A general theory of classificatory sorting strategies
1. Hierarchical systems

By G. N. Lance and W. T. Williams*

It is shown that the computational behaviour of a hierarchical sorting-strategy depends on three properties, which are established for five conventional strategies and four measures. The conventional strategies are shown to be simple variants of a single linear system defined by four parameters. A new strategy is defined, enabling continuous variation of intensity of grouping by variation in a single parameter. An Appendix provides specifications of computer programs embodying the new principles.

Introduction

Terms such as "computer classification" and "cluster analysis" have been used by divers authors to cover such a wide variety of fundamentally different numerical techniques that any attack on the general theory must specify the precise field under investigation. Classificatory programs in general can conveniently be considered as falling into four major groups, viz: (i) Methods involving simplification, usually by principal component analysis, followed by essentially subjective decisions; such are the methods of Tryon (1955) and of Mattson and Dammann (1965); (ii) Methods resulting in "overlapping classifications" such that a given element can appear in more than one group; examples are the work on clumps by Needham and his collaborators (Needham, 1962; Needham and Jones, 1964) and by Dale *et al.* (1964), the "agreement analysis" of McQuitty (1956) and the "concentration analysis" of Tharu and Williams (1966); (iii) Divisive methods such as association-analysis (Lance and Williams, 1965), dissimilarity analysis (Macnaughton-Smith *et al.*, 1964), and the methods of Rose (1964) and of Edwards and Cavalli-Sforza (1965); (iv) Agglomerative methods. It is solely this last group that we discuss in this paper.

The agglomerative strategies (which are always polythetic) can themselves be subdivided; and we shall adhere to the distinction made in our previous paper (Lance and Williams, 1966*a*): by *clustering* strategies we imply those that optimize some property of a group of elements; by *hierarchical* strategies those that optimize the route by which groups are obtained. Hierarchical strategies have attained a far higher degree of elaboration and sophistication than have clustering strategies; in this article we therefore propose to outline a general theory of hierarchical strategies, and shall defer consideration of clustering strategies until a subsequent communication.

All such methods involve two considerations. First, there is defined a measure of group-density or of inter-group likeness. Examples of the latter type of measure (the so-called "similarity coefficients") are legion; the best-known have been reviewed by Goodman and Kruskal (1954, 1959), Dagnelie (1960) and Sokal and Sneath (1963), but it is doubtful whether even these extensive collections are complete. For general consideration, suppose that two groups (i) and (j) fuse to form a group (k); then, extending (and slightly altering) the symbolism of Williams, Lambert and Lance (1966), we shall need to distinguish between three types of measure: (i)-measures, which define a property of a group, (i, j)-measures, which define a resemblance or difference between two groups, and $(i\ j,\ k)$-measures, which define some difference between the original two groups, considered jointly, and that formed by their fusion. Of these, (i)-measures are confined to clustering techniques except in so far as they may be incidentally required in the course of calculation of $(i\,j,\ k)$-measures.

Secondly, the chosen measure has to be incorporated into a "sorting strategy" whereby groups of elements are extracted. Selected sorting strategies have received some comparative study (Sokal and Michener, 1958; Sokal and Sneath, 1963; Williams and Dale, 1965; Williams, Lambert and Lance, 1966), but until recently have been regarded as separate and largely unrelated systems. In a recent brief communication (Lance and Williams, 1966*b*) we have pointed out that the five best-known hierarchical strategies are, for at least one of the (i, j)-measures, variants of a single linear system which will, moreover, generate an infinite set of new strategies. It is with the further generalization of this system that we are now concerned.

General properties

The properties of sorting strategies are not invariant under change of measure, and the measures to be considered must therefore be declared. We shall confine our attention to the four measures of our previous paper (Lance and Williams, 1966*a*); of these, three (Euclidean distance, the correlation coefficient and the "non-metric coefficient") are (i, j)-measures and can be considered jointly. The fourth, the information statistic

* *C.S.I.R.O. Computing Research Section, Canberra, A.C.T., Australia.*

in the form of information-gain (ΔI), is an $(i\,j, k)$-measure, and will be considered separately. We shall, for convenience in exposition, use a spatial model; but there would be no great difficulty in translating the concepts we shall use into a probabilistic context such as is used, for example, by Macnaughton-Smith (1965). We believe that the general properties of hierarchical strategies can usefully be regarded as of three types, which we consider in turn.

(*a*) *Combinatorial or non-combinatorial.* We assume two groups (i) and (j) with n_i and n_j elements respectively and with inter-group distance an (i, j)-measure denoted by d_{ij}. We further assume that d_{ij} is the smallest measure remaining in the system to be considered, so that (i) and (j) fuse to form a new group (k) with n_k $(=n_i + n_j)$ elements. Suppose the matrix with all d_{ij} as entries to be held by columns, with the n_i values as an additional row, and consider a third group (h). Before the fusion, the values of d_{hi}, d_{hj}, d_{ij}, n_i and n_j are all known and are all included in the (i) and (j) columns of the matrix. If d_{hk} can be calculated from these five values, than a (k) column can be derived from the original (i) and (j) columns; the computer need operate only on pairs of columns and, since all measures can be calculated from pre-existing measures, the original data need not be stored after the first set of measures has been calculated. Such a strategy we call *combinatorial*; our original linear example postulated the relation

$$d_{hk} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{hi} - d_{hj}| \qquad (1)$$

where the parameters α_i, α_j, β and γ determined the nature of the strategy. We also pointed out that when $\gamma = 0$ the string of measures associated with successive hierarchical fusions will be monotonic provided that

$$(\alpha_i + \alpha_j + \beta) \geqslant 1. \qquad (2)$$

In contrast to such a system, a *non-combinatorial* strategy is one in which the new measures cannot be calculated from the old, so that the data must be retained for the calculation of measures required later in the analysis. Combinatorial strategies have manifest computational advantages.

(*b*) *Compatible or incompatible.* A *compatible* strategy is one in which measures calculated later in the analysis are of exactly the same kind as the initial inter-element measures; they have the same dimensions (if any), are subject to the same constraints, and can be illustrated by an exactly comparable model. An *incompatible* strategy is one in which some at least of these properties are lost; the ensuing difficulties in interpretation render incompatible strategies undesirable.

(*c*) *Space-conserving or space-distorting.* The primary inter-element measures may be regarded as defining a space with known properties. When groups begin to form, it does not follow that the inter-group measures define a space with the original properties. If they do so, and the original model remains unchanged, we describe the strategy as *space-conserving*. However, with certain strategies the model will behave as though the space in the immediate vicinity of a group has been contracted or dilated; these are the *space-distorting* strategies. In a space-contracting system a group will appear, on formation, to move nearer to some or all the remaining elements; the chance that an individual element will add to a pre-existing group rather than act as the nucleus of a new group is increased, and the system is said to "chain" (for a measure of chaining, see Williams, Lambert and Lance, 1966). In a space-dilating system groups appear to recede on formation and growth; individual elements not yet in groups are now more likely to form nuclei of new groups. Such a strategy will group any data for which all d_{ij} are not identical. It is inherently likely to produce "non-conformist" groups of peripheral elements; an example is given in Watson, Williams and Lance (1966).

We now proceed to examine the major existing strategies with these considerations in mind.

Standard strategies

(*a*) *Nearest-neighbour.* This is the oldest of the conventional strategies. The distance between two groups is defined as the distance (normally an (i, j)-measure) between their closest elements, one in each group. It is combinatorial, in that it is only necessary to pick out the smaller measure on fusion; it is immediately derived from Eqn. 1 by the condition $(\alpha_i = \alpha_j = +\frac{1}{2};\ \beta = 0;\ \gamma = -\frac{1}{2})$. It is compatible under all (i, j)-measures, since all inter-group measures are to be found in the initial inter-element matrix. As a group grows it must appear to move closer to some elements and further from none; it is thus a space-contracting strategy, and its consequential chaining tendencies are notorious (*vide*, e.g., Williams, Lambert and Lance, 1966).

(*b*) *Furthest-neighbour.* This was suggested (*in litt.*) by P. Macnaughton-Smith for possible use when a relatively intense grouping strategy was needed. It is the exact antithesis of the foregoing, in that the distance between two groups is now defined as that between the most remote pair of elements, one in each group. It, too, is combinatorial and is derived from Eqn. 1 by the condition $(\alpha_i = \alpha_j = +\frac{1}{2};\ \beta = 0;\ \gamma = +\frac{1}{2})$; it is similarly always compatible. Since on growth a group will recede from some elements and move nearer to none, it is markedly space-dilating.

(*c*) *Centroid.* The earliest use we know of this strategy is that of Sokal and Michener (1958) under the name "weighted-group method". Algebraically, the group is considered as defined in Euclidean space and is replaced on formation by the co-ordinates of its centroid. Its combinatorial properties are not invariant under change of measure, and we now proceed to establish these for our three (i, j)-measures.

(*i*) *Squared Euclidean distance.* Since this is additive over attributes, we need consider only a single attribute x. Let the co-ordinates of the centroid of (i) be denoted by x_i. Then the centroid of (k) will be at $(n_i x_i + n_j x_j)/n_k$

and, by definition,

$$d_{hk} = [x_h - (n_i x_i + n_j x_j)/n_k]^2.$$

By multiplying up and rearranging it is easily shown that the right-hand expression is identically equal to

$$\frac{n_i}{n_k}(x_h - x_i)^2 + \frac{n_j}{n_k}(x_h - x_j)^2 - \frac{n_i}{n_k}\cdot\frac{n_j}{n_k}(x_i - x_j)^2$$

$$= \frac{n_i}{n_k}\cdot d_{hi} + \frac{n_j}{n_k}\cdot d_{hj} - \frac{n_i}{n_k}\cdot\frac{n_j}{n_k}\cdot d_{ij}.$$

The strategy for this measure is thus obtained from Eqn. 1 when

$$\alpha_i = n_i/n_k,\ \alpha_j = n_j/n_k,\ \beta = -\alpha_i\alpha_j \text{ and } \gamma = 0. \quad (3)$$

(*ii*) *Correlation coefficient.* The problem of correlating sums of elements seems first to have been posed and solved by Spearman (1913); but his familiar solution requires access to the complete set of initial measures. The measure can be made combinatorial by storing the appropriate covariances and variances in place of correlation coefficients and group-sizes. We write cov_{ij} for the covariance of (i) and (j), and v_i for the variance of (i); the correlation coefficient is constructed when required from the definition $r_{ij} = \text{cov}_{ij}/(v_i v_j)^{1/2}$. From the definitions of the quantities concerned, it is easily shown that

$$\text{cov}_{hk} = \text{cov}_{hi} + \text{cov}_{hj}$$

$$v_k = v_i + v_j + 2\,\text{cov}_{ij}.$$

These two equations will serve to define a combinatorial solution, although this cannot be based on the correlation coefficients themselves, and cannot be derived from Eqn. 1.

(*iii*) *Non-metric distance.* A combinatorial solution for this measure would require a solution of the following problem: given real positive quantities a, b and c, and given the values of $|a - b|$ and $|a - c|$, to derive the value of $|(a - b) + (a - c)|$. This is obviously impossible; there is therefore no combinatorial centroid solution for this measure, and the original data must be held in store.

The centroid strategy is compatible for all coefficients and is space-conserving. The ensuing simplicity of the overall model has not unnaturally tended to endear the system to users, but the strategy is not without inherent disadvantages. In particular, the monotonicity requirement of Eqn. 2 is not met, and reversals, particularly with some measures, can be extremely troublesome (Williams, Lambert and Lance, 1966).

(*d*) *Median.* A further disadvantage of centroid is that, if n_i and n_j are very disparate, the centroid of (k) will lie close to that of the larger group, and remain within that group; the characteristic properties of the smaller group are virtually lost. (It is for this reason that Sokal and Michener, 1958, described it as a "weighted" strategy.) The strategy can be made independent of group size by arbitrarily putting $n_i = n_j$; the apparent position of (k) will now always lie between (i) and (j), and the parameters of Eqn. 3 reduce to $\alpha_i = \alpha_j = +\frac{1}{2}$; $\beta = -\frac{1}{4}$; $\gamma = 0$. In the Euclidean model, the new group is sited at the mid-point of the shortest side of the triangle defined by (i), (j), (h); d_{hk} lies along the median of this triangle, and it is for this reason that Gower (1966), who first suggested this strategy, proposed the name "median", and derived its properties from the theorem of Apollonius. It is available as an optional strategy in the mixed-data program CLASP on the Rothamsted Orion computer.

It is combinatorial by definition, and fully compatible for squared Euclidean distance, which was the case for which Gower defined it. Since the non-metric measure can be regarded (Williams and Dale, 1965) as a distance in a non-Euclidean space, this too may be treated as if it were compatible. Although the correlation coefficient could be manipulated (in the form of $(1 - r_{ij})$) in the system, we are unable to assign to it any useful geometrical meaning and we think the strategy should be regarded as incompatible for this measure. The system is space-conserving, though the apparent position of a group may swing widely. The condition of Eqn. 2 is not met and hence the strategy is liable to failure of monotonicity.

(*e*) *Group-average.* In our preliminary note (Lance and Williams 1966*b*) we ascribed this method to Sokal and Michener (1958). Re-examination of their work shows, however, that they used it only for element/group, and not for group/group, relationships; nevertheless, the system has interesting properties which will repay investigation. It is combinatorial over all coefficients; for, if s_{hi} represents a single inter-element measure between (h) and (i), we have by definition:

$$d_{hk} = \frac{1}{n_h n_k}\cdot\sum_{h,k} s_{hk}$$

$$= \frac{n_i}{n_k}\cdot\frac{1}{n_h n_i}\cdot\sum_{h,i} s_{hi} + \frac{n_j}{n_k}\cdot\frac{1}{n_h n_j}\cdot\sum_{h,j} s_{hj}$$

$$= \frac{n_i}{n_k}\cdot d_{hi} + \frac{n_j}{n_k}\cdot d_{hj}.$$

The system is therefore obtained from Eqn. 1 when $\alpha_i = n_i/n_k$; $\alpha_j = n_j/n_k$ and $\beta = \gamma = 0$. It is fully compatible, providing the concept of an average measure is acceptable. The concept of an average correlation coefficient is not entirely happy, and a more satisfactory solution might be provided for this case by putting

$$d_{ij} = \cos\left[\frac{1}{n_i n_j}\cdot\sum_{i,j}\cos^{-1} s_{ij}\right]$$

but we are not aware that this form has been used. The system is less rigorously space-conserving than is centroid but, since it has no marked tendencies to contraction or dilation, it may be regarded as a conserving strategy. Since $\alpha_i + \alpha_j + \beta = 1$, Eqn. 2 is satisfied and the resulting tree is necessarily monotonic. This strategy has not received the attention it deserves.

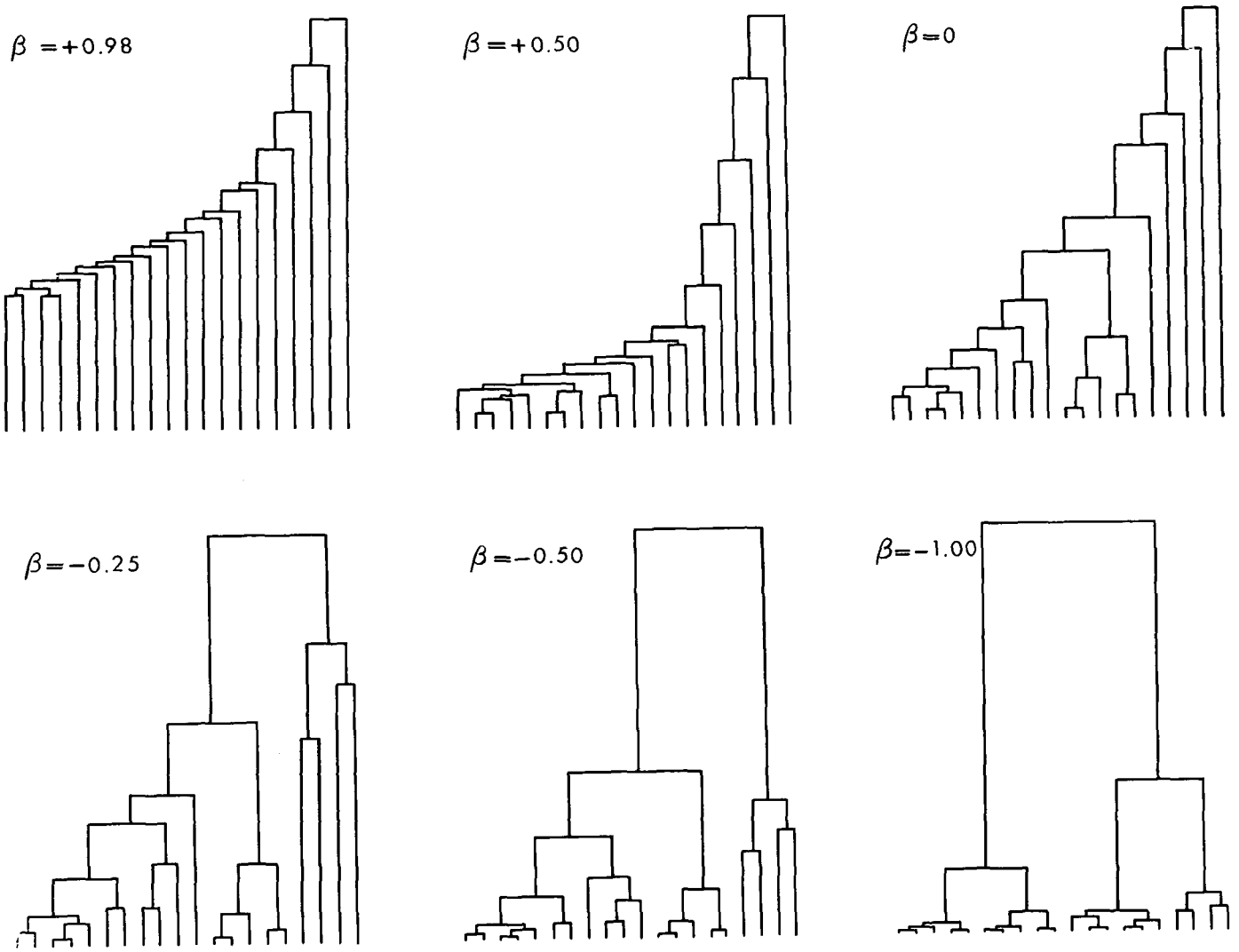

Fig. 1.—Effect of varying β in flexible sorting strategy. Data: 20 elements specified by 76 binary attributes. Measure: squared Euclidean distance

Flexible strategy

By this we mean the system derived from Eqn. 1 by the quadruple constraint ($\alpha_i + \alpha_j + \beta = 1$; $\alpha_i = \alpha_j$; $\beta < 1$; $\gamma = 0$); it is combinatorial by definition. It is compatible for Euclidean distance though the strictly Euclidean property is lost; the 1/0 constraint may fail with the non-metric measure, but with so arbitrary a quantity we believe this to be of no importance. It is meaningless if applied to the correlation coefficient, and for this measure it is completely incompatible. Its flexibility lies in its space-distorting properties. As β approaches unity, it is increasingly probable that, after only a single fusion, the apparent distance from this first group to the nearest element will always be less than any remaining element/element distance. The system, in fact, becomes increasingly space-contracting, and, apart only from initial ambiguities, can be made to chain completely by taking β sufficiently close to unity. As β falls to zero and then becomes negative, the system ceases to contract and becomes increasingly space-dilating, and the elements correspondingly more intensely-grouped. In **Fig. 1** we show the 20-element data of our previous paper (Lance and Williams, 1966*a*) processed with six values of β. We have been unable to define rigorously the value of β for which the strategy could be regarded as space-conserving; comparison with the known conserving strategies suggests that this would correspond to a small negative value of β, but the exact point probably depends on the values of the measures under examination.

The information-statistic case

Information-gain is an $(i\,j, k)$-measure, being derived from the information contents before and after fusion

by the simple relationship

$$\Delta I_{(i,j,k)} = I_k - I_i - I_j.$$

If the fusion is between two individual elements, this collapses, in the conventional (a, b, c, d) notation for a 2×2 contingency table, into $2(b + c) \log 2$—i.e., into a multiple of squared Euclidean distance—and is indistinguishable from an (i, j)-measure. We can then distinguish between three situations: (i) Strategies in which the quantities manipulated never rise above the element/element level — nearest-neighbour, furthest-neighbour, group-average. These are combinatorial and compatible for the information statistic, but its use would be pointless; the answers would simply be those for Euclidean distance obtained by an unnecessarily cumbersome computation. (ii) Median and flexible sorting. The solution here would be incompatible; element/group and group/group values would remain (i, j)-measures, and would no longer represent information gains. (iii) Centroid sorting. The solution currently used in information analysis is the non-combinatorial centroid strategy, though the centroid proper is never formed; the definition of the entire group under study is used. This is the only compatible strategy available for this measure. The measure is traditionally based on a probabilistic model; a spatial model would be quasi-metric, in that the triangle inequality fails. It behaves, however, as a space-dilating strategy of considerable power.

Conclusions and recommendations

Computational considerations

To encompass the complete set of six strategies and four measures here discussed, two separate computer programs are needed, which we now specify. *Program I* would be based on the combinatorial strategy of Eqn. 1; all six strategies are then available by specifying α_i, α_j, β and γ on entry. It would be confined to the (i, j)-measures Euclidean distance, non-metric distance and correlation coefficient. Some restrictions on measures would still be required: the correlation coefficient must be excluded from median and flexible, and both the correlation coefficient and the non-metric measure must be excluded from centroid (the non-metric measure because no combinatorial solution exists, the correlation coefficient because it would involve the complication of providing an alternative to Eqn. 1). *Program II* would use the non-combinatorial centroid solution of our previous paper (Lance and Williams, 1966*a*), and would incorporate the information statistic (for binary data only), and the correlation coefficient and non-metric measure (for binary or continuous data). This is substantially our original program CENTROID with the Euclidean distance facilities removed. Both programs now exist on the Control Data 3600 computer at Canberra; a specification for Program I and the revised specification for Program II are given in the Appendix.

There remains the question of the mixed-data classificatory program MULTIST, which is based on the non-metric measure. The existing form uses the non-combinatory centroid solution, and is therefore rigorously compatible and space-conserving. It seems to us likely that, in the case of data which group poorly, a user might well be prepared to relax the strict compatibility of this solution in order to obtain the increased grouping potentialities of the flexible strategy. A version of MULTIST using this strategy has therefore been prepared, and has proved very successful in biological applications.

User considerations

A user may reasonably expect to receive advice as to the measure and strategy most appropriate to his particular problem. Since the strategy may be restricted by the measure, the latter should be chosen first. The only fully-developed mixed-data program at our disposal uses the non-metric measure; the use of this measure is therefore inevitable if the mixed nature of the data is to be preserved. If the user is interested in interrelationships between attributes rather than their absolute values—if, in taxonometric parlance, he is interested in "shape" rather than "size"—the correlation coefficient is appropriate. If the data is binary and a probabilistic solution is desired, the information statistic should be used. If none of these restrictions applies, there are obvious advantages in Euclidean distance, because of its compatibility over all strategies. In the past we have not advised its use, since with the centroid solution it groups only weakly; but the use of the flexible strategy enables any desired degree of grouping to be obtained. In the binary case we normally discourage standardization of attributes prior to analysis, since the consequent increase in the importance of the presence of rare attributes, or the absence of common ones, seems to be unduly exaggerated for most purposes.

We cannot believe that a user who requires a classification will want classificatory boundaries weakened; we therefore believe that there is now no place for space-contracting strategies. Consequently, we submit that nearest-neighbour sorting should be regarded as obsolete, and we do not expect there to be any requirement for the flexible strategy with positive β. The main problem is to decide whether to use a conserving or dilating strategy; if the latter is desired, and if the flexible strategy is compatible with the measure in use, it is then necessary to decide the degree of dilatation desirable or acceptable.

All dilating strategies are inherently likely to produce one or more "non-conformist groups", whose members have little in common beyond the fact that they are rather unlike everything else, including each other. Whether this tendency is acceptable, or even desirable, depends on the precise use to which the classification is to be put. If, as is common in ecology and land survey, the analysis is required to reveal discontinuities which will then be the subject of later study, the discontinuities are required to be as sharp as possible, and a dilating

strategy is desirable. It is for this reason that the powerfully-dilating strategy of information analysis has been so successful in ecological work. However, if the project is genuinely taxonomic, in that every element is to be allocated to a group in the best possible way, strong dilatation is unacceptable. In such cases we advise a first run with a space-conserving strategy such as centroid or group-average; if the resulting picture is too fragmentary for ease of interpretation, we recommend the use of sufficient dilatation—i.e., a small negative value of β—to "clean up" the hierarchy but no more. We are currently using $\beta = -0{\cdot}25$ for this purpose; but a firm recommendation as to the optimal value of β must await further experience of these programs.

Appendix

A. Specification of Program CLASS. (Program I of p. 377)

The rationale of the method and details of the formulae used are given above; this appendix is designed to give prospective users sufficient information to (*a*) prepare data for the program and (*b*) understand the results.

1. *Data preparation.* Two types of data are allowed: (i) precalculated coefficients; (ii) basic data relating to individuals. Every set of data is preceded by two cards, the first contains four integers I, J, K and L punched in the FORMAT (4(I4, 2X)). I is the number of individuals, J is the number of attributes, K = 0 for normal analysis and K = 1 for inverse analysis, and L is the total number of final groups to be provided on the plotter. If all groups are required then L = 0, i.e. it may be omitted. The second card—which comes just before the data proper—is a title card and may contain any identifying information and if the data are precalculated coefficients ((i) below) must *not* contain the word DATA in cols. 1–4 whereas if the data are basic ((ii) below) the word DATA *must* be punched in cols. 1–4.

(i) *Precalculated coefficients* are punched using the FORMAT (7(F9.5,*,*)). Each row of the upper triangle of coefficients is punched—omitting the irrelevant diagonal element and each row must start on a new card. Cols. 73–80 may be used for identifying information and, of course, a total of $n(n-1)/2$ coefficients are punched.

(ii) *Basic data* may be either qualitative or numerical.

Pure qualitative data are punched to indicate, for each individual, which attributes are present, thus 2–5/8–10/21/34.

This means that in this individual attributes 2, 3, 4, 5, 8, 9, 10, 21 and 34 are present. Note that the *order* is unimportant i.e. 8–10 could follow 21 but sequences must be in ascending order, i.e. 10–8 is *not* permitted. Blanks may be inserted anywhere but the . after the last attribute is essential. The first 72 cols. can be used for data and continuation cards are allowed, thus cols. 73–80 are available for identification and are ignored on output. Each individual must start on a new card and the final individual has an * *immediately* following its . .

Numerical data are in the same FORMAT as precalculated coefficients; again each individual must start on a new card.

2. *Control Card.* A control card precedes the data deck and others, if required, follow the data. It contains parameters which determine the analysis to be performed. They are IC, IP, IW, IS, IO, BETA and are punched with FORMAT (5(I4, 2X), F7.0). They have the following meanings:

IC	EFFECT
0	No more cases to perform—end of calculation.
1	Use data again for a further case.
2	Read the data which follows punched in qualitative form.
3	Read the data which follows punched in numerical form.

IP	TYPE OF COEFFICIENTS COMPUTED
1	Correlation.
2	Squared Euclidean distance.
3	Non-metric distance.
4	Use SUBROUTINE SPECIAL to calculate coefficients. (This is to allow the user to extend the options available).

IW	EFFECT
0	No weighting of data.
1	Data standardized to zero mean and unit variance, for this run only.

IS	TYPE OF SORTING
1	Nearest neighbour.
2	Furthest neighbour.
3	Median (Gower).
4	Centroid.
5	Group average (Sokal).
6	Flexible.

IO	EFFECT
0	Coefficients are not printed or punched after they have been computed.
1	Coefficients are printed and punched and the analysis is completed.
2	Coefficients are printed and punched but analysis not completed.

BETA is the value of that quantity which is used only in the Flexible strategy i.e. when IS = 6. The field can be left blank in all other cases.

Many different analyses may be performed using the same set of data by simply adding control cards.

The IO facilities are included because we often have a requirement for the coefficients only so that they can be used for other purposes.

3. *Problem size.* The maximum problem which can be accommodated is given in the case of K = 0 (normal analysis) by

$$17I + J + I \times J + \tfrac{1}{2}I^2 + 3 < 17{,}000$$

or in the case of K = 1 (inverse analysis) by

$$17J + I + I \times J + \tfrac{1}{2}J^2 + 3 < 17{,}000.$$

4. *Output*

(i) *Printed.* The first line of output is the identifying information on the title card. This is followed by *Normal Analysis* or *Inverse Analysis* as appropriate and the data, which is output in the FORMAT (10(F9.4, 2X)). This Format is used even when the data were input in qualitative form. If a normal analysis has been requested the data are printed by individuals, i.e. as they are input from the cards, but, if inverse analysis is required, the data are output in the transposed form i.e. by attributes. If IW = 1 the standardized data is also output preceded by *Standardized.* At the top of the next page the type of coefficient used is printed and this is followed by the sorting strategy used. Next come the results themselves in the form

$$p + q = s \qquad C$$

where p and q are the element (or group) numbers which combine at level "C" to form a new group numbered s.

If printing of coefficients is requested (IO = 1 or IO = 2) these are output before the results which, of course, are suppressed if the coefficients only are needed (IO = 2).

After the last line of results has been obtained the next control card is read and the requested computation is performed. A blank control card causes termination of the job. If IC = 1 (same data to be used again) then the data are *not* output again but only the identifying information specifying the coefficient and sorting strategy.

(ii) *Plotted.* An hierarchical table is plotted after each set of results has been computed. The size of the table depends on the value of L on the control card; if L = 0 the whole table is plotted but if L $\neq$ 0 (but $\leqslant$ I or J for normal or inverse analyses, respectively) then only the "top" L final groups are displayed. However, in this case, the constitution of the final groups plotted is printed on the printer. The use of this facility is recommended when I (or J) $>$ 20 otherwise the table becomes a little difficult to interpret.

5. *Diagnostics*

Problem too large—when the appropriate condition of section A3 is violated.

Inadmissible case—when, for example, a control card of the form IC = 1, IP = 1, IS = 4 is supplied. Combinations which give the diagnostic are:

Correlation coefficient (IP = 1) with median (IS = 3), centroid (IS = 4) or flexible sorting (IS = 6) and Non-metric coefficient (IP = 3) with centroid sorting (IS = 4).

6. *General.* Note that it is not possible to perform an inverse analysis after a normal one and vice versa, without re-reading the data.

B. Specification of Program CENTBET (Program II of p. 337)

The theory behind this program is given in Lance and Williams (1966*a*) although certain features have been removed from this latest version because these are now included, more conveniently, in CLASS.

1. *Data preparation.* Data are prepared in exactly the same way as for program CLASS *except* that there is no provision for the use of precalculated coefficients.

2. *Control card.* A control card precedes the data deck and others, if required, follow the data. These cards contain parameters which determine the analysis to be performed. They are IC, IP, IW and are punched with Format (3(I4, 2X)). These quantities are defined as in specification of CLASS, *except* that IP = 2 is not permitted in CENTBET. Furthermore, when IP = 4 the Information Statistic is used but IP = 4 is only permitted if the data are purely qualitative, i.e. IC = 3.

3. *Problem size.* The maximum problem that can be accommodated is given, in the case of K = 0 (normal analysis) by

$$18I + J + I \times J < 17{,}000$$

or, in the case of K = 1 (inverse analysis) by

$$18J + I + I \times J < 17{,}000.$$

4. *Output.* This is similar to that described above for CLASS except that remarks relating to printing of coefficients do not apply.

5. *Diagnostics*

Problem too large—when the appropriate condition of section B3 is violated.

Inadmissible case—when, for example IP = 2 is specified; or IP = 4 with numerical data.

6. *General.* As before it is not possible to perform an inverse analysis after a normal one and vice versa, without re-reading the data.

References

DAGNELIE, P. (1960). "Contribution à l'étude des communautés végétales par l'analyse factorielle," *Bull. Serv. Carte Phytogeog.*, *B*, Vol. 5, p. 7.

DALE, A. G., DALE, N., and PENDERGRAFT, E. D. (1964). "A programming system for automatic classification with applications in linguistic and information retrieval research," Paper No. LRC 64 WTM-5, Linguistics Research Center, Univ. Texas.

EDWARDS, A. W. F., and CAVALLI-SFORZA, L. L. (1965). "A method for cluster analysis," *Biometrics*, Vol. 21, p. 362.

GOODMAN, L. A., and KRUSKAL, W. H. (1954). "Measures of association for cross-classification," *J. Amer. Statist. Ass.*, Vol. 49, p. 732.

GOODMAN, L. A., and KRUSKAL, W. H. (1959). "Measures of association for cross-classification, II," *J. Amer. Statist. Ass.*, Vol. 54, p. 123.

GOWER, J. C. (1966). "A comparison between some methods of cluster analysis," *Biometrics*, (in the press).

LANCE, G. N., and WILLIAMS, W. T. (1965). "Computer programs for monothetic classification ("Association analysis")," *Comp. J.*, Vol. 8, p. 246.

LANCE, G. N., and WILLIAMS, W. T. (1966*a*). "Computer programs for hierarchical polythetic classification ("Similarity analyses")," *Comp. J.*, Vol. 9, p. 60.

LANCE, G. N., and WILLIAMS, W. T. (1966*b*). "A generalized sorting strategy for computer classifications," *Nature*, Vol. 212, p. 218.

MACNAUGHTON-SMITH, P. (1965). *Some statistical and other numerical techniques for classifying individuals.* (Home Office Res. Rpt. No. 6) H.M.S.O., London.

MACNAUGHTON-SMITH, P., WILLIAMS, W. T., DALE, M. B., and MOCKETT, L. G. (1964). "Dissimilarity analysis: a new technique of hierarchical subdivision," *Nature*, Vol. 201, p. 426.

MCQUITTY, L. L. (1956). "Agreement analysis: classifying persons by predominant patterns of responses," *Brit. J. Statist. Psychol.*, Vol. 9, p. 5.

MATTSON, R. L., and DAMMANN, J. E. (1965). "A technique for determining and coding subclasses in pattern recognition problems," *IBM J. Res. Dvlpmnt.*, Vol. 9, p. 294.

NEEDHAM, R. M. (1962). "A method for using computers in information classification," *Proceedings of I.F.I.P. Congress* 62, p. 284.

NEEDHAM, R. M., and JONES, K. S. (1964). "Keywords and clumps," *J. Documentation*, Vol. 20, p. 5.

ROSE, M. J. (1964). "Classification of a set of elements," *Comp. J.*, Vol. 7, p. 208.

SOKAL, R. R., and MICHENER, C. D. (1958). "A statistical method for evaluating systematic relationships," *Kans. Univ. Sci. Bull.*, Vol. 38, p. 1409.

SOKAL, R. R., and SNEATH, P. H. A. (1963). *Principles of Numerical Taxonomy*, W. H. Freeman: San Francisco and London.

SPEARMAN, C. (1913). "Correlations of sums and differences," *Brit. J. Psychol.*, Vol. 5, p. 417.

THARU, J., and WILLIAMS, W. T. (1966). "Concentration of entries in binary arrays," *Nature*, Vol. 210, p. 549.

TRYON, R. C. (1955). *Identification of social areas by cluster analysis.* California: University of California Press.

WATSON, L., WILLIAMS, W. T., and LANCE, G. N. (1966). "Angiosperm taxonomy: a comparative study of some novel numerical techniques," *J. Linn. Soc.*, Vol. 59, p. 491.

WILLIAMS, W. T., and DALE, M. B. (1965). "Fundamental problems in numerical taxonomy," *Adv. Bot. Res.*, Vol. 2, p. 35.

WILLIAMS, W. T., LAMBERT, J. M., and LANCE, G. N. (1966). "Multivariate methods in plant ecology, V.," *J. Ecol.*, Vol. 54, p. 427.

Book Reviews

Analogues for the Solution of Boundary-Value Problems, by B. A. Volynskii and V. Ye Bukhman, 1965; 460 pages. (Oxford: *Pergamon Press*, 90s.)

This book could be of value to anyone interested in the solution of partial differential equations with boundary values. Although the emphasis in the book is on analogue methods of solution, this is because the authors have found that for many problems that they have studied in detail analogue techniques are preferable. Although the book was originally written in 1960 this first British edition has a specially written preface and a chapter describing the later work in the USSR. Even in the original edition the authors made a plea for hybrid computers, and their desire now is for a hybrid computer that contains digital, electronic analogue and network analogue, all fully programmable.

The layout of the book is rather unusual in that after the introduction in the first chapter the next two chapters contain, respectively, examples of how problems should be formulated and the mathematical methods available for the solution of such problems. Chapter 4 deals with the computational problems, while the following five chapters deal with specific analogue methods, mainly by the use of networks.

The book is a pleasure to read; it is written with an enthusiasm and humility that the translator and editor have carried over into the present edition so that the book is more "live" than are many translations. The publisher is to be complimented on the printing and format. The bibliography is poor.

J. S. GATEHOUSE