dimensions* (orthogonal axes which span the multidimensional space), the *regions* (or concentrations or high density of points, differentiated from others by empty regions) and the *simple structures* (identifiable one- and two-dimensional simple patterns).

4.3.1 *Spanning dimensions*

A set of orthogonal reference axes is necessary to locate any set of points. As we have seen, in the case of simple Euclidean distance the orientation of the axes is arbitrary yet a good deal of effort has been devoted, especially in the factor analysis tradition, to 'identifying' or naming them.

The important characteristics of a dimension, within this tradition, are that it represents a higher order organising construct ('factor'), which can be thought of as varying continuously and is bipolar (i.e. varies in both a positive and negative direction), and defines a major pattern of variation in the data (cf. Rummel 1970, ch. 21).

Having decided upon a set of reference axes, the dimensional analyst first separates out the objects or stimuli with the most extreme (positive and negative) co-ordinates on each dimension compared to those nearest zero, in order to establish which are the relevant and which the irrelevant objects for identifying the dimension concerned. Then the objects with the highest co-ordinates are compared to those with the lowest co-ordinates in order to identify the bipolarity of the dimension, or the contrast involved, if such there be.

The process of naming the dimension is by its very nature difficult to systematise. In effect, the researcher is performing a cognitive task analogous to that which social scientists often ask of their subjects (see 2.1). That is, 'In what way(s) do the high and low points differ?' Also, 'What property/properties do these points share, which others (in the zero position) do not possess?' The answer—the term used to label the dimension—depends in part on the researcher's verbal or conceptual abilities or on the accessibility of Roget's *Thesaurus*. In factor analysis and MDS, the most frequently encountered set of reference axes are principal components (see 3.5.5).

While it is useful to find out what the direction of maximum variation in the configuration actually is (the first principal component), there is no reason to suppose that it will be substantively significant or meaningful. On the other hand it may direct the user's attention to a readily identifiable, significant, general variable or factor underlying the configuration—general intelligence, general occupational prestige, the overall size of specimens, the left-right political continuum, the evaluative factor in word connotations, tough-tender-mindedness in personality studies—all have been much-heralded primary dimensions of variation.

What of the other dimensions of variation? It depends principally upon whether the user wishes to keep the dimensions orthogonal to each other. If so, a number of techniques exists for finding a set of reference axes which will often lead to a more interpretable set of dimensions, by (rigidly) rotating the principal components to a new position where, for instance, the co-ordinates of the points on a given axis tend to either unity or zero* (a 35° counter-clockwise rotation of axes in Figure 1.1

*Kaiser's 'varimax' rotation criterion. See Rummel 1970, pp. 391–3 or Maxwell 1977, p. 54 et seq. for a simple introduction and exposition.

above comes close to satisfying this requirement). If the MDS analyst wishes to pursue the identification of 'best' axes, then the detailed technology developed in the factor analytic tradition (cf. Rummel 1970, chs. 14–21) may be used. On any account, the naming of dimensions on purely internal criteria can be a hazardous and subjective undertaking, at least for the basic MDS model.

4.3.2 *Graphical interpretation*

A configuration may be interpretable in differing ways in different regions and there is no guarantee whatever that one particular type of structure will best describe the entire configuration. But if we hope to detect different types of local structure we shall need a procedure that is sufficiently general to cover other methods as special cases. A technique that has been found to be very serviceable in detecting different types of local structure is graphic analysis (Kendall 1971a; Waern 1971).

The basic idea is to represent the highest similarity values in the data by drawing a line in the MDS configuration between each pair of objects involved. This can be done either by deciding upon a cut-off value (say the top quartile of data, or all similarity values greater than 0.70) or by ordering the data from the highest to the lowest similarity values. If the data are ordered by size, the link between the highest similarity pair is drawn in the configuration first, then the next highest pair is linked and so on, until the researcher has sufficient information about the local structure. In this way the growth of various types of structure becomes apparent as the researcher moves down the ordered data list. These structures may include:

(i) *chains*: successive links which form a path through the configuration. The path may be (approximately) linear, signalling a vector or dimension-like property, or non-linear, a parabola, circle or some other simple regular pattern, or even an irregular, zigzagging, but connected, sequence.

(ii) *clusters*: links which occur within particular regions and build up to form locally-connected subgraphs or clusters based upon a few points.

In many instances both types of structure may appear, producing a linked set of clusters and often a residual set of isolated points.

4.3.2.1 *Sequences and seriation*

The historical scientist is naturally interested in inferring or detecting the time-sequence of a set of objects or events. This process is usually referred to as 'seriation' or 'ordination' (see Renfrew 1976, ch. 2; Hodson 1971, pp. 173–290; Hubert 1974). A good example of the success of the graphical method in finding a sequence is reported in Kendall's paper (1971b) in which he analysed an 'abundance matrix' consisting of co-occurrence counts of various artefacts such as type of pottery and jewellery common to pairs of graves in the neolithic cemetery of the la Tène culture at Müssingen-Rain. By using graphical analysis as described above, he showed that a chaining emerged whose shape led to its being described as a 'horseshoe'. Kendall hypothesised that the sequence of the points along the horseshoe indicated the historical order of the burials. This hypothesis was subsequently confirmed in an independent study by Hodson.

Why does a continuum (such as, here, time) become distorted in this way? There

is no definite answer, but the evidence suggests that it is due in part to the fact that some permissible monotone transformations of the straight line generate a semi-circular MDS two-dimensional configuration (Shepard 1974, van de Geer, 1971, pp. 239–42). There is also increasing evidence (Kendall 1971a, p. 225) that it may be due to data collection procedures imposing an upper ceiling on values which the dissimilarities can take.

Thus, for example, in a rating exercise where only five or seven categories are allowed, the 'totally dissimilar' or 'totally unlike' category restricts the subject's ability to distinguish between very dissimilar and the very, very dissimilar. This, in turn, restricts the number of distinct values in the data matrix and the ability of the non-metric procedure to make the distance between points at opposite ends of the continuum as large as they 'should' be, thus producing the noticed horseshoe shape. This phenomenon can be overcome, i.e. the horseshoe sequence straightened out by treating the data as being only *locally* Euclidean, or by assuming that the largest distances are of least (or no) value in determining a solution, and relying only on proximate information. This may be done by using continuity mapping (5.2.2) or local monotonicity (5.2.1.1). The apparently simpler expedient of scaling in one dimension to recover such structures is usually misleading and is not recommended. This is partly due to the fact that highly irregular non-linear continua occur. As Shepard (1974, p. 386 et seq.) comments:

In analyses of many different sets of data that were known to be basically one-dimensional, I have found that two-dimensional solutions, when attempted, characteristically can assume either the simple C-shape or the inflected S-shape ... Evidently, by bending away from a one-dimensional straight line, the configuration is able to take advantage of the extra degrees of freedom provided by additional dimensions to achieve a better fit to the random fluctuations in similarity data.

Users should therefore be on their guard against using one-dimensional solutions to recover an ordering and should use graphic procedures to help detect strong non-linearities of this sort.

Perhaps the best known, and independently replicated, instance of a C-shaped continuum in a 2-space is of adjudged musical intervals: the intervals separating the points correspond very well to the number of intervening semitones when they are projected onto the horseshoe, but no linear dimensional interpretation is possible (Levelt et al. 1966; Shepard 1974, pp. 386–7).

4.3.2.2 *Simple structures*

Other simple graphical structures have been studied and identified in scaling solutions, either from empirical patterns in the data or, in the Guttman tradition of facet theory (see 4.5) from the structure of the original mapping sentence (Lingoes and Borg 1979; Maimon et al. 1980). Among the more commonly encountered structures, in Guttman's terminology, are:

the simplex (a chain, a linear or non-linear dimension or sequence);

the circumplex (a circular arrangement of points);

the radex (a combination of the simplex and the circumplex), consisting of two or more concentric circles with lines emanating from the centre, dividing the circles

into sectors and thus the sheep from the goats. In three dimensions a set of stacked radexes is termed a *cylindrex*.

A simple example of facet-theoretic interpretation occurs in Levy and Guttman (1975) and a simplified version of the relevant mapping sentence is presented in section 4.5. The two-dimensional MDS solution for fifteen of the items on a USA sample of subjects is presented in Figure 4.1 and includes its radex interpretation. In such an interpretation the configuration is divided up into a number of regions within which particular object points possess a particular characteristic, or facet. In this instance, where the radex defines a number of sectors, item 15, life in general (LG), functions as the centre of the circle, the facet state *vs* resources form the inner and outer circular bands and the facet 'area of life' divides the circle into eight sectors corresponding to those of the original mapping sentence.

A similar simple structure was found in a re-analysis (Coxon 1974) of the Bollen-Delbeke data (Delbeke 1968) on preferences for families consisting of different size and composition (of boys and girls). Using the basic (distance) MDS model, the two constituent dimensions (number of boys and number of girls) were identifiable as two somewhat distorted lines, reflecting the empirical fact that people tend to prefer mixed family compositions. Using the vector model (see 5.3.2 below) for analysis of the same data, a radex structure was apparent. In this case, the inner and outer circular bands represented mixed *vs* unmixed composition, and the sector lines divided points in terms of the overall size of the family. These data and their analyses are discussed at greater length later, in sections 6.2.2 and 6.2.3.

A fascinating example of a cylindrex pattern ('stacked' circles) occurs in Heider

INTERRELATIONSHIPS AMONG FIFTEEN VARIABLES* OF SATISFACTION WITH LIFE AREAS IN THE UNITED STATES

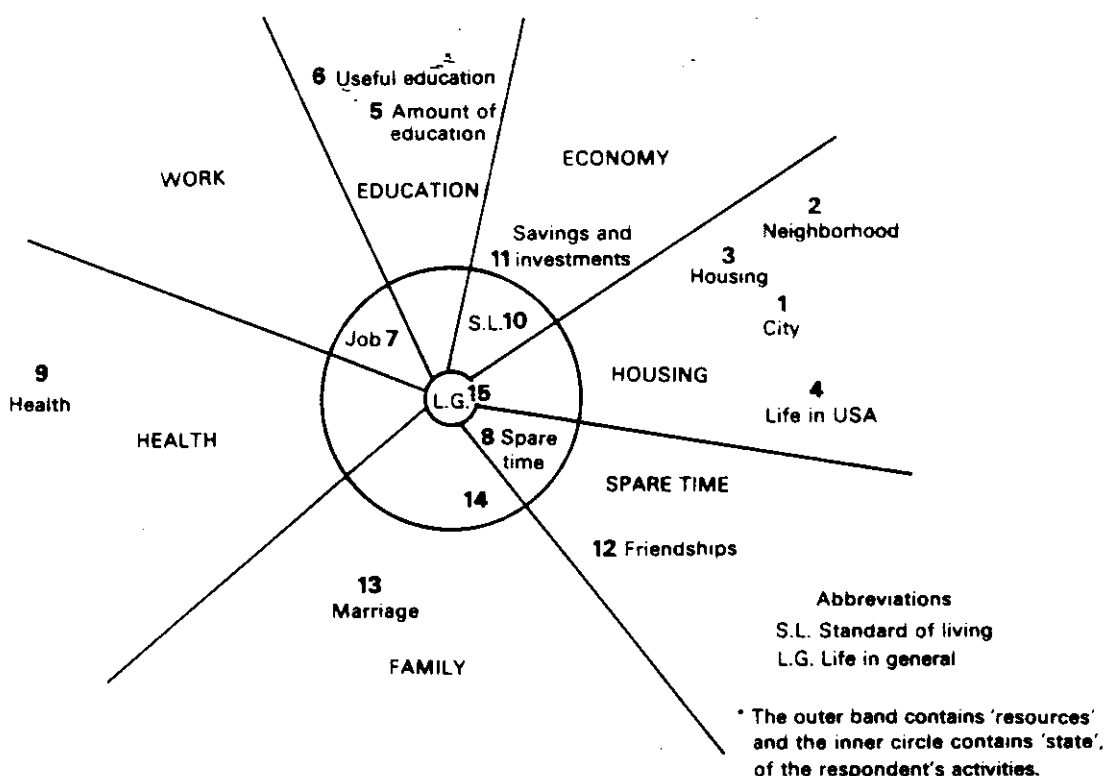


Figure 4.1 *Interrelationships among fifteen variables of satisfaction-with-life areas in the USA*

and Olivier's analysis (1972) of the structure of colour naming and memory in two markedly different languages. (The so-called Whorf hypothesis asserts that categories of naming used in different languages in some sense 'cause' different perceptions of reality, hence the interest of this study.) In this case, as in other studies of colour perception, the two basic dimensions of the space (brightness and hue) have in the MDS configuration become 'wrapped round' the brightness axis to form a circular pattern in 3-space, bringing together the red and purple ends of the hue dimension. The substantive interest in the Heider and Olivier study is to see whether the restricted categorisation of the Dani (a New Guinea people who use only two basic colour terms) affects their ability to recognise and identify different colour chips, compared to an English-speaking sample. Although the effects of restricted categorisation can be seen in the naming task configurations, it seems that retention of colour image in memory is unaffected by the very considerable cultural and semantic differences of the two languages. In both cases, colours of differing hue but the same brightness form a circular pattern, and the circles stack at different levels of brightness.

4.3.3 *Regions of high and low density of points*

Casual inspection and/or graphical analysis of an MDS configuration usually reveals that the points are not evenly distributed over the space. Rather, points tend to clump or cluster together, reflecting their high similarity, and are separated from other clusters by empty or sparsely-populated regions. That is, there exist subsets of points the relationships within which are stronger than those between them (Lingoes 1977, p. 116). Users may wish to check whether these clusters display any more formal structural properties.

A whole family of models, as extensive as those of MDS, exists for identifying and relating clusters of similar objects, namely cluster analysis (Wishart 1978; Everitt 1978). However, cluster analysis gives no information on the *extent* of separation of the clusters, and for this reason it is often advantageous to combine clustering analysis with MDS, which, of course, represents distances directly.* The best strategy is therefore to analyse the data *separately* by both a clustering and an MDS model and then represent the clusters *within* the MDS solution configuration. This is done in a 2-space by drawing a closed contour around the points contained in a cluster. The regions so enclosed will represent areas of high density, and the extent of their dissociation will be the distance in the configuration. This is done in a two-dimensional MDS plot: extension to three dimensions is usually not impossible, but poses problems in graphical portrayal.

Two varieties of clustering are extensively used in conjunction with MDS analysis—*hierarchical clustering schemes* (Johnson 1967, implemented as HICLUS in the MDS(X) series) and overlapping or *additive clustering* (Shepard and Arabie 1979, implemented as MAPCLUS; see section 8.2).

*Some data analysts recommend scaling followed by a clustering of the resulting *distances*. This practice is not recommended since it capitalises on the weaknesses of both methods. As we have seen (4.1), MDS solutions are least stable in their fine-grain local structure, on account of the existence of isotonic regions, but this is the very information from which clustering initially proceeds and which significantly determines the final clustering.

4.3.3.1 Hierarchical clustering schemes (HCS)

An hierarchical clustering scheme takes a matrix of dissimilarity measures between a set of objects and represents the objects as being gathered into clusters on the basis of this information. It describes not one clustering but rather (for p points) p different clusterings, referred to as *levels* of a single total hierarchical scheme. At the highest level, all the objects are contained in one cluster, at the next highest there are two and so on until, at the lowest level, there are as many clusters as there are points. The defining characteristic of a hierarchical scheme is that once a point is incorporated into a cluster at a lower level it may not 'leave' that cluster at a higher one. Thus the clusters form a hierarchical scheme in the sense that each level is a special case of the next highest. We now consider in some detail the method of hierarchical clustering.

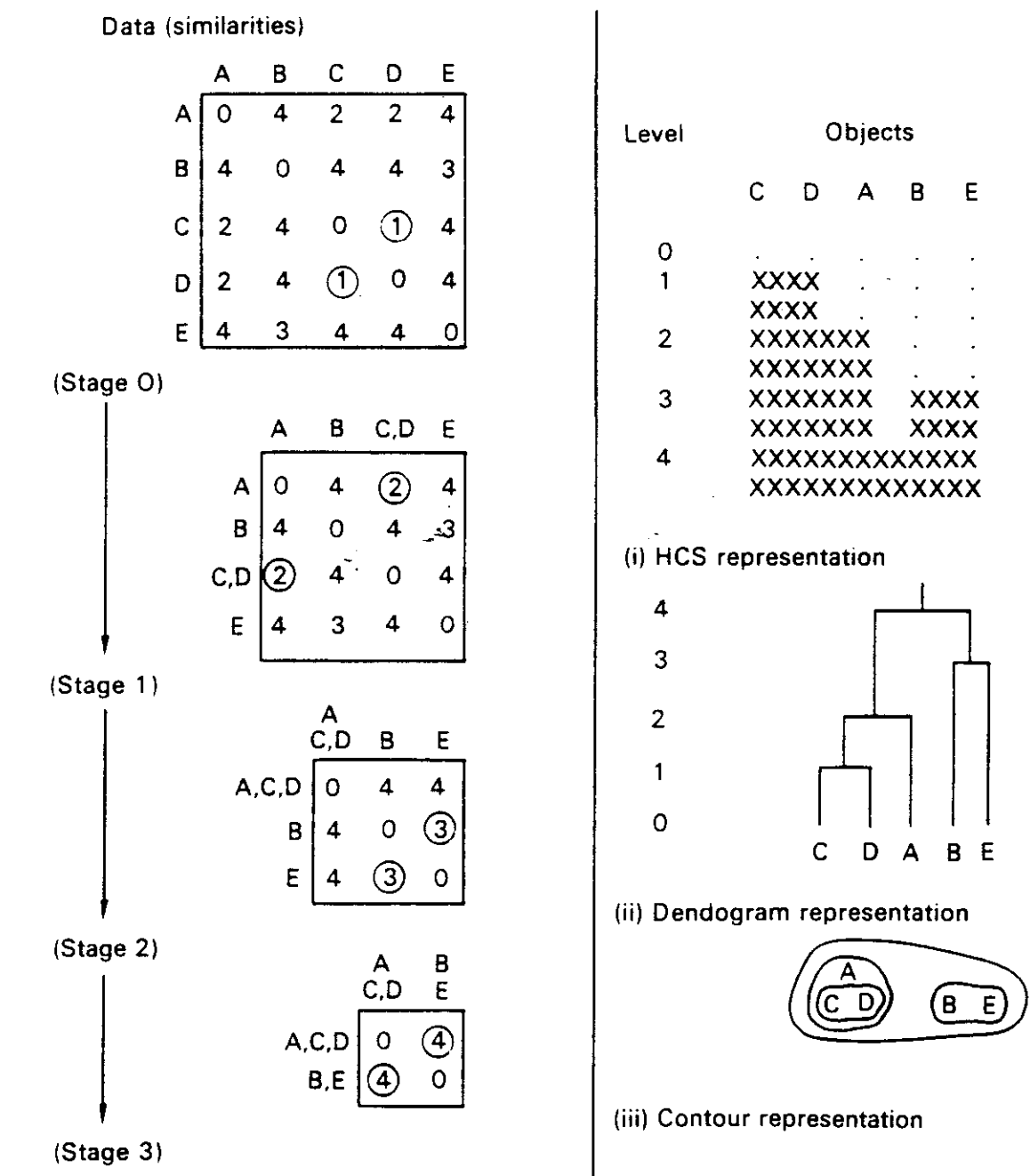


Figure 4.2 Illustrative example of the HCS procedure and forms of representation

The Method

Stage 0 The process of clustering begins by inspecting the original data matrix of dissimilarities and identifying first the most similar (or least dissimilar) pair of objects (C and D) and then merging them into a cluster. We now have the closest cluster of two points: (C, D) .

Stage 1 Points C and D are from now on treated as a single *object* and the data matrix is reduced by removing the row and column of C and D and substituting one representing the cluster (C, D) . In this example, the dissimilarity between C and each other object is the same as that between D and the same object—e.g. $\delta(C, A) = \delta(D, A) = 2$. (Normally this will not be the case.) The smallest entry in the reduced matrix, the currently most similar link, is now identified, and turns out to be between the cluster (C, D) and A . So the new structure consists of the dense cluster (A, C, D) and a set of unlinked points.

Stage 2 The new reduced matrix consists of the cluster (A, C, D) and points B and E . The smallest entry is now between B and E . This pair now form a new distinct cluster; the structure at this stage consists simply of the two clusters: (A, C, D) vs (B, E) . A new reduced matrix is formed.

Stage 3 In this final stage, the remaining two entities—the two clusters (A, C, D) and (B, E) —are merged, forming the final clustering of all the points.

The process of hierarchical clustering—forming clusters at decreasing levels of compactness—gives considerable insight into the regions of the space. In this case, we can see that the basic contrast is between the (C, D, A) and (B, E) cores of clustering.

As in other areas of data analysis, especially block modelling of social networks (see White et al. 1976 and Breiger et al. 1975) 'holes', the empty areas, frequently turn out to be quite as significant as the clusters, and both aspects have to be represented in any structural analysis. Empty regions represent two types of significant information: differentiation or dissociation between clusters on the one hand, and/or the significant absence of objects on the other hand, which might mean that certain stimuli have been neglected or overlooked in a study or that no objects actually exist which have a particular combination of attributes.

Clearly, the most significant *clustering* information is contained in the initial stages, and the most significant *dissociation* information is contained in the later stages of a clustering.

In the above example there was no ambiguity in defining the distance between a newly-formed cluster and existing objects (clusters or points), but this will not usually be the case. Consider the simplest case where we have a cluster formed of two points A and B and a third point C . There will be two distances, namely those between A and C , $\delta(A, C)$, and between B and C , $\delta(B, C)$; and we have to decide how we are going to use these to define the distance between (A, B) and C , that is $\delta((A, B), C)$. If we want the procedure to produce identical clustering schemes when the data are monotonically transformed we cannot take the obvious step of averaging $\delta(A, C)$ and $\delta(B, C)$. Johnson (1967) suggests two contrasting ways of defining this distance in this instance:

The *maximum* method (otherwise known as the diameter or complete link method) defines the distance $\delta((A, B), C)$ to be the *maximum* of $\delta(A, C)$ and $\delta(B, C)$.

The alternative *minimum* method (also known as the connectedness or single-link method) conversely defines the distance between the new cluster and the extraneous point to be the *minimum* of the distances between the extraneous point and each of the points in the cluster.

When the data satisfy the ultra-metric inequality (see 6.1.6) and are therefore perfectly representable as an HCS, the two methods produce identical hierarchical clusterings. Otherwise, the two HCSs will differ—often not markedly, but sometimes significantly.

The maximum (diameter) method picks out the largest distance within a cluster as 'the' distance and seeks to minimise the diameter (largest distance between the objects) within a cluster. This tends to produce a fairly small number of compact clusters.

The minimum method, by contrast, selects the smallest distance as 'the' distance and seeks to minimise the largest link needed to produce a chain or connected path between the objects. It tends to produce rather a large number of broken clusters and is often marked by chaining—the continued addition of a single element to a cluster.

In practice, the minimum method is usually to be preferred to the maximum method in exploring the hierarchical structure of a set of data (although both methods should be inspected to see how far the data may legitimately be represented this way.* The chief use of the HCS procedure is to examine not only relatively dense 'local' structure of highly proximate points (the lower levels of the clustering) but also the open or 'global' structure of spaces which separate or dissociate the clusters (the highest levels).

Hierarchical clustering then, possesses a number of useful characteristics:

- it presents not one, but a whole series of linked clusterings of increasing density, from a 'clustering' where each point is a separate cluster to the one where all the points are in a single cluster:

- it includes two commonly used types of clustering as special cases and therefore gives the user some idea of how well the data fit the assumptions of the clustering model:

- the HCS procedure is non-metric, in the sense that any ordinal rescaling of the data will produce identical results.

4.3.3.2 Clustering in high-dimensional space

The simple representation of HCS solutions within MDS configurations is only really feasible in two- or, at most, three-dimensional space. What if the user has a higher-dimensional solution and wants to gain some insight into the differential density of points in that space? An ingenious procedure is suggested by Andrews (1972; also see Everitt 1978, pp. 81–6) to represent each point as a wave form.

*Holman (1972) has shown that a set of errorless data will never perfectly satisfy both the Euclidean distance model and the hierarchical model, but will always satisfy one of the models to some extent.

Given a set of points in r -dimensional space, each point x is defined by its r coordinates: $\mathbf{x} = (x_1, x_2, x_3, \dots, x_r)$. It is then represented as a Fourier series function of the form: $f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \dots$, and the function is plotted for $-\pi < t < \pi$. The proximity of points can then be studied in terms of the similarity of the wave forms:

If some plotted functions form a band by remaining close together for all values of t then the corresponding points are close together in the Euclidean metric.
(Andrews 1972, p. 133)

Because the wave function preserves Euclidean distances the points which are close together have wave forms that have highly similar wave shape, whereas distant points have wave functions whose shape is different.

Because the wave function is affected by *all* the dimensions, the salient features of a high-dimensional configuration can be studied, and the procedure is therefore very useful for detecting isolates or outliers (which have markedly different wave forms) and clusterings (which have markedly similar wave forms).

4.4 External Methods of Interpretation

So far we have made use only of the original data in seeking to detect structure in the MDS solution.

When the researcher possesses additional external information about the points in the configuration, the task of interpretation is made much easier. These external variables, or 'properties' as they are often called, may come from a variety of sources. They may be relevant physical characteristics; they may be judgments made by respondents separately from the judgments used in the scaling; or they may simply represent the hunches or hypotheses of the researcher about the nature of the configuration. In any event, each property is assumed to consist of a set of numerical (interval or ordinal) values of the variable concerned, one for each point in the configuration.

A number of procedures exists for representing or 'embedding' each property within the already-obtained configuration, in a simple and easily recognisable manner. Two commonly used forms are as a *vector* and as a *point*. In essence a vector is a line drawn through the solution space and pointing the direction in which higher values of the property occur. Thus if our points were geographical sites, one relevant property might be 'northness', i.e. each point would have a value which was its north latitude. The property would then be fitted into the configuration map so that it was directed towards the pole. Such a representation would be adequate for points within areas of the northern hemisphere such as Scotland or the USA, but if the configuration actually contained the north pole, such as one consisting of sites in North America and Asia, in fact a map drawn from a vantage point above the pole, then the property of 'northness' would have to be represented not as a vector but as a *point* at the pole from which this property 'north' would decrease uniformly in all directions. Notice that representation in terms of vector or point depends on the characteristics not only of the property but also of the configuration.

These two forms of representing external information in a configuration are illustrated in Figure 4.3, with reference to the same 5-point configuration. Figure

4.3a represents a property of the points as a vector pointing in a north-easterly direction as the property values increase. Figure 4.3b represents another property in the same configuration of points, with the highest occurrences of the property somewhat left and below the origin of the space and systematically declining in all directions from this 'peak'.

These two ways of representing an external property within a configuration are now taken up in turn.

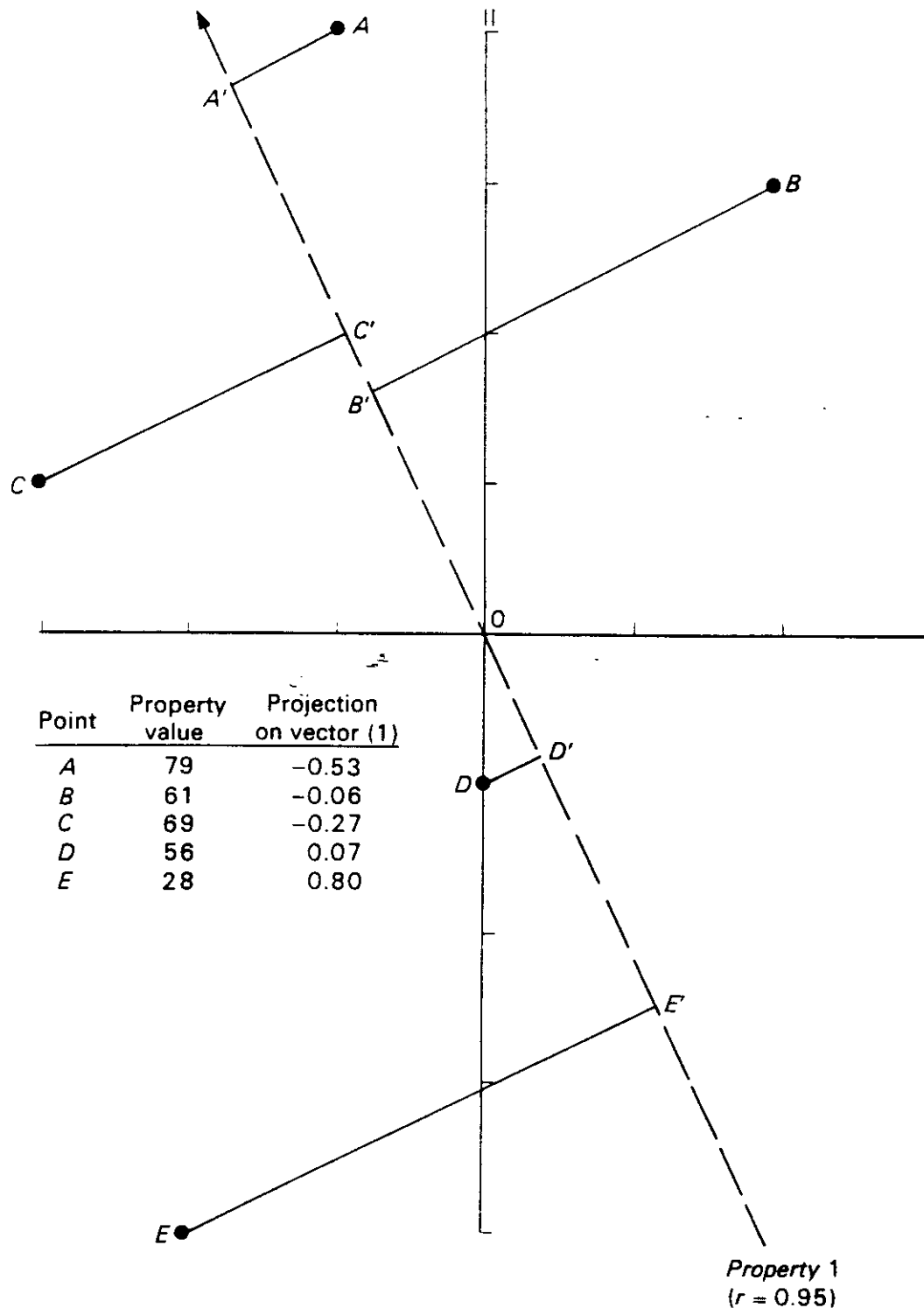
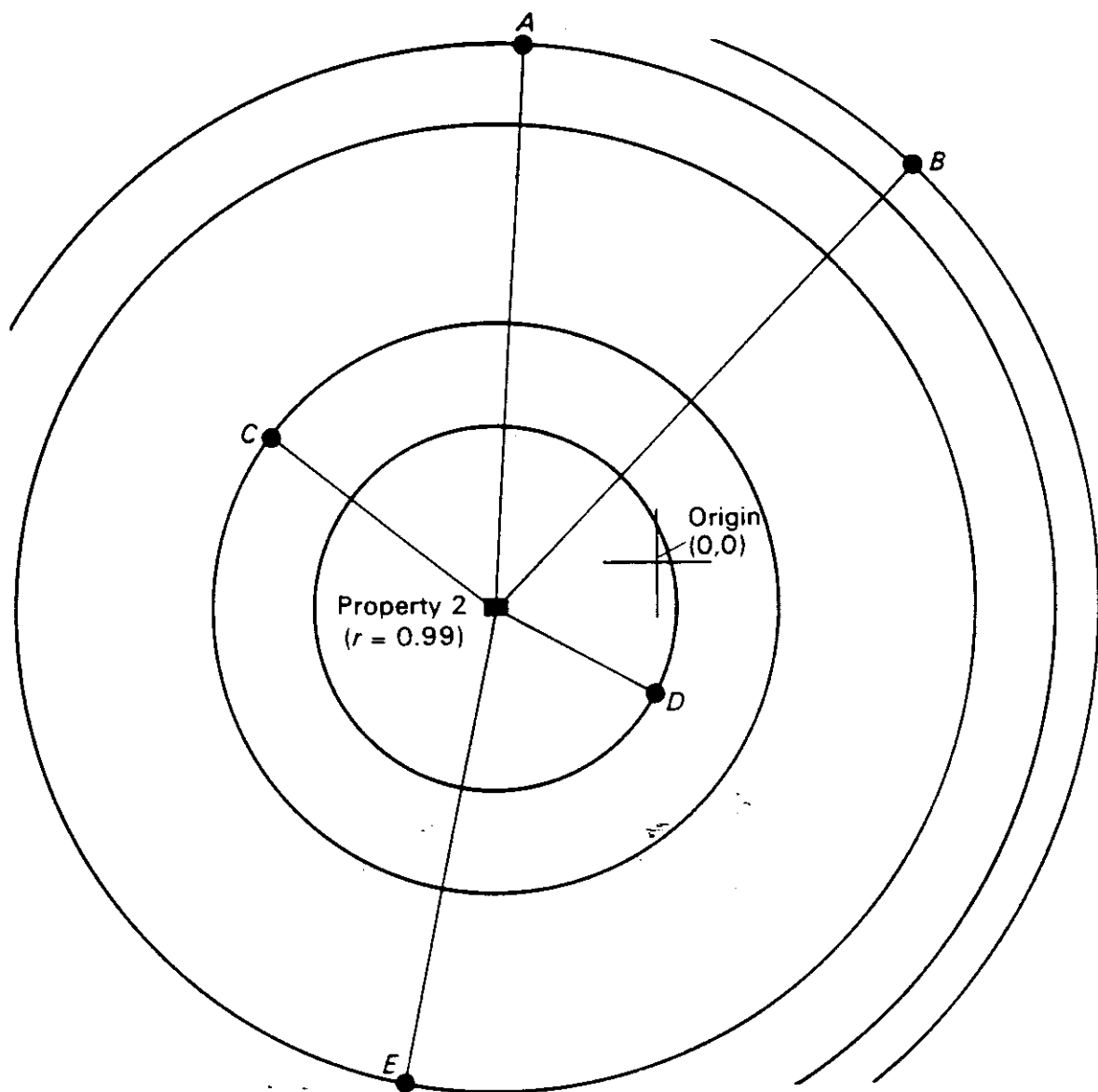


Figure 4.3a Representation of an external property as a vector



Point	Property value	Distance from point (2)
A	52	1.09
B	63	1.26
C	28	0.28
D	18	0.12
E	45	0.83

Figure 4.3b Representation of an external property as distances from a maximal point

4.4.1 Properties as vectors

Locating a single property in a space as a vector is similar to finding a dimension in the space, except that a vector is unipolar and not bipolar as is a dimension. A property consists of a value for each point in a configuration and the aim of the procedure is to position the vector in the space (like an axis) so that the projections onto the vector (like co-ordinates on the axis) correlate with the property values in some well-defined sense.* Such a procedure has several uses: it accurately locates a

*Clearly, the match between the original property values and the projections on the property vector should be as close as possible. Current options in MDS(X) include maximising *ordinal* fit (PREFMAP IV with monotone option), *linear* fit (PREFMAP IV with linear option, or PROFIT with linear regression) and *non-linear 'continuity'* fit (PROFIT with continuity option). See section 6.2.1.

direction in the space for each property; it gives the researcher some assurance as to how accurate are her hunches about properties of the space, and it provides a useful way of mapping additional information into the space.

When more than one property is fitted into the same space, attention is focussed additionally upon the relationship between the fitted vectors. The basic information is quite simply the angle separating them. Thus a right angle represents independence (zero correlation), in which case they are equivalent to a set of axes; 180° represents perfect *negative* association; and a 45° angle represents a linear correlation of 0.707 ($\cos 45^\circ$) and so forth.

Obviously, property-fitting can also be used to *identify* axes of a configuration by inspecting how close an external property comes to pointing in the same direction as the axis concerned. An example is presented in Carroll and Chang (1969, pp. 290–3), where they identify the three dimensions of a configuration obtained from scaling judgments of a set of tones as modulation frequency, modulation percentage and modulation waveform, to within 8° , 5° and 14° of the axes of the configuration.

An example

An instructive example of the uses to which vector property-fitting can be put is Rosenberg's studies of implicit personality theory (Rosenberg and Sedlak 1972a, 1972b). Each subject was given a set of 60 trait names (such as reserved, good-natured, submissive, humourless, etc.) and asked to sort them into groups, each of which was to represent a different person that they knew. A co-occurrence measure was used as a basis for constructing a dissimilarity measure between pairs of traits, and then scaled, yielding a two-dimensional solution, with stress₁ of 0.09. The subjects had also been asked to rate each of the 60 traits in terms of 5 general semantic differential scales (Osgood et al. 1965): hard/soft; good-intellectual/bad-intellectual; active/passive; good/bad; good-social/bad-social. The averaged ratings formed five properties, which were then fitted into the configuration as vectors (see Figure 4.4). Rosenberg and Sedlak's interpretation well exemplifies the use of property fitting and merits extended quotation:

Five properties are obviously not needed to interpret a two-dimensional space. Moreover, there are alternative pairs of properties, all with high R (linear correlation) values, which can be used to interpret this space.

It is possible, for example, to interpret the two-dimensional space in [the] figure [4.4] with the two general semantic differential factors, good-bad and hard-soft. Also, if we consider the three-dimensional solution, the R value for active-passive increases to 0.585 ($p < 0.001$), and the angle between the fitted axes for the three semantic differential factors are:

	good-bad	hard-soft
hard-soft	83°	
active-passive	92°	76°

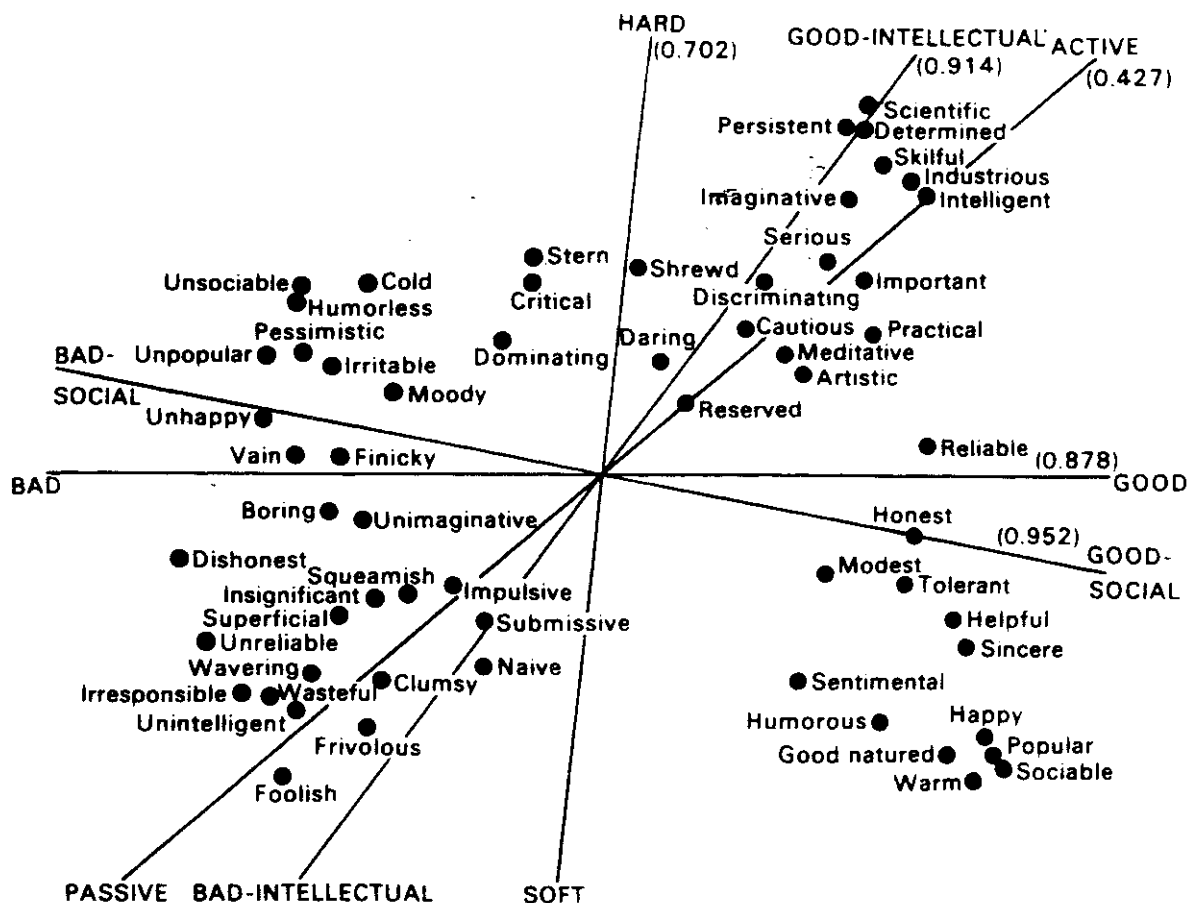
Thus, the results from the three-dimensional solution support the presence of the three semantic differential factors in personality perception with each factor more-or-less orthogonal angles ($\cong 90^\circ$) to the others.

An alternative interpretation, at least for the two-dimensional space in [the] figure [4.4] is the use of the two descriptive-evaluative properties, social good-bad and intellectual good-bad. It is interesting to relate this interpretation to Hays' (1958) findings that the extreme traits on one of his rank-order

dimensions were warm and cold (social good-bad?), and on the second rank-order dimension they were intelligent and stupid (intellectual good-bad?).

These two descriptive-evaluative properties are not orthogonal, however, either in the two-dimensional (65°) or the three-dimensional (67°) spaces. Nevertheless, while orthogonality is convenient, it is not a necessary feature of an acceptable interpretation of a space. Nor are the two sets of interpretations of the two-dimensional solution incompatible. The heuristic distinction between connotative and denotative meaning is relevant here. The interpretation based on hard-soft and good-bad might be thought of as concerned with connotative meaning, whereas the interpretation of the space in terms of social and intellectual desirability is concerned with denotative meaning. A multidimensional analysis of the basic denotative meanings of traits may not result in orthogonal co-ordinates. Indeed, intellectual traits are likely to be perceived by college students as relevant for social activities.

(Rosenberg and Sedlak 1972b, p. 253)



2 dimensional configuration of the 60 traits scaled by Rosenberg *et al.* showing the best-fitting axis for five properties. Each number in parentheses indicates the multiple correlation between projections on the best-fitting axis and property values. Reproduced by permission.

Figure 4.4 *Rosenberg's trait configuration interpreted by 5 property vectors*

When a property is represented as a vector the assumption is that the property is increasing across the space in the direction indicated by the vector and that the limit of the property is at infinity. Thus the lines joining points equidistant from the limit or ideal point are parallel to the vector. If the limit is brought in from infinity

to the boundary of the space, i.e. the property is increasing but with a finite bound, then the equidistant lines assume the familiar convex shape of the indifference curves—iso-preference contours—of micro-economic analysis. If the limit or ideal point is brought within the boundary of the space, i.e. becomes not only finite but accessible, then the iso-preference or iso-similarity lines become circles around the point. As in our example of the property 'north' above, such a representation may make assumptions not only about the characteristics of the property but rather about the characteristics of the space. Thus, whereas in the case of vector representation the vector is positioned so that the perpendicular projections (coordinates) onto the vector of the points matches the property values, in the case of the point representations it is the distances from the 'property point' to each of the points in the configuration which are matched to the property values.

4.4.2 *Properties as points*

If the user collects data which relate a new object to each of the existing ones, and then interprets the information as distances, it is straightforward to locate it as a point in the original configuration (the SSAM and PREFMAP III programs allow just such an option). A useful application occurs when the user wants to position some new points in a configuration that is either already known or where the information for some points is more reliable than for others.

An example which illustrates such use is Tobler and Wineberg's study (1971), based upon the co-occurrence of the names of a number of Bronze Age merchant colonies in Cappadocia on a set of some 800 cuneiform tablets. The co-occurrence frequencies were taken to be a function of the size of the colonies and of their geographical separation, and the data were scaled in two dimensions. Unfortunately—but hardly surprisingly—the location of the great majority of the 65 Bronze Age towns finally used in the analysis were not known. But had the location of a significant fraction been known, they could have been treated as a known, fixed, geographical framework. The co-occurrence information for the remaining towns could then have been treated as a set of external properties and located as points within the known configuration. In this way it would be possible, in principle, to identify the location of colonies which had subsequently disappeared. (This is also illustrated in Kendall's famous paper (1971b), 'Maps from marriages').

Probably the most common use of property-fitting as points occurs in preference studies. In this case, a subject's numerical evaluations of a set of objects are located as a point of maximum preference in a configuration which has been previously obtained from judgments of similarity of the same objects. (This use is discussed in 5.3.3.1).

In all these instances—adding new objects to a configuration, mapping subjects' preferences as 'ideal points' or representing an external property as a point as an aid to interpreting a configuration—the basic principle is the same. The additional information is viewed as giving a set of relative distances from the new point to all the existing ones. The task of the scaling program is then to position the point so that it best reproduces those distances (or their rank order). To date, little use has been made of this way of representing a property as a 'high point' in the configuration, but it has considerable advantages.

4.5 Facet Theory as a Framework of Interpretation

There exists a methodology which views the whole of the measurement process, up to and including the interpretation of scaling solutions, as an integral whole. This is Guttman's 'facet theory', Space does not permit a detailed exposition here, and the interested reader is referred to Borg (1977). The core of facet theory is the 'mapping sentence' which relates various aspects ('facets') of the population of subjects, or objects/stimuli, and of alternative questions, and maps them into the actual response made. A simplified example relating to a quality of life study might be as follows (cf. Borg 1977, p. 75). (The bracketed and labelled components are referred to as facets, and within each facet the alternatives are separated by a slash (/) symbol.)

A	B
The (cognitive/affective) assessment of the (state of/government's treatment for)	
	C
the well-being of the respondent's social reference group (self/government)	
	D
with respect to its (primary internal/primary social/primary resource/secondary state) environment concerning a general aspect of life area:	
	E
(family/on the whole/economic/...) according to the respondent's normative criterion for that life area	
IS MAPPED INTO THE RESPONSE-SET: (very satisfactory/.../very unsatisfactory)	

An actual questionnaire item can be thought of as being an instance of a particular combination of these five facets (A, B, C, D, E). Consider for instance, the following three questionnaire items:

I 'Generally speaking, are you happy these days?' can be viewed as combining the facets (a_2, b_1, c_1, d_1, e_2).^{*} This specification in brackets is often referred to as a 'structuple'.

II 'In general, how do you evaluate your family life?' combines (a_2, b_1, c_1, d_1, e_1).

III 'What is your opinion of the way the government handles the economic problems of the country?' combines (a_1, b_2, c_2, d_4, e_3).

By inspecting the item specifications it is clear that items I and II are inherently most similar, differing only on the E facet (on the whole *vs* economic), whilst item III differs from items I and II and A, B, C, D as well. Clearly, facet theory makes the researcher be more specific about the content of her data collection procedures, and this is itself a good thing. It also makes it possible to inspect the *a priori* or theoretical similarity and relationship of questions *before* obtaining the data. The facet theory approach also formulates a number of useful expectations or

^{*}Subscripts refer to the sequence number of the alternative within a facet.

hypotheses (or 'rules') about the data, e.g. 'The stimuli which are more similar in their facet structure will also be more similar empirically', which would lead us to expect their greater proximity in a distance model solution.

This abbreviated account is sufficient to show that the virtue of facet theory when it comes to interpretation of MDS solutions is that it alerts the researcher to characteristics to be looked for in the configuration, and to the type of structures (e.g. clusters of proximate items) to be expected. To this extent, facet theory can be used to assist the researcher to move beyond simple exploratory investigation towards a confirmatory approach.

In this chapter the simple structures to which Guttman and others have drawn attention have been discussed, but since they were originally developed within facet theory, interested users will profit from inspecting the full context (see Guttman 1971 and Lingoes 1977).

4.6 An Example of Interpretation: Occupational Similarities

An example drawn from one's own experience is most useful in conveying the detail and feel of the process of interpretation. I have therefore chosen work done in the Project on Occupational Cognition with Charles Jones, and concentrated on the interpretation of the basic configuration* of occupational similarities.

Data Pairwise judgments (using a 9-point rating scale) of the similarity of 16 occupational titles were obtained from 287 subjects. Both titles and subjects were selected from a fourfold typology of occupations chosen to contrast level of educational requirements and nature of the job—basically 'People' *vs* 'Data and Machines'.

Configuration Data were scaled in an aggregate (averaged) form, and in unaggregated form. The present example refers to the 2-D projection of the 3-D solution obtained from an Individual Differences Scaling (see Chapter 6) of a subset of 68 subjects' scalings. The basic configuration is that used as the basis for Figure 4.5.

The Resources available for interpreting the configuration were as follows:

- (a) *Researchers' 'internal' intuitions and hunches.*
- (b) *Data-based summaries for internal analysis*—in this case, the matrix of averaged similarities.†

These are presented in Table 4.1.

- (c) *Subjects' verbalisations.* As subjects completed their data ratings, they were encouraged to give the basis of their judgments, and these were either tape-recorded or written on the schedule. In each case, the comments were identifiable as referring to a particular pair of occupations.

- (d) *External information from subjects.* Subjects (including a number who had

*The substantive analysis is contained in Coxon and Jones 1978a, chapters 3 and 4 and the technical and methodological material is contained in chapters 2 to 4 of Coxon and Jones 1979. The data on which the analysis is based are available from the SSRC Survey Archive at the University of Essex.

†The original aggregate matrices (mean average, root mean square average and standard deviation of judgements) are presented in T3.8 in Coxon and Jones (1979).

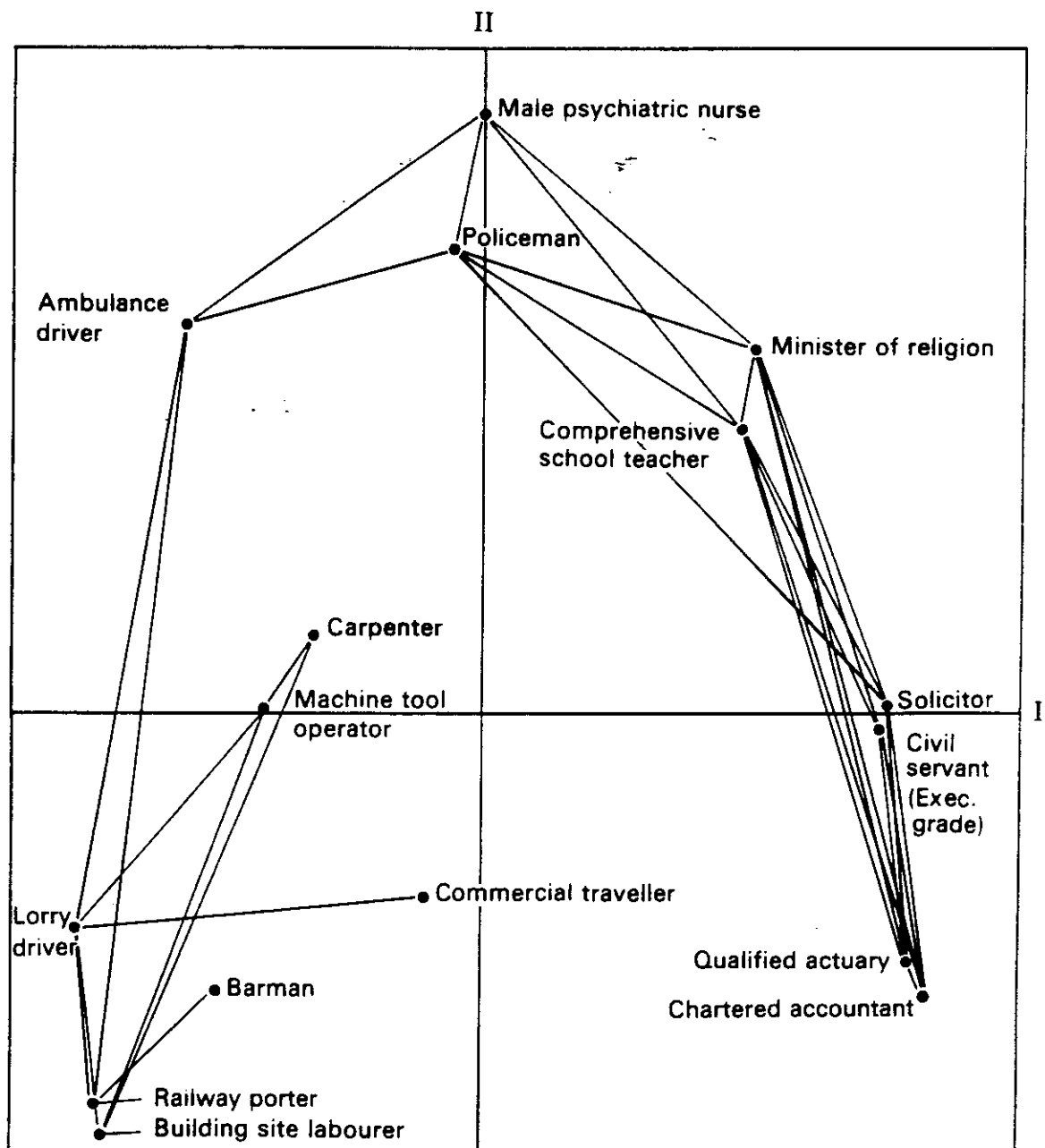
provided similarities data) were also asked to rate and rank the same occupations in terms of a set of characteristics including:

- 1 Social usefulness.
- 2 Prestige and rewards they ought to receive
- 3 Social standing (involving the standard sociological definition in terms of general standing in the community)
- 4 Monthly earnings (estimated income), together with a measure of
- 5 Cognitive distance. This last characteristic is not considered further here.

The process of interpretation

(i) *Dimension-naming*

The INDSCAL configuration is by definition already rotated to a non-arbitrary orientation (see 7.2.1), given by the east-west (Dimension I) and north-south



(Dimension II) directions on Figure 4.5. The occupational titles located at the extremes form the contrasts, and discontinuities are indicated by the gaps.

Contrast				
	Positive Pole	<i>vs</i>	Negative Pole	Discontinuity
Dim. I	(CA, QA, CSE, CS)		(BSL, RP, LD)	MPN and CST
Dim. II	(MPN, PM, AD, MIN, CST)		(BSL, RP, QA, CA, BM)	C and CST

The terms used by the *subjects* to describe the contrast involved in Dimension I were retrieved by looking at what they said about the pairs concerned, e.g. Accountant *vs* Labourer; Accountant *vs* Porter; Civil Servant *vs* Porter; Solicitor *vs* Lorry Driver. The concepts they employed included 'qualifications', 'skills required', 'education', and we, as researchers, decided that the common core to the descriptions made the tag 'educational qualifications' a reasonable one. Of course, other connotations such as status and income were also present and other labels could just as easily have been chosen. Dimension II was named as 'service orientation' by a similar process. Note that, in the instance, we used subjects' descriptions of the contrast as a fundamental resource, but decided upon the final identifying label ourselves.

(ii) *External properties*

In the study, we treated occupational judgment as involving the analytically separable components of cognition and evaluation, the former operationalised in terms of similarity judgments and the latter in terms of the first three characteristics listed earlier. The subjects' ratings were then averaged, thus providing four 'external properties'.

Both point-distance location and vector representations were used (by means of the PREFMAP III and IV models). Without any doubt the vector representation gave a better fit to the data for each property, and these are presented in Figure 4.6a. An interesting point is that the most explicitly evaluative property, 'social usefulness', is independent of (at right angles to) estimated earnings: on average, people judge the pay for an occupation to be unassociated with its social worth. Moreover, the 'good' end of the vectors—whether of social utility, status or earnings—all point towards the professional side of the configuration.

Turning now to the differential regional density of the points, it is worth noting that, if the largest single gap on each spanning dimension is marked as a dividing line, the four quadrants each include the two occupations chosen in the original design to be examples of a four-fold typology combining 'educational requirements' and 'people-orientation'. (These occupations are emphasized in Figure 4.6b.) In effect, the subjects view the occupations very similarly to the researchers, which is a useful bonus.

The diameter HCS clustering of the averaged similarities data is presented in contour form in Figure 4.6c, up to the final two levels. The pattern is clear with either HCS: the major divide is between the left and right hand sides the manual and the professional occupations, with the Barman and Commercial Traveller largely unassociated.

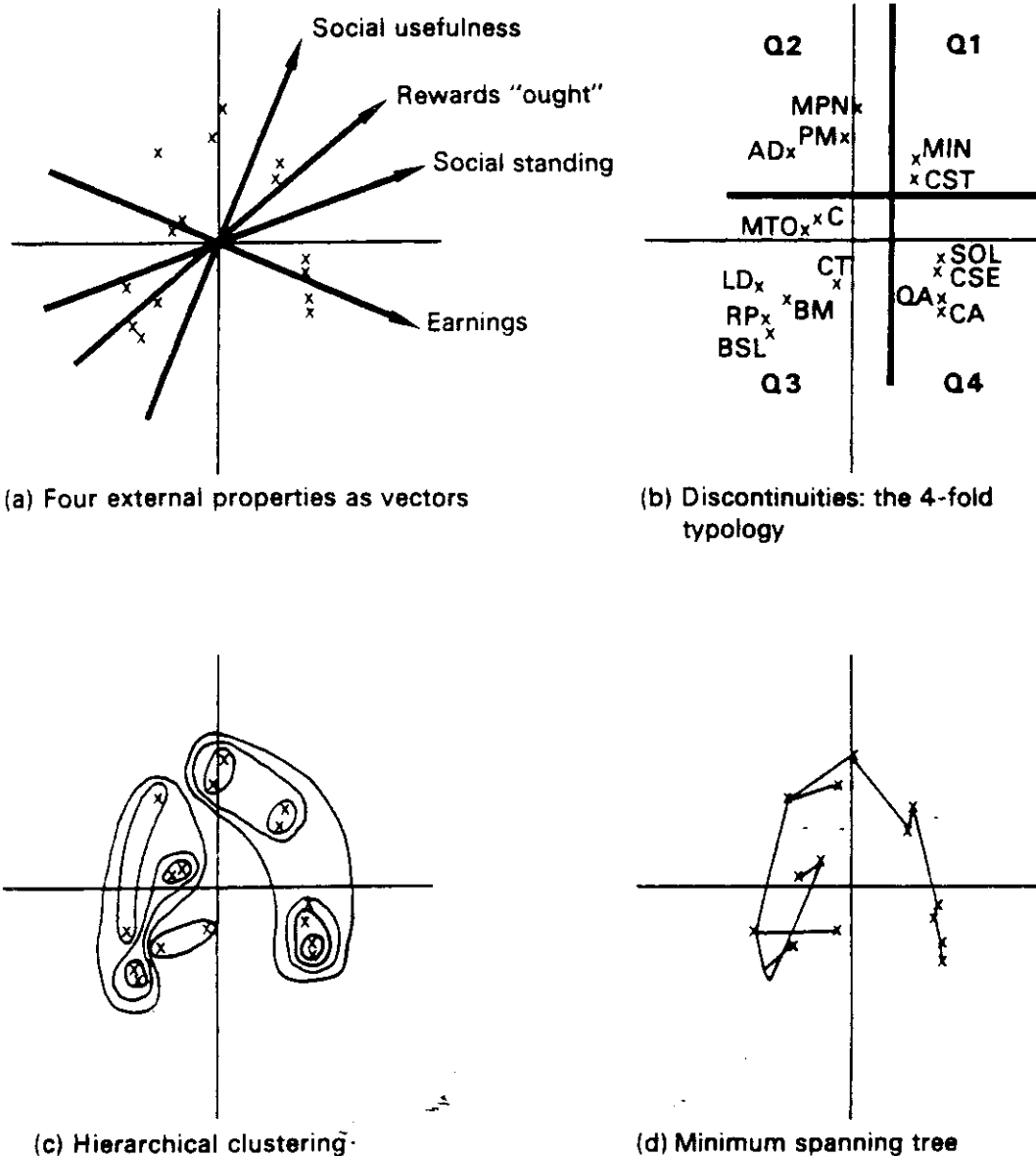


Figure 4.6 Various interpretations of single configuration

The dense clusterings occur round the two professional groups (people- and non-people oriented) and within the skilled and unskilled manual groups. It is very obvious that the clustering follows very closely the horseshoe sequence shown in Figure 4.5 and not any dimensional direction.

(iii) Graphic analysis

Finally, graphic analysis was used to detect local structure and it was at this point that the horseshoe pattern became very evident. When mapping the lines into the configuration in sequence from the highest similarity down, it became very obvious that the pattern was that of a linked set of clusters, rather than a simple sequence of occupations. In addition to mapping the top quartile of similarities, the minimum spanning tree (MST) was also constructed. The purpose of constructing a MST (Prim 1957) is to connect the points by a network (or tree) of *connected* links which have the smallest *overall* distance. The MST is illustrated in Figure 4.6d. Once again, the horseshoe sequence is very apparent as the first ten links join up the

points from the lower right-hand corner *in sequence* round to the lower left-hand corner.

(iv) *Final interpretation*

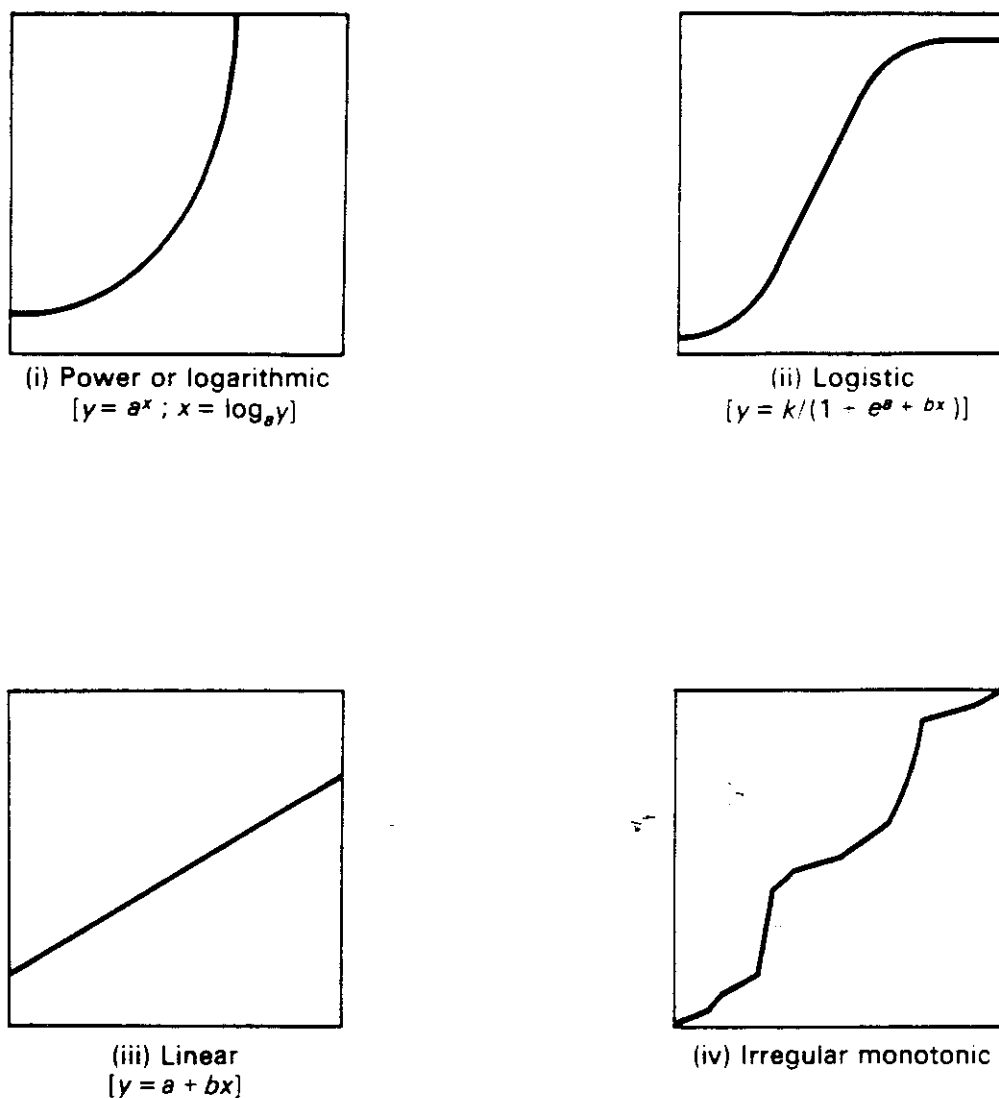
But what does the horseshoe mean? The clue to this came from looking at the HCS and matching clusters of points at each level with what the subjects said about each cluster. Thus Actuary and Accountant were regularly described as 'concerned with figures'; when Solicitor and Civil Servant are added as 'instrumental occupations' or 'managing', and when joined with the Minister and Teacher ('teaching') and the Policeman and Nurse ('custodial'), the entire group are repeatedly called 'professions'. It thus became clear that different descriptions (predicates) existed at different levels of generality. The range of generality of each of the most common predicates was obtained by finding out to which occupations a given predicate typically applied. This was looked at and mapped into the horseshoe sequence, with the length of the arrow indicating the range of generality (see Figure 3.6 in Coxon and Jones 1978a). It then became clear that:

The predicates repeatedly change as one moves along the 'horseshoe', making it very difficult to interpret the map as only involving a single contrast or dimension, even one as general as 'social status'. Yet there is a sequence, at least in the sense that predicates tend to develop and rarely appear outside their 'typical' range of applicability ... In moving from cluster to cluster along the sequence there are certainly many correspondences but equally, common features drop out and others re-appear.

(ibid, p. 92)

Thus, while there is a very clear sequence present in the configuration—a highly non-linear one, not to be confused with a dimension—it turns out not to be a sequence of continuous meaning. Rather, it is a family of resemblances where characteristics or properties apply at different segments so that the sequence is more like a chain of associations or what Wittgenstein (1958, section 60) describes as a 'family resemblances' theory of meaning.

However abbreviated, this example makes it clear that different methods of interpretation often provide convergent evidence about both local and global aspects of the structure or configurations which, with a little imagination, can be invaluable in aiding interpretation.

Figure 5.1 *Monotone curves*

concentrated in the largest distances and this fact needs to be taken into account in reading the map. The same principle applies in MDS solutions derived by use of this criterion.

There are two cases where it is useful to use the local monotonicity criterion, one being when ceiling effects occur in the data producing the familiar C-shape, or horseshoe shape, discussed in section 4.3.2.1. To overcome this effect, the remedy is fairly simple: ignore or down-grade the importance of the largest data dissimilarities. This can be implemented as follows:

- (i) by choosing the local monotonicity option in SSA(M); or
- (ii) by using a program such as PARAMAP which implements 'continuity' transformation, one of whose features is to act like a local monotonicity constraint (see 5.2.2 below).

Both options have similar and often dramatic effects—in 'unbending' highly non-linear simple structures.

The second use of local monotonicity is to map a high-dimensional solution into a space of lower dimensionality. This procedure is acceptable if local structure is of primary interest and larger distances which will be distorted can be ignored. It

should be noted, however, that Graef and Spence (1979) have shown that the largest distances do most work in producing an MDS solution and they can be critical in the satisfactory recovery of a configuration.

5.2.2 Continuity (smoothness) transformations

In basic non-metric scaling a best overall monotonic fit is sometimes achieved by producing sudden changes in distance values which do not exist in the data values. If we are firmly committed to the assumption that there really is no information in our data other than the order of the dissimilarities, well and good. But if we believe that the data contain more than simple ordinal information then these sudden discontinuous jumps may well distort the local structure, producing high distance values to correspond to very close data values. In this case, it might be better to concentrate on minimising or smoothing out the jumps by making the relationship between the data and the distances of the solution as 'smooth' or 'continuous' as possible, even at the cost of worsening the overall monotonic fit. This can be done by requiring that when two data values are close to each other, then there should be little difference (or variation) in the corresponding distance.

This basic idea of continuity can best be illustrated by a simple example of a one-dimensional solution. Suppose we wish to examine the relationship between the physical loudness of a set of six tones, x_1 to x_6 , and their perceived loudness, y_1 to y_6 . Our attention will concentrate, as we move up the scale, upon whether perceived differences in loudness change in the same manner as physical differences do.

If we say that the y values seem to change in a 'continuous' manner as we move along the underlying x continuum, we are essentially saying that the change in y as we move from one x value to the next tends to be small compared to the change in y generally associated with larger jumps in x .

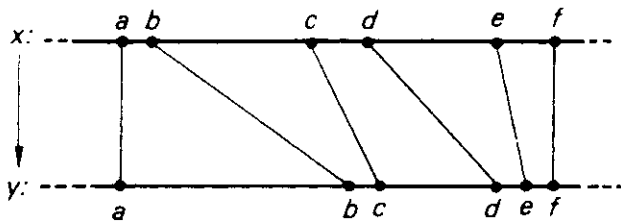
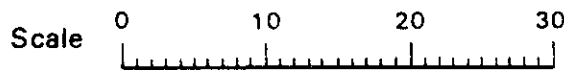
(Shepard and Carroll 1966, pp. 579-80)

In Figure 5.2 two examples of such a relationship are given: one where small changes in x are *not* associated with small changes in y , a relatively discontinuous relationship (Figure 5.2a), and another (Figure 5.2b) where small changes in x are associated with small changes in y , a relatively continuous relationship.

The extent to which the relationship between x and y is smooth or continuous can be monitored by a simple index which compares changes in y to changes in x , for each of the adjacent pairs along the scale. This can be done by taking the differences between adjacent values of both x and y and finding the ratio:

$$\frac{\text{Difference in } y}{\text{Difference in } x} = \frac{\Delta y}{\Delta x} = \frac{y_j - y_k}{x_j - x_k}$$

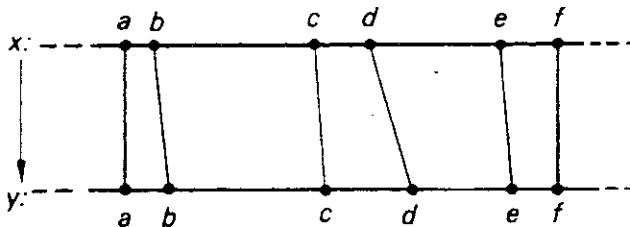
Clearly, when the relationship is smooth or continuous, Δy and Δx will be almost the same and the ratio will be about 1. But if small changes in x produce large changes in y , then the ratio will be correspondingly large. A simple overall measure of discontinuity (DISCONT) is constructed by squaring the ratio (both for computational simplicity and to draw attention to particular gross discontinuities) and then summing over the adjacent pairs:



(a) Discontinuous

Adjacent pair	Ratio of diffs: $\Delta y / \Delta x$	Squared ratio
a b	16 / 2	64.00
b c	2 / 11	0.03
c d	8 / 4	4.00
d e	2 / 9	0.05
e f	2 / 4	0.25

$$\text{DISCONT}(\text{Sum}) = 68.33$$



(b) Continuous

Adjacent pair	Ratio of diffs: $\Delta y / \Delta x$	Squared ratio
a b	3 / 2	2.25
b c	11 / 11	1.00
c d	6 / 4	2.25
d e	7 / 9	0.61
e f	3 / 4	0.56

$$\text{DISCONT}(\text{Sum}) = 6.67$$

*i.e. $(y_j - y_i) / (x_j - x_i)$

Figure 5.2 Discontinuous and continuous relations between two continua

$$\text{DISCONT} = \sum_{\substack{j,k \\ \text{adjacent}}} \left(\frac{\Delta y}{\Delta x} \right)^2 = \sum \left(\frac{y_j - y_k}{x_j - x_k} \right)^2$$

This measure is calculated in the boxes alongside the two examples in Figure 5.2. The relatively continuous relation (B) has a value close to 5 (the minimum value of DISCONT), and the discontinuous relation (A) has a value of 68.33. (Note that in this latter case the value of DISCONT is most affected when small differences in x are accompanied by large differences in y , e.g. for (a, b) and (c, d) . By contrast, where large changes in x give rise to small changes in y the contribution is very small, and this measure will be largely insensitive to them.)

5.2.2.1 Kappa as an index of continuity

The continuity transformation is used in MDS to obtain a solution where differences in the data correspond as smoothly as possible to differences in the solution differences. To do this we need the use of measure such as DISCONT, rather than stress. But in adapting DISCONT to measure the discontinuity between multidimensional spaces (rather than uni-dimensional continua) we run into a

problem. With a single line, the idea of a small change in value as we move up the continuum is easily defined: it is the difference between *adjacent* object locations. The notion of 'difference' generalises perfectly easily to 'distance' in the multidimensional case, but a little thought will convince you that there is no equivalent to 'adjacent' points in a two (and higher) dimensional space. But there is an approximation that will suffice: 'adjacency' can be replaced by 'closeness' or 'relative proximity' so long as we take care that only information relating to the immediate vicinity of each point is taken into account. In constructing an index of discontinuity in the multidimensional case, we shall therefore want to emphasise the distance involving closely proximate points and successively de-emphasise those at increasing distance. (This is obviously a further instance of *local* monotonicity described in the previous section.) In the context of MDS, the DISCONT measure is known as the 'kappa' index, symbolized by κ . The simplest measure on the analogy of stress, is referred to as 'raw kappa' and consists of two components, a discontinuity ratio and a weighting factor which restricts attention to the most proximate points:

$$\begin{aligned} \text{(raw) kappa} &= \frac{\text{discontinuity ratio}}{\text{ratio}} \times \frac{\text{local proximity}}{\text{weighting factor}} \\ \kappa &= \left(\frac{\Delta x_{jk}}{\Delta y_{jk}} \right)^2 \times w_{jk} \end{aligned}$$

Discontinuity ratio

In MDS applications we wish to ensure that small changes in the solution distances (d_{jk}) are associated with small changes in the data (δ_{jk}). Working with squared distances, as in DISCONT, the ratio becomes*:

$$\sum_{j \neq k} (\delta_{jk}^2 / d_{jk}^2)$$

Weighting factor

In the case of kappa, the weight factor is made the reciprocal of the corresponding squared solution distance:

$$w_{jk} = 1/d_{jk}^2$$

This form of weight has two useful properties: it ensures that local monotonicity is preserved (decreasing the contribution of any pair by the square of its distance, so proximate pairs contribute a good deal, and far distant ones scarcely at all), and the weights remain invariant under changes of scale.

Put together, these form the raw kappa index:

$$\text{Raw } \kappa = \sum_{j \neq k} \left(\frac{\delta_{jk}^2}{d_{jk}^2} \right) \left(\frac{1}{d_{jk}^2} \right)$$

or, in simplified form:

$$\boxed{\text{Raw } \kappa = \sum_{j \neq k} \left(\frac{\delta_{jk}^2}{d_{jk}^4} \right)}$$

*See Shepard and Carroll 1966, p. 581 et seq. In their treatment, data are referred to as (d_{jk}^2) and solution distances as (D_{jk}^2).

As in the case of raw stress, this index has the unfortunate property that an arbitrary enlargement of the solution configuration can make departure from continuity (raw kappa) as small as desired. And once again, the remedy is a normalising factor that will ensure that changes in the scale of the solution do not affect the index. Shepard and Carroll (1966, p. 583) show that the simplest effective normalising factor is:

$$NF = 1 / \left(\sum_{j \neq k} \frac{1}{d_{jk}^2} \right)^2$$

The normalised index of discontinuity (used in PARAMAP and non-linear PROFIT) then becomes:

$$\text{Normalised } \kappa = (\text{Raw } \kappa) / NF$$

$$\kappa = \sum_{j \neq k} \frac{\delta_{jk}^2}{d_{jk}^4} / \left[\sum_{j \neq k} \frac{1}{d_{jk}^2} \right]^2$$

These, and related measures, are further discussed in Appendix A5.1 and in the PROFIT and PARAMAP documentation of the MDS(X) series.

By minimising kappa, continuity scaling both preserves local structure and allows solutions to be forced down into very small dimensionality, so long as the user is prepared to disregard or downgrade large distances. The Shepard diagram resulting from continuity scaling has a characteristic fan-like form which reflects these properties. As the (solution) distances increase, the corresponding data values increase, which reflects the fact that any discrepancy in the representation of small distances is heavily penalised (i.e. local structure is being preserved), whereas even very large discrepancies in representing the largest distances are virtually ignored. Typical examples of the diagram occur in Shepard and Carroll (1966, p. 575) and in Coxon and Jones (1978b, p. 266), reproduced as Figure 5.3.

Continuity scaling is a hybrid transformation. In that it assumes that the data are a direct estimate of the solution distances (except for a possible scaling factor), so it implicitly assumes that the data are at the ratio level of measurement, and is therefore an instance of classic metric scaling (see section 5.2.3.2). But it also preserves local monotonicity, and to that extent continuity scaling can be viewed as an even weaker form of monotonicity than that assumed by non-metric scaling. However, the continuity criterion ensures that the characteristic 'steppiness' and 'angularity' of the monotone function are smoothed out.

6.2.1 External mapping by the vector model (PREFMAP IV and PROFIT)

Concisely: PREFMAP (PREFerence MAPping) (Phase IV) provides:

external analysis of two-way, row-conditional data
by a scalar products model.
using either a monotonic or a linear transformation of the data.

Concisely: PROFIT (PROperty FITting) provides:

external analysis of a configuration
by a set of property ratings or rankings in row-conditional format
by a scalar products model.
using either a linear or continuity transformation of the data.

Given

- (i) a fixed configuration of p stimulus points in a specific number of dimensions, and
- (ii) a rectangular data matrix of N rows (one for each subject or property) and p columns (identical to the configuration points),

these two programs map each of the N subjects/properties into the stimulus configuration as a vector, pointing in the direction in which the data values—preferences, property values, similarities—are increasing. For convenience, all the vectors pass through the centroid (centre of gravity) of the stimulus configuration. (Vector representation was discussed earlier in 4.4.1.)

What differentiates the two programs is simply the rescaling transformation option which can be requested:

(1) PREFMAP-IV

The rescaling transformation can either be *linear* or *monotonic* (strictly speaking, quasi-monotonic, meaning that the procedure first performs a linear transformation and then goes on to a set of iterations using monotone regression). If the monotonic option is chosen, three further options exist, depending on whether ties exist in a subject's row or data:

- (a) no ties exist (FIT(1)).
- (b) ties are treated as equalities (secondary approach, FIT(2)).
- (c) ties are treated as indeterminate (primary approach, FIT(3)).

(2) PROFIT

The rescaling transformation can be either *linear* or a *simple continuity* (kappa) function (see 5.2.2.1). In the former case, linear PROFIT is formally equivalent to PREFMAP-IV with linear fitting, although the algorithm used to produce the solution differs somewhat.

External mapping by a vector model is extensively used as a means of interpreting configurations and identifying dimensions; these uses were discussed in section 4.4.1 at some length.

External mapping is also often used to map subjects' preferences into an already obtained stimulus space. There is much to be said for analysing preference data in this way, since the evaluation of things may often be quite distinct from the cognition or adjudged similarity of things. We may agree entirely on the

characteristics and relative similarity of a set of regimes, people, foods, books, but disagree violently on their merits or over which we prefer.

A full discussion of such analyses and a number of examples of applications of this sort is provided in Carroll (1972, especially pp. 130–46).

External mapping is also particularly useful when the researcher wishes to represent information within an already known or physical configuration—such as a geographical map or the plan of a machine assembly. In this way, the directions can be estimated in which a particular plant species increases in a botanical area, or social deprivation increases in a city, or strains increase within a machine, or travel priorities lie in a country, can be readily assessed. It is also possible, using the other distance sub-models of PREFMAP, to see whether these data would be better represented as a point—an 'ideal point'—and if so, what sort of more complex distance might best fit the data.

The significant information which should be attended to in an external vector analysis is, first, how well a particular row of data (subject, property) fits in the configuration. (The goodness of fit depends, of course, on the transformation chosen: a linear relation is more restrictive than a monotonic one and therefore a vector of data is bound to fit better—or at least no worse—monotonically than linearly.) Given acceptable fit, the most important information conveyed by a vector is simply its direction (in none of these programs is the length of the vector relevant) and in comparing vectors, the angle of separation is crucial and represents the correlation between the subjects' data. Communality or concentration in a set of vectors, akin to the cluster in distance representations, will show as a tightly bound sheaf, and divergence or difference is signalled by empty sectors.

A brief illustration of the results of using PREFMAP-IV for the analysis of the occupational data described above is given in Figure 6.6. The ratings of (a) social usefulness and (b) earnings averaged within nine sub-groups of subjects were mapped into the configuration (see Figure 4.6) obtained from scaling the similarities data. (In each case, the analysis of variance in directions of the vectors gives a significant difference between the groups of less than one per cent: see Table 4.16 and 4.20 in Coxon and Jones 1978a.) Note, first, that the vectors are heavily concentrated, in each case, within a small sector. The average 'social usefulness' vector is located in a NNE direction (at 62°) and the variability in the groups' vectors is contained within 34° ; nine-tenths of the unit circle is empty. In this direction, the people-oriented, but low paid occupations ('vocations') are located, followed by less educated/lower paid and more educated/better paid through to the least socially useful. Notice the high degree of agreement between disparate groups—student teachers, engineers and clergy agreeing almost entirely (at least insofar as group averages go). By contrast, the average 'earnings' vector orders the occupations in an entirely independent manner (separated by just over 90° , representing a correlation of virtually zero) in a SE direction, but again with little spread, the extreme vectors subtending an angle of merely 30° . However, in this instance there is a detectable difference between the professionals' (more accurate) estimate and the working class groups' estimates, which tend to underestimate the professional occupations' remuneration compared to their own. It should be emphasised that present data are averages: the same conclusions do not necessarily hold when examining individual data.

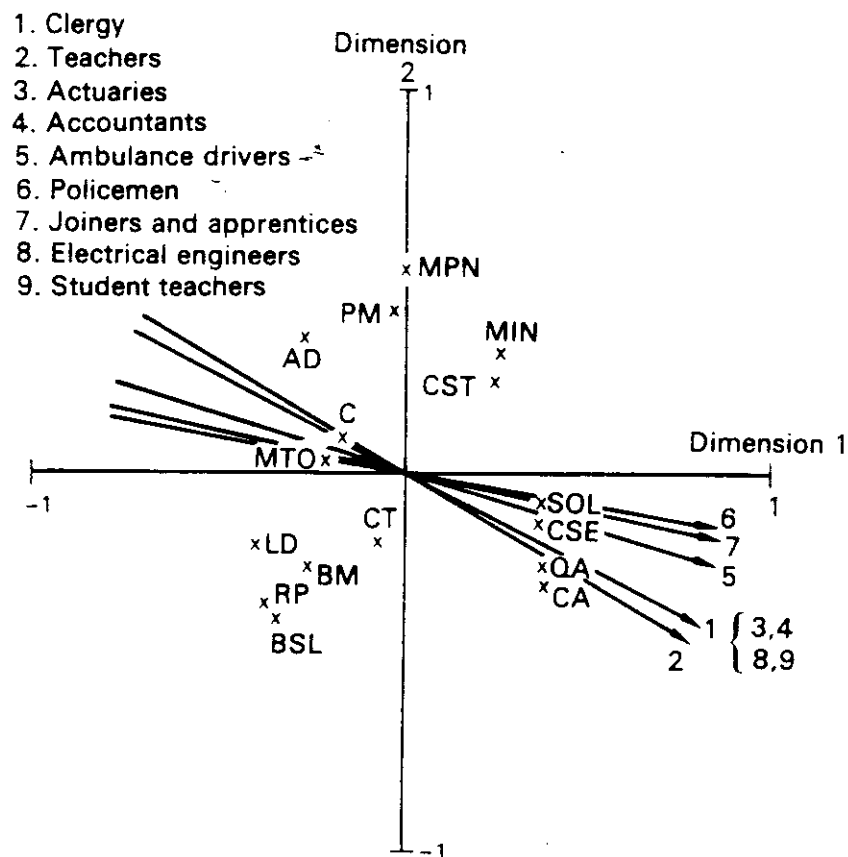
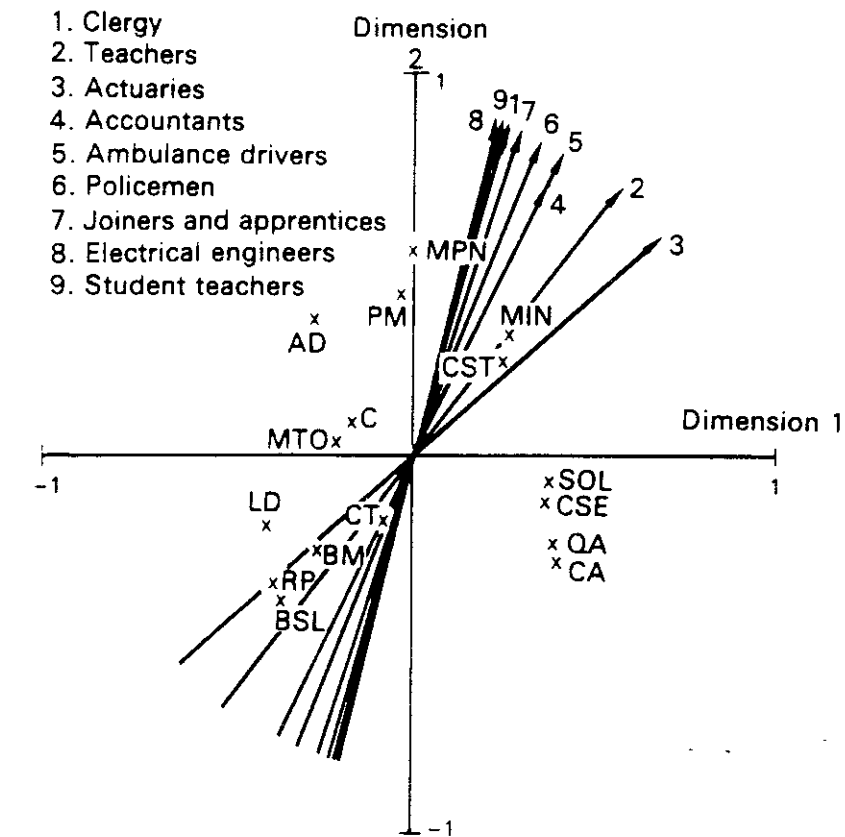


Figure 6.6 *PREFMAP-IV: external mapping of two criteria into INDSCAL configuration*

APPENDIX A5.1 KAPPA AND RELATED MEASURES OF DISCONTINUITY

All Shepard-Carroll (1966) kappa-based measures of continuity have the basic form:

$$\text{kappa} = \left[\left(\begin{array}{c} \text{smoothing} \\ \text{ratio} \\ \text{(i)} \end{array} \right) \times \left(\begin{array}{c} \text{local proximity} \\ \text{weight} \\ \text{(ii)} \end{array} \right) / \left(\begin{array}{c} \text{normalising} \\ \text{factor} \\ \text{(iii)} \end{array} \right) \right]$$

(i) *Smoothing*

The basic notion of a 'smooth transformation' consists in comparing two uni-dimensional continua, x and y , in terms of a mapping or transformation which ensures that the intervals or differences between adjacent points (i, j) in the one are of approximately the same size as those in the other, i.e. that the interval ($y_i - y_j$) is of about the same size (apart from differences in scale) as the interval ($x_i - x_j$). This is achieved by studying the *ratio* of the differences for each adjacent pair, squaring the result for purposes of convenience. Thus,

$$\left(\frac{\Delta y}{\Delta x} \right)^2 = \left(\frac{y_i - y_j}{x_i - x_j} \right)^2 \quad (1)$$

In the case of MDS, where the data 'distance' (y) are being compared to the solution distances (x), the differences in (1) become distances, and a simple overall measure of discontinuity or 'lack of smoothness' between the data and the solution is formed by summing the ratio over all $p(p - 1)/2$ pairwise data points:

$$\sum \sum_{i \neq j} \left(\frac{\delta_{ij}^2}{d_{ij}^2} \right) \quad (2)$$

(ii) *Local proximity weight*

Local proximity weights w_{ij} are the extension to the multidimensional case of the restriction to adjacent pairs (representing changes in value) in the uni-dimensional case. Shepard and Carroll (1966, p. 582) show that only weights having the form

$$w_{ij} = d_{ij}^s, \quad \text{with } s < 0$$

can ensure both that local monotonicity is enforced ($s < 0$), and that the ratio of any two weights remains invariant under change of scale, since solutions are unique only up to similarity transforms. As they indicate, simplicity and experience show that $s = -2$ is a sensible choice, yielding weights of the form: $(1/d_{ij}^2)$, which when multiplied into the discontinuity ratio (1), yields the basic measure of discontinuity.

$$\text{raw kappa} = \sum_{i \neq j} \frac{(\delta_{ij}^2)}{d_{ij}^4} \quad (3)$$

(iii) *Normalising factor*

Shepard and Carroll (1966, p. 582) define a normalising factor which ensures that kappa reaches a minimum when the solution distances d_{ij}^2 are proportional to the data distances δ_{ij}^2 except for a similarity transform. The simplest such factor is

$$\sum_{i \neq j} (1/d_{ij}^2)^{-2}. \quad (4)$$

The product of (3) and (4) yields the basic (normalised) kappa index. Gower (1979, p. 3) shows that this normalised kappa measure can be written in a particularly simple and interpretable form:

$$\text{normalised kappa} = \sum_{i \neq j} w_{ij} \left(\frac{1}{\delta_{ij}^2} - \frac{1}{d_{ij}^2} \right) \quad (5)$$

If the d_{ij} and δ_{ij} are of approximately the same order of magnitude, the weighting factor is approximately equal to $(1/d_{ij}^6)$ —which emphasizes fairly starkly how drastic a weighting function it is, giving long distances virtually no influence in determining the final configuration, and giving short (proximate) distances enormous weight. Even small differences in short distances will have very considerable effect on the size of the kappa measure: many users may prefer a less punitive weighting factor.

A5.1.1 Generalised forms of continuity index

Normalised kappa is a special case of the family of continuity indices referred to as 'kappa star'

$$\kappa^* = \sum_{i \neq j} \frac{(\delta_{ij}^2)^a}{(d_{ij}^2)^b} \left[\sum_{i \neq j} (d_{ij}^2)^c \right]^{-b/c} \quad (6)$$

(Normalised kappa is the case where $a = 1$, $b = 2$ and $c = -1$.) If the normalising factor is to keep the kappa index invariant under a similarity transform on the solution space, then the exponents must satisfy the condition $b + c - a = 0$, and c should be negative. (The exponent values can be varied within the PARAMAP

program, where the default values produce the normalised kappa index.)

In terms of the components of the index, a and b affect the continuity ratio, b (and, more indirectly, a) affects the strength of the local monotonicity weight, and b and c affect the normalising factor. Kruskal and Carroll (1969) have argued for $a = b = 1$, thus minimising the local monotonicity weighting and making all 'changes' and distances of equal importance in the minimisation process. They also make the case for reducing the size of the exponents, suggesting two further possibilities:

(i) $a = \frac{1}{2}$, $b = 1$, when the ratio takes the especially simple form of (δ_{ij}/d_{ij}^2) , which still preserves local monotonicity weighting, but not in a way that so severely reduces the effect of larger distances: and

(ii) $a = b = \frac{1}{2}$, which removes the local monotonicity weighting and concentrates the effect on the simple ratio of the two distances (δ_{ij}/d_{ij}) .

In general, it is necessary that $b > a$ if local monotonicity is to be maintained: the greater the inequality, the more severely discrepancies in representing local structure are penalised, and the less the balancing effect of more distant points.