A general theory of classificatory sorting strategies

II. Clustering systems

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

Current clustering programs (i.e., non-hierarchical classificatory programs) are examined, with particular reference to the internal consistency of the methods used for initiation, allocation and reallocation. It is shown that almost all existing methods are open to serious objection, and that no method fully exploits the potentialities of such systems. The desirable properties of a clustering program are examined *de novo*, and suggestions made for optimum lines of further development.

1. Introduction

We shall in this paper adhere to the definition we have advocated previously (Lance and Williams, 1966, 1967): by a *clustering* strategy we imply a classificatory method which optimizes intra-group homogeneity, as distinct from optimizing a hierarchical route from individual elements to population. Such a strategy possesses two theoretical advantages over a hierarchical system . First, all agglomerative hierarchical strategies suffer from what has come to be known as the "migration problem": fusions rightly made in the early stages of the process may later prove unprofitable, in that a small number of elements may, at the working level finally adopted, be manifestly misclassified. Since a clustering strategy is (or should be) entirely concerned with group homogeneity, and since it may in addition make provision for a final reallocation procedure, this failing should be altogether lacking in a fully-developed cluster system. Secondly, a hierarchical classification of n elements must necessarily begin with the calculation of $\frac{1}{2}n(n - 1)$ inter-element measures; although a clustering system *may* begin in this way, it does not do so of necessity. The consequent possibility of beginning the classification process with a subset of the population should render these systems applicable to larger populations than can conveniently be handled by their hierarchical counterparts.

Unfortunately, clustering strategies are still in a relatively primitive stage of development, and no completely satisfactory system yet exists. Our intention in this communication is to examine the systems now current, and to establish their inter-relationships and their distinctive properties, with a view to the possible development of more rigorous methods. We have made our survey as complete as we are able; but work of this type is largely scattered through user journals or even circulated only in duplicated form, and we cannot be certain that no method has been overlooked.

2. Strategy and measures: general

A complete clustering system may in principle consist of four distinct processes, viz.:

(*a*) a method of *initiating* clusters;

(*b*) a method of *allocating* new elements to existing clusters, and/or of fusing existing clusters;

(*c*) a method of determining when further allocation may be regarded as unprofitable, so that certain elements remain unallocated as single-element clusters;

(*d*) a method of *reallocating* some or all of the elements to existing clusters when the main classificatory process is completed; this is intended to redress any misclassification produced by the "migration" process referred to in the previous section.

All systems necessarily involve (*a*) and (*b*); but in any particular system either (*c*) or (*d*), or both, may be lacking.

The measures on which the methods for (*a*) and (*b*) are based have received little or no critical attention; and, in particular, the earlier workers seem to have been insensitive to the desirability of using measures that will define an internally-consistent model with predictable properties. Since initiation must of necessity work at the element-element level, the measures involved are all, in the symbolism of our previous paper, (i, j)-measures, and present no special difficulties; but the problem of later allocation is complex. We first note that, in the great majority of existing systems, inter-cluster fusions are specifically forbidden; allocation measures are thus normally confined to specifying a relationship between an existing cluster and an individual element. If we are right in our contention that the important feature of true cluster-analysis is its ability to optimize a group rather than a route, it would seem logical to expect the use of an (i)-measure: an element would be added to that group with the smallest consequent (i)-value, providing this did not exceed a prescribed maximum.

More commonly a spatial model is used and a cluster regarded as being situated at a point, usually its centroid; an (i, j)-measure is then defined between this point and an element. However, such a measure may be regarded as a crude estimate of the largest radius of the cluster, in whatever space is defined by the measures in use; it can therefore be handled as if it were an (i)-measure. Finally, some workers have used the increase in an (i)-measure

* *C.S.I.R.O. Computing Research Section, P.O. Box* 109, *Canberra, A.C.T., Australia.*

on addition of an element as a decision-function for allocation. We are not attracted to such systems; for any such function is essentially an (ij, k)-measure, the (i)-measure of a single element being taken as zero. The system is therefore hierarchical, but lacks the elegance of overtly hierarchical systems.

The concept of space-distortion, which we introduced in our earlier paper, is also relevant to clustering systems. For example, suppose (in a qualitative system) that the total information content, I, was used as an (i)-measure; this measure is strongly space-dilating, and, as a cluster grows, it will be increasingly difficult for it to capture new elements. As the analysis proceeds there will therefore be an increasing tendency for an element to be allocated to the smallest existing cluster, almost irrespective of its attribute-structure. Conversely, suppose that the (i)-measure was the mean of all intracluster inter-element distances; this is a space-contracting measure and consequently elements will increasingly tend to be allocated to the largest existing cluster. The purpose of space-dilation in hierarchical systems is to delay the fusion of large groups and so reduce "chaining"; but since group fusions are normally not permitted in clustering systems, this property is no longer relevant. Moreover, neither in cluster nor in hierarchical systems is there any advantage in space-contraction; and it follows that it is in general desirable that all clustering measures should be space-conserving.

It is not possible to separate the allocation systems from the measures used to implement them, and neither provides a useful basis for classifying existing systems. The latter do, however, differ strikingly according to the method of initiation they employ; we shall therefore, in the section which follows, consider them on this basis. Reallocation we shall deal with in a separate section.

3. Specific strategies

(*a*) Methods operating on an inter-element similarity matrix

These all begin in the same way as an agglomerative hierarchical system, i.e., by the computation of an (i, j)-measure between all $\frac{1}{2}n(n - 1)$ pairs of elements. The non-hierarchical sorting procedure which follows may be carried out subjectively, as appears to be the case with Kaskey *et al.* (1962); more usually, each element in turn (in random order) is treated as a potential cluster centre, and those elements added to it which meet the allocation requirement prescribed. Two important systems of this type are the "Clustering Program II" of Bonner (1964), and the probabilistic system of Goodall (1964, 1966a); we deal with these in turn.

Bonner: "Clustering Program II"

Here the initial (i, j)-measure is extremely crude: whatever measure is computed, it is dichotomized by the declaration of an arbitrary threshold, so that every element either is, or is not, "similar" to every other. Core-clusters simply contain those elements similar to each other on this 1/0 basis. Two difficulties immediately arise: first, it is inevitable that the same cluster and/or its subsets will be found more than once. Bonner points out that there is no computational difficulty in recognizing such cases prior to output, so that clusters which are identical with, or subsets of, clusters already delimited are suppressed. However, there may still be an inconveniently large number of *almost* identical clusters; and Bonner has therefore suggested a means (his "Clustering Program I") of reducing the incidence of this duplication. The method is as follows: the original data-matrix of n elements specified by s attributes is used as the basis of the calculation of an $n \times n$ similarity matrix which is then dichotomized by reference to an agreed threshold. The resulting binary matrix is treated as though it were a data-matrix of n elements specified by n attributes; a new similarity matrix is calculated and dichotomized in its turn, and the process is iterated an arbitrary number of times. We know of no formal investigation into the properties of this system, and it does not even seem to be known whether it is convergent over all values of the threshold; but there is empirical evidence that it is in fact effective in "sharpening" the picture, and so reducing the number of near-identical clusters.

The second difficulty is that the clusters so produced will be "clumps" in the sense of Needham (1962), in that the same element can, and usually will, appear in more than one cluster.* If, as is perhaps most commonly the case an exclusive classification is required, Bonner defines an auxiliary program for refining the clumps into "disjoint core clusters"; this ends the initiation procedure.

Bonner's allocation system is complicated by the fact that he does not distinguish between the allocation of single elements not yet in core-clusters and the reallocation of the members of small clusters to larger clusters. However, since "small" must be arbitrarily specified, we will take the case when it represents a single element, i.e., simple allocation. Since every element can only be like or unlike every other, Bonner defines an element/cluster (i, j)-measure as the proportion of elements within the cluster which resemble the element under examination. The element is allocated to that cluster for which this proportion is highest; an arbitrary value of the proportion can be used as a stopping-rule. It seems probable that ambiguities would cause difficulty during the early stages of the analysis.

So far the system has been internally consistent; but Bonner now wishes to decide whether a given cluster should, or should not, be removed from the parent population. He therefore defines an extremely sophisticated probabilistic test (based on a conservative estimate of X^2) which uses the original exact values of the attributes, considered as co-ordinates in a Euclidean space. With characteristic candour, he discusses some of the

* Good (1965) has suggested that the term "clump" should be confined to clusters derived from a non-Euclidean model, but this usage appears unlikely to gain general acceptance.

resulting difficulties in the context of a specific example. Finally, he completes the process by defining a descriptive measure for the internal clustering strength of a cluster (I_{xx}) and the "interaction" between two clusters (I_{xy}). These are in fact simply the averages of the internal and external 1/0 similarities respectively; I_{xx} is a highly space-contracting (i)-measure, but it does not appear to be used for decision purposes. The system was published before the recent discussions on the place of probability theory in classificatory manipulations (Williams and Dale, 1965; Williams and Lance, 1965; Macnaughton-Smith, 1965; Goodall, 1966b); but it is in any case clearly undesirable that a test should be based on a model different from that by means of which the clusters were extracted.

Goodall

This is one of the few completely internally-consistent strategies in the literature. The initial (i, j)-measure is the complement of a probability, obtained as follows: let there be m ($= \frac{1}{2}n(n - 1)$) such measures, let them be ordered and let r of them exceed a stated value, and let p be the probability that this value is exceeded. Then if we accept a non-parametric model, r/m may be taken as the best available estimate of p. The resulting "probabilistic similarity index" is developed to a high degree of sophistication, and this is one of the very few systems that make provision for mixed data—i.e., data which contains qualitative, quantitative, multistate and/or ranked attributes. A computational drawback of such systems is that the inter-element measures depend on the properties of the population as a whole; consequently, if an element is added to, or withdrawn from, the population the entire similarity matrix must be recalculated.

Goodall first finds all clusters within which the maximum dissimilarity is less than the figure expected for such a maximum, given the set size, and a specified significance level. For each cluster all possible single-element accretions are examined by reference to a similar significance test, and the process continues until no further significant additions are possible. Of the various clusters resulting from this process, the largest is selected (or the most homogeneous if there are several of the same size). This cluster is removed, a new similarity matrix is computed for the residue, and the process repeated. A similarity matrix is also computed for the cluster itself, and sub-clusters are sought. We are informed by Goodall (*in litt.*) that the precise strategy of allocation is still under development.

In this case the (i)-measure of group homogeneity is, in effect, the probability that a set of the given size will contain the range of dissimilarities that it in fact does. The acceptable degree of dissimilarity will thus rise with group size; for this reason we should expect Goodall's strategy to have mild space-contracting properties, and to exhibit a slight tendency to "chaining". However, judgement of its properties must remain in abeyance until the strategy is available in a completely-developed form, and more experience is available of its application. However, it seems inevitable that the repeated computation of similarity matrices will remain an integral part of the procedure; the computation is thus unusually heavy, and use of the system is likely to be confined to problems of modest size for which its special properties are considered desirable.

(b) **Methods operating serially on all individual elements**

The essential feature of these methods is that they do not necessarily involve the computation of the entire similarity matrix, though this may be undertaken for a subsidiary purpose. The simplest system of this type known to us is again due to Bonner (1964), and constitutes his "Clustering Program III"; more complex systems have been devised by Hyvärinen (1962) and by Rogers *et al.* (Rogers and Tanimoto, 1960; Rogers and Fleming, 1964). We deal with these in turn.

Bonner: "Clustering Program III"

An element is chosen at random; its similarities with all others are computed, and those sufficiently similar are allocated to it to form a cluster. In principle, it would now be possible to exclude all elements in this cluster from further consideration. However, owing to the random start, such a cluster might well be sub-optimal; Bonner therefore imposes only the restriction that the members of this primary cluster may not be used again as the starting-points for new clusters. An "outside" element is therefore again chosen at random, and allocation made as before; it follows that the process will eventually exhaust the population, and that the resulting clusters will be clumps. Further processing would be necessary to render them exclusive.

The allocation (i, j)-measure is squared Euclidean distance; and Bonner again uses the X^2 test devised for his Clustering Program II. However, since this too is based on a Euclidean model, it is now completely consistent with the primary model, and our previous reservations concerning its use no longer apply.

Hyvärinen

There are obvious objections to a purely random start; and it was only to be expected that attempts would be made to ensure that the elements selected to initiate the process would have a high probability of being located in a dense region. Hyvärinen seems to have been the first worker to begin by placing the elements in an order of "typicality". His system, which is applicable only to qualitative or multi-state data, begins by defining an information content, I, of the entire group; the expression he gives is simply the multi-state form of the measure we similarly used for the qualitative case (Lance and Williams, 1966). His intention is to withdraw each element in turn and note the n consequent information losses; the element resulting in minimum loss must resemble a considerable number of others, and may rightly be regarded as "typical" (the system is, in fact, the exact antithesis of the "dissimilarity analysis" of Macnaughton-Smith *et al.*, 1964). We feel, however, that his statistic has not been used to its best advantage; for he calculates the ratio I/I_{max} (the "diversity") and defines typicality by the reduction in this ratio. However, I itself is a strictly additive quantity, and the information loss produced simply by subtraction, ΔI, is distributed nearly as $\frac{1}{2}X^2$; we consider that ΔI would therefore have been a better measure of typicality. The most typical member having been found, others are allocated to it in accordance with a criterion we consider below; when this can no longer be satisfied, the entire cluster so formed is excluded from further consideration, and a new typical element extracted from the remainder.

Unhappily, the allocation measure is an (i, j)-measure quite unrelated to the information statistic used for initiation. It is additive over attributes; for a single multi-state attribute, the contribution to the measure for two elements is r if both are in the rth state, 0 if they are in different states, and 1 if the state of either or both is unknown. It is not clear to us why the information model was abandoned at this point, since three compatible possibilities exist: an element might be allocated to that group which produced the lowest consequent I, ΔI or I/I_{max}. It is true that I and ΔI are space-dilating, and that there is reason to believe that I/I_{max} would be space-contracting; but these possibilities would undoubtedly repay further study.

Rogers and Tanimoto

This method is based on a theoretical system devised by Tanimoto (1958). Before considering it in detail we must first establish a simple general principle. Consider, for example, an inter-element similarity matrix using an (i, j)-measure constrained between 1 for identity and 0 for extreme difference. If this is added by columns, it follows that the element with the highest column sum must lie close to a number of others. It need not be at the centre of a cluster—in fact, if there happen to be two rather close clusters, the element in question is likely to lie between them—but it must lie in a dense region, and cannot be an outlier. The principle can be modified without difficulty for other types of measure, and no elaborate mathematical transformation is necessary or desirable. The objection to such a system for the establishment of typicality is simply the time involved in computing the entire similarity matrix.

This is, however, the method used in the system under discussion. Their measure s_{ij} is, in the normal (a, b, c, d) symbolism of a 2×2 contingency table, the coefficient $a/(a + b + c)$, which seems first to have been used by Jaccard (1908). If this were simply added by columns the element with maximum Σs_{ij} must lie in a dense region. However, Tanimoto ignores zeros, converts all non-zero values to $-\log_2 s_{ij}$ and adds. The rationale of this cumbersome procedure seems to be twofold: (*a*) the resulting quantities are measures of entropy, and so provide more fundamental estimates of the degree of disorder of the system, and (*b*) "since all the s_{ij}'s are independent probabilities", their product is the likelihood of the system.

To deal with these in reverse order, we must first state that we are not satisfied that all s_{ij} can be regarded as independent. For let the ith element possess b_i positive records out of a possible total of B attributes, and consider the general case where $b_j \neq b_i$ and let $b_j > b_i$. Then s_{ij} is constrained above, since it cannot exceed b_i/b_j; and if $(b_i + b_j) > B$ it is constrained below to $(b_i + b_j - B)/B$. The $\frac{1}{2}n(n - 1)$ quantities s_{ij} are thus constrained by the n quantities b_i; but these are not similarly constrained, since b_i/B can take any value between 0 and 1. The usual practice (vide, e.g., Bonner, 1964; Macnaughton-Smith, 1965) would of course be to regard *attributes* as potentially independent. Secondly, although it is true that Shannon (1948) states that any monotonic function of p can be used as an information measure, he gives compelling reasons for the use of $-\Sigma p \log p$ as an entropy function. For the qualitative case, the concepts of entropy and of likelihood are then simply related. If the jth of s attributes occurs in a_j out of n individuals, the best available estimate of the probability of its presence is a_j/n; the likelihood, L, of the entire population is then immediately given by

$$L = \prod_{j=1}^{s} \left(\frac{a_j}{n}\right)^{a_j} \left(1 - \frac{a_j}{n}\right)^{n-a_j}.$$

Furthermore, if the entropy of the system be denoted by H, and the information content $(= nH)$ by I, we have already shown that the Shannon recommendation results in the expression

$$I = sn \log n - \sum_{j=1}^{s} [a_j \log a_j + (n - a_j) \log (n - a_j)]$$

and it is immediately obvious that $I = -\log L$.

The allocation-measures used are based on the Tanimoto system; they are, however, complex (for an exposition, see Sokal and Sneath, 1963) and quite beyond any simple theoretical analysis of their potential space-distorting properties. The most striking feature of allocation is that, for an element to join an existing group, no less than three independent requirements must be met. These are symbolized by EN, β and δ. EN defines element-group similarity, and is an (i, j)-measure; β defines the resulting group homogeneity, and is an (i)-measure; δ defines the resulting change in group homogeneity, and is an (ij, k)-measure. It seems likely that the outcome of this triple sieve will be to make clustering impossible; all elements may be expected to appear members of a single ungroupable population. This is exactly what has happened in their example, 28 species of the algal genus *Halimeda*; and we find it inconceivable that any existing genus of living organisms of this size possesses no infra-generic structure. The entire system appears to have nothing to recommend it.

(*c*) Methods operating on a subset of elements

As the usefulness of numerical methods of classification has been increasingly widely admitted, users have submitted ever larger populations for processing. It is not surprising that workers in this field have increasingly wished to free themselves from the necessity of computing all $\frac{1}{2}n(n-1)$ pair-functions, or even all n single-element functions. As a result, no less than four closely-related methods based on subset procedures have been developed independently and almost simultaneously. All involve the prior declaration of the number of groups sought, k; this may or may not change during the course of the analysis. All require a definition of a group mean; this has currently been based on a Euclidean model, although MacQueen (1966) has pointed out that the concept of the centroid can easily be manipulated in non-Euclidean space. All use Euclidean distance for allocation.

The simplest, and perhaps the earliest, of these systems is that due to Forgy (1965). The population is arbitrarily partitioned into k groups, and the mean of each is found. It is now possible for an element to be nearer the mean of another group than it is to its own; all elements are therefore allocated afresh to the k means. A new set of means is calculated and the process is repeated; in practice it is found to converge rather quickly. An almost identical system has been proposed by Jancey (1966). This also uses constant k, but, using a strictly Euclidean model, begins by generating k points at random in the space under study. The elements are then allocated to these points, which are moved to the new means, and the process in essence proceeds as in the Forgy system. The initial partition is in this case less arbitrary than that of Forgy, since it represents a partition of a space and not of a population; the process would therefore be expected to converge more quickly.

A more flexible variant is the "k-mean" system of MacQueen (1966). In this, k of the elements are arbitrarily selected to act as primary centres; but, as each element is allocated in turn, a new mean is calculated. As a result, the k means shift as allocation proceeds; if two come closer than a predetermined value, they are fused, and k is reduced. An upper limit to allocation is also fixed; if a given element cannot therefore be allocated to the existing means, it forms a new nucleus and k increases. The programs ISODATA (Ball and Hall, 1965) and its modified form ISODA (Ball, 1965) resemble the k-mean system very closely; the only major difference lies in the mechanism whereby k, when necessary, is increased. In Ball's programs *all* elements are allocated at each iteration and the resulting clusters examined for heterogeneity, which is defined by the variance in each dimension considered independently. If an arbitrary value is exceeded in any dimension, and if certain criteria of group-size are satisfied, the cluster is split and k thereby increased. Fusion of clusters, as in the MacQueen system, depends only on inter-mean distance. It is not obvious why different criteria are used for fission and fusion, and the simpler, and internally-consistent, system of MacQueen appears to be preferable.

There must always be misgivings concerning systems which begin with a random subset or an arbitrary partition; and it is natural to enquire as to the extent to which the final answer may be dependent on the precise choice of starting-point. Jancey has suggested that the system can be improved by allowing the means, on taking up their new positions, to "over-swing" slightly, thus giving them an opportunity of escaping from sub-optimal positions. MacQueen has subjected the k-means system to an empirical test of stability: a population of 250 elements, with 5 attributes and with $k = 18$, was analysed three times with different starting-points. The results agreed to within 7 per cent, which is reassuring.

It is at first sight surprising that no attempt has been made to begin with a part-optimized sample. MacQueen has suggested this possibility; but, if we have not misunderstood his example, he is concerned, not with adding more elements, but predicting the behaviour of further attributes. This omission is probably due to the fact that, until recently, *all* methods—whether hierarchical or clustering—have been slow and uncertain; in particular, the widespread use of "nearest-neighbour" sorting in hierarchical systems, with consequent poor grouping, has rendered them inherently unsuitable as starting-points. Now that hierarchical methods are rapid and can be made to cluster as intensely as is wished (Lance and Williams, 1967) it might well be advantageous to begin by a hierarchical classification of one or more subsets, using the groups so produced as nuclei for the allocation of further elements. We shall return to this possibility in the final section of this paper.

4. Reallocation

The concept of reallocation is one of the most valuable features of clustering systems, and its successful application might go far to overcome the difficulties currently associated with all agglomerative systems. It is not, however, always practicable, and we must distinguish two situations.

First, if the system is such that a group mean or centroid is defined (or even definable), the problem is simple. At the end of the clustering process, the then means are taken as fixed and final points, and all elements are individually allocated to these points as cluster centres. This method is used by MacQueen, and could be used in ISODATA. It would be superfluous to attempt to apply the process to the Jancy and Forgy systems; for in their methods, since the number of centres is fixed, every iteration is in fact a reallocation of precisely this type. It is, however, important to note that this system of reallocation could be used with equal profit in a hierarchical analysis, provided the problem were such that the interest lay in the structure of the groups rather than in the discontinuities between them, and that the user had determined the level of subdivision at which he wished to work.

If, however, a cluster is defined by a "nodal" member (as with Hyvärinen and in Bonner's Clustering Program III), or if it is simply defined *per se* (as in Bonner's Clustering Program II), this procedure is impossible.

All that can be done is to compare every element in turn with every cluster in turn, usually with emphasis on reallocating members of small clusters; Bonner does in fact define an algorithm for a "cluster adjustment program", but his misgivings as to convergence may be deduced from his instruction, "Iterate . . . with the hope that stability will be eventually obtained". Nor does it seem to us that the concept of reallocation can be usefully applied to Goodall's system. In fact, if reallocation is a definite requirement, we see no practicable alternative to defining a cluster-centroid.

5. Conclusions

It will now be clear that few existing clustering methods can be recommended for use as they stand. Of those beginning with the inter-element similarity matrix, only that of Goodall is internally consistent. It is, however, computationally very heavy—it requires, in fact, more computation than would a hierarchical system involving the same number of elements, and it is likely to remain restricted to small-scale critical problems. We do not intend any criticism of the method; but since our own declared interest is to devise empirical methods for classifying very large quantities of data, we shall not consider it further. None of the methods beginning with the string of single elements is satisfactory, though that of Hyvärinen might well repay further development with particular reference to the use of an internally-consistent allocation procedure. The "subset" methods are in a different category; they are all consistent, and capable of extension to large populations. The most attractive is undoubtedly MacQueen's "k-mean" system, since new elements can be allocated individually, k is variable, and there is provision for a final reallocation procedure.

As overall conclusions we summarize what we consider to be the two most promising lines of future development:

1. A reallocation system should be developed for use with hierarchical systems; these, on account of their speed and flexibility, would then be used for any problem where initial computation of the inter-element similarity matrix was acceptable.
2. A clustering system should be devised on the following lines: given a large population, a random subset of appropriate size would be classified hierarchically. At a subdivision-level that requires investigation the process would be terminated and the means of the groups so obtained employed as nuclei for the MacQueen k-mean system, followed by reallocation. As a last operation, the final set of means could themselves be classified hierarchically to elucidate higher-level relationships.

We are ourselves proposing to initiate work along these lines.

References

BALL, G. H. (1965). Data analysis in the social sciences: what about the details? Stanford Research Institute, California.

BALL, G. H., and HALL, D. I. (1965). ISODATA, a novel method of data analysis and pattern classification, Stanford Research Institute, California.

BONNER, R. E. (1964). On some clustering techniques, *IBM J. Res. Dvlpment.*, Vol. 8, p. 22.

FORGY, E. (1965). Cluster analysis of multivariate data: efficiency vs. interpretability of classifications, W.N.A.R. meetings, Univ. California.

GOOD, I. J. (1965). Categorization of classification, In *Mathematics and computer science in biology and medicine*, H.M.S.O., London.

GOODALL, D. W. (1964). A probabilistic similarity index, *Nature*, Vol. 203, p. 1098.

GOODALL, D. W. (1966a). Numerical taxonomy of bacteria—some published data re-examined, *J. Gen. Microbiol.*, Vol. 42, p. 25.

GOODALL, D. W. (1966b). Classification, probability and utility, *Nature*, Vol. 211, p. 53.

HYVÄRINEN, L. (1962). Classification of qualitative data, *B.I.T.*, Vol. 2, p. 83.

JACCARD, P. (1908). Nouvelles recherches sur la distribution florale, *Bull. Soc. Vaud. Sci. Nat.*, Vol. 44, p. 223.

JANCEY, R. C. (1966). Multidimensional group analysis, *Aust. J. Bot.*, Vol. 14, p. 127.

KASKEY, G., KRISHNAIAH, P. R., and AZZARI, A. (1962). Cluster formation and diagnostic significance in psychiatric symptom evaluation, *Proc. Fall Jt. Computer Conf.* (*AFIPS Conference Proc.*, Vol. 22), p. 285.

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

LANCE, G. N., and WILLIAMS, W. T. (1967). A general theory of classificatory sorting strategies. I. Hierarchical systems, *Computer J.*, Vol. 9, p. 373.

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

MACNAUGHTON-SMITH, P., WILLIAMS, W. T., DALE, M. B., and MOCKETT, L. G. (1964). Dissimilarity analysis: a new technique of hierarchical subdivision, *Nature*, Vol. 202, p. 1034.

MACQUEEN, J. B. (1966). Some methods for classification and analysis of multivariate observations, Western Management Science Institute, Univ. California, Working Paper No. 96.

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

ROGERS, D. J., and FLEMING, H. (1964). A computer program for classifying plants. II. A numerical handling of non-numerical data, *Bioscience*, Vol. 14, p. 15.

ROGERS, D. J., and TANIMOTO, T. T. (1960). A computer program for classifying plants, *Science*, Vol. 132, p. 1115.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell System Tech. J.*, Vol. 27, p. 379 and p. 623.

Sokal, R. R., and Sneath, P. H. A. (1963). *Principles of numerical taxonomy*, W. H. Freeman: San Francisco and London.

Tanimoto, T. T. (1958). An elementary mathematical theory of classification and prediction, *IBM Taxonomy Application M. & A.*6, Vol. 3, p. 30.

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

Williams, W. T., and Lance, G. N. (1965). Logic of computer-based intrinsic classifications, *Nature*, Vol. 207, p. 159.

Book Review

Advances in Computer Typesetting: proceedings of the 1966 International Computer Typesetting Conference: (published May 1967), edited by W. P. Jaspert; 306 + xiv pages. (London: *The Institute of Printing*, 44 Bedford Row, London W.C.1. £5 net. Los Angeles (U.S. agents): *Composition Information Services*, 1605 North Cahuenga Boulevard, Los Angeles, California, 90028, $14.00.)

This book is a record of the proceedings of the International Computer Typesetting Conference held at the University of Sussex, and the Fairfield Halls in Croydon, in July 1966. The conference was attended by experts in computer typesetting from four continents, who spent two days in closed session at the University of Sussex and then presented their findings to a meeting in Croydon attended by 500 delegates. During the closed sessions, the experts endeavoured to assess the present state of the art of computer typesetting and to determine areas of future development. The closed sessions were sub-divided into four themes; Input equipment, Editing systems, Software and Hardware, and Graphic arts equipment. It contains all the papers presented at the conference, including those presented at Croydon, and an edited version of the discussions.

The book was produced by computer typesetting methods, using a system designed by Richard Clay & Co. Ltd. and Elliott-Automation Ltd., the latter company being responsible for the computer programs. Unjustified keyboarding of the majority of the text was carried out at Southwark Offset Ltd. in London; a second tape containing corrections and make-up instructions was perforated amd merged with the original tape on an Elliott 903 computer. The resultant tape was used to produce the final film on a Photon-Lumitype 713 in single column format for phototypesetting at Southwark. Some of the tabular work was set directly on Photon-Lumitype 540 machines at Southwark and because of the specialized nature of the texts, four papers were set "Monophoto" by the Universities Press, Belfast. The remainder of the reproduction and make-up were carried out at Southwark Offset, and the volume was printed and bound by Richard Clay (The Chaucer Press) Ltd., Bungay, Suffolk. Pages were made up using a standard page-layout grid system designed by Maurice Goldring (co-ordinating designer) who also devised a simple method of obtaining correct-size illustration copy for reproduction; (some of the smaller illustrations are not printed clearly and they resemble bad copies from an office copying machine).

The conference was organized by the Computer Typesetting Sub-Committee of the Institute of Printing, under the chairmanship of Mr. C. J. Duncan (Newcastle University).

Over sixty papers were presented during the sessions at Sussex. The section on *Input Equipment* includes six papers on keyboard training, four on coding of tape and keyboard design, three on character recognition, 13 on systems analysis and programming, and discussion. The section on *Editing Systems* contains 14 papers and discussion. *Software and Hardware* includes four papers on autocodes used with compilers, four on computer hardware and five on programs for typesetting and text handling. The section on *Graphic Arts equipment* contains eleven papers.

The *one-day conference* at Croydon contributed about 24 pages of discussion. There is (unfortunately) no index other than to contributors' affiliations and experience.

This book is not recommended for general reading, but it will be essential, if somewhat dated, material for those becoming involved with the application of computers to the practical side of the printing industry: the mathematical programmer who finds himself in this position might find a foretaste of problems to come, at page 54, with a brief description of the handling of mathematical formulae: the need for extensive keyboard conventions becomes apparent and several other papers are devoted to this.

In the software section the papers deal with special compilers required to translate the coded data recorded by the operative with typesetting knowledge into typesetting signals, there is a movement towards real-time working with problems ranging from input editing and proof correction to line justification with a minimum of hyphenated words and insertion of graphical illustration along with text.

The many problems met with in these developments appear in nearly all the papers. For example, to name only one, the importance of the electronic engineer appreciating what is meant by "acceptable quality" and the typographer/printer understanding the limitations of CRT character display *at the beginning* of the development of a cathode-ray tube typesetting process, rather than later.

The chairman of the sub-committee (Mr. C. J. Duncan) summarized the atmosphere of the conference

> ". . . . in general the picture was of piecemeal patchwork using mixtures of programs and hardware inadequate for an elegant and rational solution. It is not surprising therefore that there was little firm information on the economics of the processes employed, although clearly there was now a much larger number of people who were willing to take the risk of losing money with the hope of eventual advantage."

The reader of this book must therefore look around for later developments, since the conference, to keep up-to-date.

H. W. Gearing