WORKING PAPER 349

CLUSTER ANALYSIS AND Q-ANALYSIS

S.M. MACGILL

WORKING PAPER

School of Geography

University of Leeds

CLUSTER ANALYSIS AND Q-ANALYSIS

1. Introduction

Q-analysis has emerged over the past decade as a relatively new approach to data analysis and framework for thought for research within the social sciences. Partly due to its idiosyncratic notation and conceptual novelty it may be recognised by name only to many and not recognised at all by others. Several researchers, however, have been increasingly impressed both by its practical and its intellectual utility (Johnson, 1982, Gould, 1981, 1983). Different introductions to aspects of the method can be found in Atkin (1981), Chapman (1981), Gould (1980), Beaumont and Gatrell (1982) or Macgill (1982a). The original development of Q-analysis was the work of Atkin (1974, 1977).

In sharp contrast to the generally accepted novelty and distinctiveness of Q-analysis, the writer has recently been informed (private communication) that the characteristic algorithm of Q-analysis is no more than one of the oldest and best known methods of cluster analysis: the single-link (or nearest neighbour) method, assuming a similarity coefficient calculated by counting common descriptors. The purpose of this paper is to enable this hitherto unpublicised observation about the correspondence of Q-analysis with a particular method of cluster analysis to be more widely shared and to explore some of the immediate implications that this raises. There would appear to be a number of immediate benefits from so doing.

(i) A reduction in the insularity of Q-analysis from other methods of analysis. Q-analysis has been developed in relative isolation from other model-based and quantitative approaches in the social sciences and this inhibits researchers from being able to appraise its potential suitability in particular contexts. By relating Q-analysis to possibly more familiar territory - a particular branch of cluster analysis - the insularity of Q-analysis may be partially reduced. (That the consequent reduction in the insularity of Q-analysis may be only partial is important to appreciate; to interpret the methodology of Q-analysis simply as a clustering algorithm is obviously to misunderstand the essence of the approach.)

(ii) An additional perspective into the components generated by the standard Q-analysis algorithm. We may note in this respect that although the most distinctive contribution of Q-analysis as a new methodological perspective would appear to lie well beyond its clustering potential (being bound up with the use of a particular style of qualitative mathematical ideas associated with so-called traffic, and changes in methods of thinking that resulted in today's hard physical science - see the literature cited in the introduction above), in some cases Q-analysis has been used only for its clustering powers, ie. purely as a taxonomic device. In such cases the Q-analysis algorithm has been presented as an essentially new method of analysis (Beaumont and Beaumont, 1982, Gatrell, 1981) and only a minimum amount of comparison with other clustering methods has been made (Pinkava, 1981, Johnson, 1981b, Johnson, 1981c, Macgill, 1982b). It would now seem appropriate however to refer to longer experience with the same algorithm in traditional cluster analysis work.

(iii) New insights into a traditional clustering method. The traditional method with which Q-analysis has now been recognised to correspond is one which, as far as the author is aware, is not commonly used by social scientists. The popular CLUSTAN package (Wishart, 1978), extensively used by the latter, is based on quite a different set of clustering principles. However, in other fields of numerical taxonomy, the classification of plant communities for example, the traditional method explored in the present paper is apparently more popular. Thus the development of Q-analysis in social science may, as an incidental offshoot, have given new life to an inappropriately neglected method. Furthermore, standard outputs of the Q-analysis algorithm (eccentricities, structure vectors, q-nearness graphs, though not considered explicitly in this paper) potentially provide additional summary measures and sign-posts (additional to those implicit already in any clusters generated) for guiding an analyst back through his or her original data. Furthermore, the deeper philosophy of Q-analysis (via its distinctive interpretation of 'traffic' and other features alluded to under (ii) above) may facilitate a number of new insights to be given into groups found by traditional cluster analysis methods.

2. Cluster analysis and Q-analysis: a shared algorithm

Data for both cluster analysis and Q-analysis arises in the form of a set of N entities X_i, $i = 1, N$, to which a further set of p measurements or further entities α_j, $j = 1, p$ may be related (the α_j's may be descriptive features that the X_i's possess; or they may be other entities which the X_i's interact with or correspond to in some sense). Given information on whether or not entity α_j is related to entity X_i (see for example, table 1), it is possible to derive matrices depicting the degree of similarity or association (via shared α_j's) of the entities X_i to each other; see table 2. The four parts of this table will be considered in turn. In table 2a the number of common descriptors between each pair of entities is given. In cluster analysis terminology this would be called a similarity matrix. In Q-analysis a so-called shared face matrix is derived from table 2a by subtracting 1 from each entry (see table 2b). This subtraction is due to a pre-occupation in Q-analysis with the dimension of spaces that can contain objects. A 3-dimensional 'object' exists in 2-dimensional 'space', an n + 1 dimensional 'object' in n dimensional 'space'). In table 2c the values from table 2a are scaled so that they all lie between 0 and 1. Finally in table 2d the same information is depicted in a so-called distance matrix (by inverting everything). For the purpose of clustering elements according to a set of descriptors, the absence of a descriptor may be considered as significant as its presence. Thus when deriving the similarity and shared face matrices (eg. table 2) shared 0's between entities are counted as well as shared 1's. This is at variance with some applications of Q-analysis as a clustering algorithm (for example, Beaumont and Beaumont, 1982), where only the presence of an attribute is counted, though its absence could perhaps be equally significant.

Given the information in table 2 the elements X_i may be clustered according to the number of attributes (the characteristics α_j) they have in common with each other. This is given in algebraic and pictorial form in figure 1.

Entities having a given number of attributes are listed at the level corresponding to that number grouping together any pair of entities that share those attributes with each other. This is a simple procedure and can be readily computed. Equivalently from table 2d we may cluster the elements by fusing them according to the

Table 1. The relation between entities X_i and descriptors α_j ($i = 1,7$; $j = 1,12$)

	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8	α_9	α_{10}	α_{11}	α_{12}
X_1	0	0	1	1	0	0	0	0	0	1	1	0
X_2	1	1	0	0	0	0	0	0	0	0	0	0
X_3	0	1	0	1	0	0	1	0	1	0	0	1
X_4	0	0	0	1	1	0	0	0	0	0	0	1
X_5	0	0	0	1	0	0	0	0	0	0	1	0
X_6	0	0	0	0	0	0	0	1	1	0	1	0
X_7	0	0	0	1	0	1	1	1	1	0	0	0

Table 2a. <u>The similarity matrix</u>

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	12	6	5	7	10	7	5
X_2		12	7	7	8	7	5
X_3			12	8	7	6	8
X_4				12	9	6	6
X_5					12	9	7
X_6						12	8
X_7							12

Table 2b. <u>The shared face matrix</u>

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	11	5	4	6	9	6	4
X_2		11	6	6	7	6	4
X_3			11	7	6	5	7
X_4				11	8	5	5
X_5					11	8	6
X_6						11	7
X_7							11

Table 2c. <u>An equivalent similarity matrix</u>

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	1	$^6/_{12}$	$^5/_{12}$	$^7/_{12}$	$^{10}/_{12}$	$^7/_{12}$	$^5/_{12}$
X_2		1	$^7/_{12}$	$^7/_{12}$	$^8/_{12}$	$^7/_{12}$	$^5/_{12}$
X_3			1	$^8/_{12}$	$^7/_{12}$	$^6/_{12}$	$^8/_{12}$
X_4				1	$^9/_{12}$	$^6/_{12}$	$^6/_{12}$
X_5					1	$^9/_{12}$	$^7/_{12}$
X_6						1	$^8/_{12}$
X_7							1

Table 2d. <u>An equivalent distance matrix</u>

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
X_1	0	6	7	5	2	5	7
X_2		0	5	5	4	5	7
X_3			0	4	5	6	4
X_4				0	3	6	6
X_5					0	3	5
X_6						0	4
X_7							0

Figure 1. Q-analysis algorithm (equivalently, single link cluster analysis algorithm) results

Q level		Cluster level
q = 7 and below	(X_1 X_2 X_3 X_4 X_5 X_6 X_7)	4
q = 8	(X_2) (X_3) (X_1 X_4 X_5 X_6) (X_7)	3
q = 9	(X_2) (X_3) (X_4) (X_1 X_5) (X_6) (X_7)	2
q = 10, 11	(X_2) (X_3) (X_4) (X_1) (X_5) (X_6) (X_7)	1

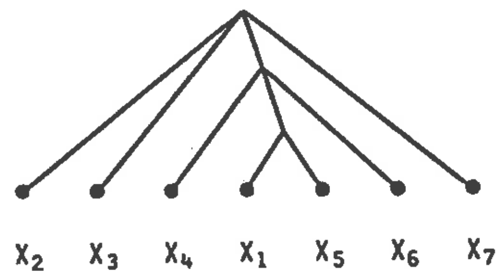
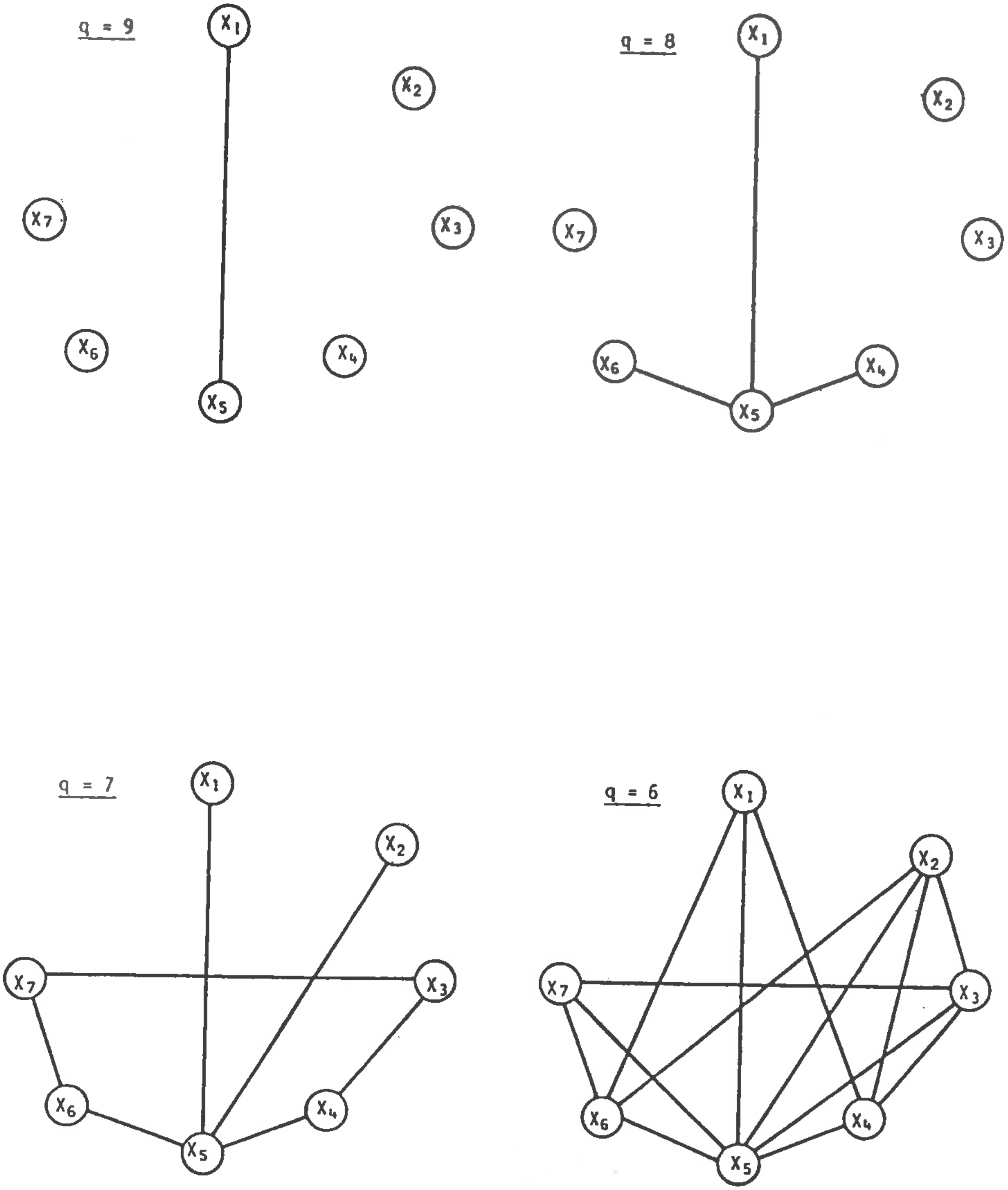


Figure 2. Q-nearness graphs for q = 9, 8, 7 and 6 *



* These graphs represent information from the face matrix and reveal the internal fabric of each of the clusters given in Figure 1. Thus the Q-analysis procedure reveals both a vertical (Figure 1) and a horizontal (Figure 2) structure of the original data set.

distance between their nearest member, the groups with the smallest distance being fused. Note that in the resulting pattern of clusters, two elements X_i and X_k, say, may appear in a given cluster at a given level either because they have the appropriate number of descriptors in common or because there is an indirect 'chain of connection' via one or several intermediate elements (eg. $X_i \to X_j \to X_k$ or $X_i \to X_l \to X_m \to X_k$). In Q-analysis terminology, the resulting clusters are sometimes called Q-connected components.

The pattern of clusters given in figure 1 can be interpreted by reading upwards from the foot of the representation. At cluster level 1 ($q = 11, 10$) it may be seen that all seven entities are distinct. For $q = 11$ this reflects the fact that each entity has a unique combination of the twelve possible attributes and for $q = 10$ it reflects the fact that each entity has a unique combination of eleven of the twelve. At cluster level 2 ($q = 9$) entities X_1 and X_5 are paired together to reflect the fact that they have ten attributes in common with each other. At level 3 ($q = 8$) X_4 and X_6 are included in this group due to having nine attributes in common with X_5. (The specific web of linkages within the relatively large cluster at this level can be seen in the q-nearness graph (for $q = 8$); see figure 2. The pivotal position of X_5 is notable here.) Finally at level 4 ($q = 7$ and below) all entities are in the same group reflecting the fact that each entity has at least eight attributes in common with some other entity. In general we may note that in the context of the present data set there is relatively little discrimination between entities in relation to the number of attributes. In principle given twelve attributes twelve different cluster levels could be defined, the seven possible additional levels referring to combinations of common occurrence of 6, 5, 4, 3, 2, 1 or 0 attributes. For the present data set, however, there is no hierarchical discrimination at these other possible levels, though the specific web of linkages within clusters varies (in general becomes richer) as the q-level decreases - see the graphs for $q = 7$ and $q = 6$ in figure 2.

The starting point above was a binary matrix. The single-link method is not confined to binary data, though where it is not, its correspondence with Q-analysis generally no longer holds. (More specifically, it is possible to carry out the above 'single link' algorithm on a similarity matrix that has been derived in some other way, based, for example, on product moment coefficients). Q-analysis is also not confined to binary data: in this case, if the original relation between X_i and α_j is numerical, a so-called slicing procedure would be used to discard any values below a given level of significance, replacing the discarded values by 0, and the remaining values by 1. Due to a possible arbitrariness in the initial choice of slicing levels, several alternatives can be tried, with the analysis repeated for each. It is not necessary to choose the same level of significance for all elements in the original data matrix; there could be a different level for each row, each column or even each cell.

3. Additional properties

The Q-analysis algorithm has been termed a 'data friendly' technique (Beaumont and Gatrell, 1982, Gatrell, 1981, Beaumont and Beaumont, 1982). It clusters the data without first working on or transforming them in any way. Thus, unlike many clustering methods, it does not impose illegal structure onto data. Moreover, it exploits finer aspects of connectivity than, for example, traditional matching score methods (see Johnson, 1981b for a brief comparison). We may see from the correspondence remark noted at the start of this paper and from a broader view of the clustering literature, that such properties are not unique to Q-analysis, being already possessed by a certain version of the single-link method. Everitt (1980) in a review of main-stream cluster analysis literature notes that other authors (for example Jardine and Sibson, 1971) have given different emphasis to the possible advantages which the single-link method holds over other agglomerative hierarchical techniques, noting that it is the only one to satisfy certain mathematical criteria (continuity, minimum distortion, etc.)

> "The major difficulty would seem to be that the similarity and distance measures calculated rarely have strict numerical significance. Because of the arbitrariness involved in scaling and combining different variables, there is rarely any justification for using the particular values rather than values obtained from some monotonic transformation; for example, their logarithms or square roots. Usually the values have only ordinal significance, and the only agglomerative hierarchical techniques applicable to coefficients with only *ordinal* significance are single link and complete link clustering." (Everitt, 1980, p. 68-9)

The same author notes that there may be other mathematical objections to complete link methods, and therefore, following Jardine and Sibson (1971), suggests that single link clustering may be the method of greatest mathematical appeal. Since single-link clustering and the basic Q-analysis algorithm are one and the same procedure (as long as they are based on equivalent* similarity and shared face matrices), the above remarks must also hold for Q-analysis.

Other authors have been less impressed by the need to satisfy strict mathematical requirements in choosing between different clustering methods, and give higher priority to other characteristics - non-hierarchical format, number of clusters required, computational efficiency, for example (Openshawe, 1982). Being confined to analysis of large data sets, that author argues strongly for computationally efficient clustering methods and is somewhat dismissive of slower hierarchical methods such as that considered in this paper. Rather than providing *a priori* grounds for universal exclusion of particular clustering approaches, it is suggested by the writer that criteria of mathematical significance and computational efficiency are more usefully to be seen as but two of the many considerations that ought to be brought to bear before choice or defence of a particular method is made. In a more catholic vein it may be reasonable to explore a given data set via a variety of different clustering procedures. This may be particularly suitable if the aim is to gain familiarity with and genuine insight into a given data set, rather than achieving a pre-specified format or structure of 'result'. We may note in passing that justification for the latter course may need more careful consideration than it is sometimes given. More detailed review of the properties and relative merits of different clustering procedures would go far beyond the scope of the present paper. Indeed, full monographs on the subject (Everitt, 1980, Duran and Odell, 1974, Jardine and Sibson, 1971) are apparently constrained by space availability.

---

*This means equivalent up to a monotonic transformation, and is an important qualification in terms, for example, of the philosophy of Q-analysis because a similarity matrix not satisfying this will introduce some distortion to the original data.

The aim of a cluster analysis may be stated as being

> "To devise a classification scheme for grouping objects into classes such that objects within classes (clusters) are similar in some respect and unlike those from other classes." (Everitt, 1980, p. 1)

The general goal of seeking classifications of objects into groups has been expanded by the same author following Ball (1971) into seven possible uses, namely: (i) finding a true typology; (ii) model fitting; (iii) prediction based on groups; (iv) hypothesis testing; (v) data exploration; (vi) hypothesis generating; (vii) data reduction.

The Q-analysis algorithm on the other hand was independently developed for use in the context of revealing natural or latent inter-linkage or connectivity of different dimensional strengths (corresponding to the different q-levels such as those given in figure 1) within a given data set with a view to identifying the scope for communication, activity or influence on and between the entities represented by the data (see ideas of traffic discussed in the basic introductions to Q-analysis referred to above and, for a more advanced analysis, Johnson, 1982). In the case of Q-analysis, the general aim of identifying clusters or Q-connected components within a given data set is thus pursued in order to be able to interpret these as parts of a latent but hitherto unseen multi-dimensional structure within the original data set. This in turn may act as a kind of backcloth and thus play a role in constraining or influencing 'activity' or so-called 'traffic' - it could be monetary expenditure, psychological stress, disease, movement of vehicles or whatever. Indeed Johnson (1981b) stresses this distinctive feature of Q-analysis by suggesting that it is inappropriate to use the designation 'Q-analysis' for an analysis in which there is no traffic.

The way that traffic and structure may mutually influence each other is suggested (Atkin, 1974, 1981) to be analogous to the way in which 3-dimensional 'physical' space in the 'real world' constrains (in terms of available 'routes') or influences (in terms of forces) physical movement or activity (vehicles, particles, people, etc.). From such influence there may be further reciprocal effects on the structure. The Q-connected components thus reflect channels, tunnels or spaces within the structure, at given dimensional levels (different Q-values). Atkin (1974, 1981) argues that these are the only 'locations' or channels through which certain types of 'traffic' can exist or be supported, or exert influence. Johnson (1981a) broadens the concept of traffic to include any graded pattern that can be mapped onto a structure, and thus goes beyond examples that can be related to the type of physical analogy alluded to above.

In the light of the somewhat different though complementary aims of cluster analysis and of Q-analysis, it is interesting to consider what has been cited elsewhere (for example, Forgey, 1965) as one of the drawbacks of the single-link clustering method, namely a so-called 'chaining' effect. Forgey concluded that the single link method performed well with very distinct clusters of any shape, but that as soon as a moderate amount of 'noise' was added (the presence of weak linkages between relatively distinct clusters, ie. overlaps, which would 'chain' those clusters together) the results became erratic. This chaining effect may be reinterpreted from the viewpoint of Q-analysis. As a preliminary observation, we may note that if the original data set contains genuine overlap between partially distinct groups, careful justification would seem to be in order before using some mechanical device (some black-box clustering algorithm) to 'remove' the natural overlap within the data, purely for neatness and apparent convenience of achieving distinct groupings (see also Gould (1981, 1983) on this point). More particularly, the chaining of one relatively distinct group to another is evidence of the existence of a 'route' between these groups along which so-called traffic can move. In the absence of such a 'route' it would not be possible for traffic within one group to 'reach' traffic in another. Thus far from being unwanted 'noise' which it is desirable for the analyst to suppress, the elements that produce a chaining effect may be useful and significant bonds in the system represented by a given set of data, which it may be desirable for the analyst to consider further.

The significance and utility of such bonds will clearly vary according to context, but in general we may note that prescriptive suggestions may be made in the light of bonds that exist. Q-nearness graphs giving pictorial representations of the information in table 2b are useful in this respect, see figure 2. Thus in certain cases it may be desirable to seek to agument existing weak bonds in the system (by attempting to add new relations between the original sets - cf. table 1) in order to facilitate movement of traffic (to ease the spread of knowledge, information, finance, or whatever, through particular systems). In other cases, it may be desirable to delete existing bonds in order to inhibit such movement (to inhibit the spread of disease through a species or congestion through a road system) or to break up closed loops in the structure. The sentiment in either case, however, is to use the Q-algorithm to reveal hitherto unseen aspects of data as a basis for informed analysis, not to force data into preconditioned moulds (for example by prespecifying the number of clusters that are to be found) or to impose structure that did not necessarily hitherto exist. Parallels between Q-analysis and network analysis which may be recognised in some of these comments have begun to be explored in the literature (see Earl and Johnson, 1981).

The methodologies of cluster analysis and of Q-analysis are still developing, thus the wealth of methods and interpretations for each of these approaches continues to increase. Johnson (1981b) for example, has recently provided a refinement of the basic Q-analysis algorithm for use in clustering work which preserves the advantageous properties repeated above from Everitt (1980, p. 68-9)*(see also Macgill 1982b). This involves inspecting the weights from the original data matrix to see if there are grounds for further dividing the components or clusters that arise from a standard Q-analysis. It turns out that there are two pairs of weights associated with each pair of entities in a cluster. If these weights are relatively similar, there would seem to be no grounds for further discriminating the clusters produced from the basic Q-analysis. However, if these weights are relatively different, there would seem to be grounds for further discrimination.

If Q-analysis is used only for its clustering powers (without any reference to other aspects of the approach) it will be increasingly difficult for users to defend their isolation from the wider clustering literature. In particular it will be necessary to defend what will have amounted to a choice, a particular version of the single-link method, in preference to other clustering methods that are available. Conversely, it is now open for users of traditional cluster analysis algorithms to consider the interpretation of 'traffic' as an additional potential benefit from using the single link method. Before reading too much

*and restores information that would otherwise become lost in a so-called slicing procedure.

significance into any mode of interpretation of either cluster or Q-analysis results, it is finally, however, important to recall that the similarity or closeness of entities as revealed by the basic algorithm may be highly dependent on the initial choice of variables (the X_i's) and on the number and variety of attributes taken into account (the α_j's).

References

Atkin, R.H. (1974) *Mathematical structure in human affairs,* Heinemann.

Atkin, R.H. (1977) *Combinatorial connectivities in social systems,* Birkhouse, Basel.

Atkin, R.H. (1981) *Multidimensional man,* Penguin.

Ball, G.H. (1971) *Classification analysis,* Stanford Research Institute, SRI Project 5533.

Beaumont, J.R. and Beaumont, C.D. (1982) A comparative study of the multivariate structure of towns, *Environment and Planning B,* 63-78.

Beaumont, J.R. and Gatrell, A.C. (1982) An introduction to Q-analysis, CATMOG 34, GeoAbstracts, Norwich.

Chapman, G.P. (1982) Q-analysis, in Wrigley, N. and Bennett, R.J. (ed.) *Quantitative geography: a British view,* Routledge and Kegan Paul.

Duran, B.S. and Odell, P.L. (1974) *Cluster analysis: a survey,* Springer Verlag.

Earl, C.F. and Johnson, J.H. (1981) Graph theory and Q-analysis, *Environment and Planning B, 8,* 367-91.

Everitt, B. (1980) *Cluster analysis,* 2nd ed., Wiley.

Forgey, E.W. (1965) Cluster analysis of multivariate data: efficiency versus interpretability of classification, *Biometrics, 21,* 768-9.

Gatrell, A.C. (1981) On the structure of urban social areas; explorations using Q-analysis, *Transactions of the Institute of British Geographers, N.S. 6,* 228-45.

Gould, P. (1980) Q-analysis: an introduction for social scientists, geographers and planners, *International Journal of Man-Machine Studies, 12,* 169-99.

Gould, P. (1981) Letting the data speak for themselves, *Annals of the Association of American Geographers, 71,* 166-76.

Gould, P. (1983) On the road to Colonus: or theory and perversity in the social sciences, *Geographical Analysis, 15*(1), 35-40.

Jardine, N. and Sibson, R. (1971) *Mathematical taxonomy,* Wiley.

Johnson, J.H. (1981a) Some structures and notation of Q-analysis, *Environment and Planning B, 8,* 73-86.

Johnson, J.H. (1981b) Q-discrimination analysis, *Environment and Planning B, 8,* 419-34.

Johnson, J.H. (1981c) A critique of the paper 'classification in medical diagnostics' by V. Pinkava, *International Journal of Man-Machine Studies, 15,* 239-48.

Johnson, J.H. (1982) Q-transmission in simplicial complexes, *International Journal of Man-Machine Studies*

Macgill, S.M. (1982a) An appraisal of Q-analysis (forthcoming).

Macgill, S.M. (1982b) A consideration of Johnson's Q-discrimination analysis, *Environment and Planning B, 9.*

Openshawe, S. (1982) The classification of large census data sets, Paper presented at the Third European Colloqium on Theoretical and Quantitative Geography, Augsburg, September 13th-17th, 1982.

Pinkava, V. (1981) Classification in medical diagnostics: on some limitations of Q-analysis, *International Journal of Man-Machine Studies, 15,* 221-37.

Wishart, D. (1978) *CLUSTAN 1C: user manual,* PLU, Edinburgh University.

April 1983