WORKING PAPER 343

A CONSIDERATION OF JOHNSON'S Q-DISCRIMINATION ANALYSIS

S.M. MACGILL

WORKING PAPER
School of Geography
University of Leeds

A CONSIDERATION OF JOHNSON'S Q-DISCRIMINATION ANALYSIS

In this paper an elementary exploration is given of some of the features of Q-discrimination analysis, an adaptation by J.H. Johnson of R.H. Atkin's Q-analysis for use in clustering work. The method is summarised. It is shown how the performance of the method varies considerably depending upon whether the data involved are integer or non-integer, on the range of values they take and on whether they are ordinal or interval.

S.M. Macgill
October 1982

A CONSIDERATION OF JOHNSON'S Q-DISCRIMINATION ANALYSIS

INTRODUCTION

In an intriguing paper, Johnson (1982) has recently shown how the basic methodology of Q-analysis can be refined and adapted for use as a clustering device, the resulting method being known as Q-discrimination analysis. Its initial exposition, giving particular attention to the theoretical structure of the method, may be beyond the grasp of many who may wish to use it. It is therefore summarised in simpler terms below. Following this, and the original rationale for the present paper, an elementary exploration of some of the features of Q-discrimination analysis is given. This involves an examination of how the performance of the method is likely to vary according to whether the data involved are in integer or non-integer form, and to the range of values the data take. Some familiarity with Q-analysis will be assumed; see Chapman (1981), Beaumont and Gattrel (1982), Gould (1981), Macgill (1982a) or Atkin (1981) for basic introductions, the conception and development of the method being the work of Atkin (1974), (1977).

CONTEXT

Q-discrimination analysis, like other clustering methods, has been developed for use in the context of deciding when elements of a given set are to be discriminated from each other with respect to a set of descriptive features. Suppose the elements are denoted by e_i i = 1, m. say, the descriptive features by d_j, j = 1, n, say, and a matrix (M_{ij}) can be defined whose (i,j) entries show that e_i is described by d_j with weight M_{ij}. The goal of clustering methods is to summarise the wealth of factorial data given by the matrix (M_{ij}), grouping together elements that are similar with respect to the descriptors, discriminating those which are not.

The values of (M_{ij}) are often integers lying between 1 and 5, perhaps resulting from an analyst's own judgement of the weight with which each element should be described by each descriptive feature. Wider ranges, of scale values between 1 and 7, 1 and 9 or more, may also occur, or, of course, non-integer data, if the weights arise on a continuous scale, or if a number of individual judgements made on an integer scale are pooled.

Clustering and discrimination of elements may be made on the basis of the matrix of weights (M_{ij}). There is by now a wide range of clustering techniques to choose from,** and paralleling the range of techniques, a corresponding diversity in the way given values of (e_i) (d_j) and (M_{ij}) may be clu~~st~~stered or discriminated. This diversity can be argued to be a weakness of the whole concept of clustering, because if elements can indeed be clustered in a wide variety of ways, the ideal of seeking a single representative summary of data on the basis of (M_{ij}) is severely undermined. Each method, however, has its own assumptions and hence degree of suitability in particular contexts. The particular virtues of Q-discrimination analysis, and consequent justification to add to existing wealth of techniques, will be indicated following the application of the method that is worked through below.

** see, for example, Everitt (1974), Duran and Odell (1974).

Q-DISCRIMINATION ANALYSIS : A WORKED EXAMPLE

The example to be worked through here is one given by Johnson (1982) in the original exposition of Q-discrimination analysis. Duplication of this example is considered warranted in order to present it in simpler terms and include some intermediate calculations. In this case the set of elements (e_i) are six different wineglasses which are each described by seven descriptive features, the latter being (1) crystal glass; (2) not fat stem; (3) tall; (4) tulip shape; (5) not vertical sides; (6) not cheap looking; (7) coloured transfer. The values (M_{ij}) are all integers lying between 1 and 5, a value of 1 denoting a descriptor applies strongly, and 5 denoting the opposite; for example, for descriptor 3 we have:

tall 1 2 3 4 5 short

x..................x

The full weighted matrix is reproduced in Table 1, the data representing Johnson's own judgement of the weight with which particular features should describe particular glasses.

Table 1. The weighted matrix of glasses vs descriptors

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
e_1	5	1	4	2	2	2	5
e_2	5	1	5	5	1	5	1
e_3	1	1	1	1	2	1	5
e_4	1	5	2	4	4	1	5
e_5	1	5	5	4	4	1	5
e_6	5	4	5	4	5	5	5

It will be recalled that the basic methodology of Q-analysis works on a data matrix whose values are either zeroes or ones. Q-discrimination analysis involves (i) transforming the matrix given in Table 1 into a matri of zeroes and ones by applying a particular slicing procedure (termed standard ordinal slicing by Johnson (1982)),(ii) performing a Q-analysis on the resulting matrix, thus partitioning the elements into distinct clusters at various levels of resolution, and (iii) inspecting the weights from the original data matrix to see if there are grounds for further refining the partitions that arise from the preceding Q-analysis. These three stages will now be worked through.

Step (i) : The original five-point integer scale divides into three parts, the first part for the values 1 and 2, the second part for the value 3 and the third part for the values 4 and 5. In line with the general interpretation of the scale values outlined above, values of 1 and 2 would show that an element related directly to the original descriptors, values of 4 and 5 to what Johnson appropriately calls their associated anti-descriptors, and the mid-point 3, denoting that neither descriptor nor anti-descriptor applied, (though in fact it does not arise in the present example) is called the complement.* This procedure, defining a so-called slicing mapping, generates Table 2 from Table 1.

Table 2. The result of applying a standard ordinal slicing to Table 1
d_i = descriptors $\tilde{d}_i$ = antidescriptors $\not{d}_i$ = complements

	d_1	d_2	d_3	d_4	d_5	d_6	d_7	$\tilde{d}_1$	$\tilde{d}_2$	$\tilde{d}_3$	$\tilde{d}_4$	$\tilde{d}_5$	$\tilde{d}_6$	$\tilde{d}_7$	$\not{d}_1$	$\not{d}_2$	$\not{d}_3$	$\not{d}_4$	$\not{d}_5$	$\not{d}_6$	$\not{d}_7$
e_1	0	1	0	1	1	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0
e_2	0	1	0	0	1	0	1	1	0	1	1	0	1	0	0	0	0	0	0	0	0
e_3	1	1	1	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
e_4	1	0	1	0	0	1	0	0	1	0	1	1	0	1	0	0	0	0	0	0	0
e_5	1	0	0	0	0	1	0	0	1	1	1	1	0	1	0	0	0	0	0	0	0
e_6	0	0	0	0	0	0	0	1	1	1	1	1	1	1	0	0	0	0	0	0	0

The slicing procedure can be defined more generally for any weighted scale for which it is possible to define a suitable cut-off between descriptors and anti-descriptors (it need not be the same cut-off for all descriptors), and Johnson calls this standard ordinal slicing. We may note that this step is really no more than identifying explicitly the fundamental attributes of the glasses that are relevant.

* Note that the division of the scale in this way is implicit in the definition of the scale anyway i.e. the slicing reflects this *a priori* meaning.

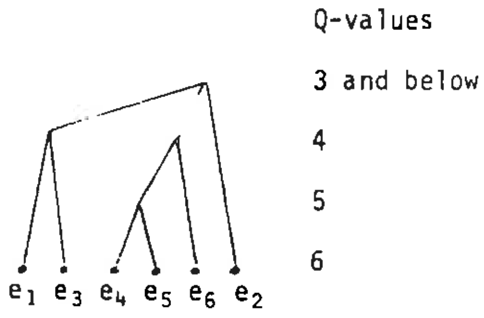
Step (ii) : A Q-analysis is performed on the data given in Table 2. The results are given in three different forms in Figure 1, parts (a), (b) and (c).

Figure 1. The results of a Q-analysis of the data from Table 2

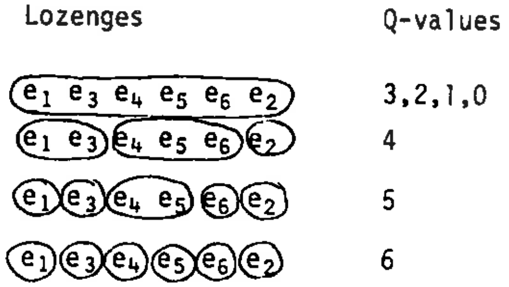
(a) Listing of components

$Q = 6$ (e_1) (e_2) (e_3) (e_4) (e_5) (e_6)
$Q = 5$ (e_1) (e_2) (e_3) (e_4,e_5) (e_6)
$Q = 4$ (e_1,e_3) (e_4,e_5,e_6) (e_2)
$Q = 3$ $(e_1,e_2,e_3,e_4,e_5,e_6)$
$Q = 2$ $(e_1,e_2,e_3,e_4,e_5,e_6)$
$Q = 1$ $(e_1,e_2,e_3,e_4,e_5,e_6)$
$Q = 0$ $(e_1,e_2,e_3,e_4,e_5,e_6)$

(b) Tree representation



(c) Lozenges



(See the Appendix to this paper for the derivation of these results). From Figure 1 we see that the pair of glasses most similar to each other are glasses 4 and 5, these but no others being grouped together at level $Q = 5$. This means that glasses 4 and 5 (but no others) have 6 descriptive features in common with each other. Since there are only seven descriptive features in all, we would expect glasses 4 and 5 to be almost the same (indeed, see Johnson (1982), figure 8). At a slightly coarser level of resolution, $Q = 4$, glasses 1 and 3 are grouped together, and in a different group, glasses 4, 5 and 6. This is because glasses 1 and 3 have five descriptive features in common with each other, and any pair of the glasses 4, 5 and 6 also all have five descriptive features in common with each other. At $Q = 3$, no discrimination can be made between the glasses. This is because each glass has at least four descriptive features in common with at least one other glass. Thus Figure 1 indicates a natural grouping of the glasses on the basis of the information given in Table 1.

Step (iii) : It will be noticed that although a reasonable amount of discrimination between the six glasses has now been derived, in basing the discrimination on the binary matrix of Table 2 rather than the original data in Table 1, a certain amount of information has been ignored. In the third step, the data from the original matrix, Table 1 is inspected, according to the procedure given below, to see whether there are grounds for refining further the partitions, or clusters, given in Figure 1.

Note that for any set of descriptors that two glasses have in common with each other there will be two sets of weights, one set being associated with each glass. As an example note that from Table 2 glasses 1 and 3 share the descriptors d_2, d_4, d_5, d_6, d_7 (and corresponding anti-descriptors $\tilde{d}_2$, $\tilde{d}_4$, $\tilde{d}_5$, $\tilde{d}_6$, $\tilde{d}_7$). However, from Table 1 for glass 1, the weights on these descriptors are 1, 2, 2, 2, 5 whereas for glass 3 they are 1, 1, 2, 1, 5. Wherever such sets of weights are relatively similar there would seem to be no grounds for further discriminating the clusters produced from the basic Q-analysis. However, if such weights turn out to be relatively different, there would seem to be grounds for further discrimination. Johnson (1982) provides one possible procedure for making this relative comparison. A number of substeps are involved. For the present example, this amounts to the following:

Firstly, the similarity of the weights for each pair of glasses is examined, for the descriptors they have in common.

For glasses e_1 and e_3 this gives (1,1) (2,1) (2,2) (2,1) (5,5)

" " e_1 and e_2 " " (5,5) (1,1) (4,5) (2,1)

and so on for all other pairs. The complete list of such terms involving glass 1 is given in Table 3. (A corresponding table for all other glasses can be derived).

Table 3. The relative weights for descriptors held in common between glass 1 and other glasses

e_1e_1	(5,5)	(1,1)	(4,4)	(2,2)	(2,2)	(2,2)	(5,5)
e_1e_2	(5,5)	(1,1)	(4,5)	(2,1)			
e_1e_3	(1,1)	(2,1)	(2,2)	(2,1)	(5,5)		
e_1e_4	(2,1)	(5,5)					
e_1e_5	(4,5)	(2,1)	(5,5)				
e_1e_6	(5,5)	(4,5)	(5,5)				

Secondly, from such a listing, it is possible to assess the relative level of similarity of all glasses to glass 1. The glass most similar to glass 1 is, of course, itself. By inspection of Table 3 the next most similar glass is glass 3, since this has more descriptors in common than any other. By a similar argument, glass 2 is the next most similar to 1. Note that glasses 5 and 6 both have 3 descriptors in common with glass 1. However, rather than putting them both at the next level of similarity, we may note that between glasses 1 and 6, two of the three weights are the same, whereas between glass 1 and 5 only 1 of the three weights is the same. Thus glass 6 is more similar to 1 than 5 is to 1. Finally, glass 4 is the least similar to glass 1. The argument of this paragraph generates levels of similarity of all glasses to glass 1 given in Table 4.

Table 4. Levels of similarity of all glasses to glass 1

Level 1	Glass 1
Level 2	Glass 3
Level 3	Glass 2
Level 4	Glass 6
Level 5	Glass 5
Level 6	Glass 4

A corresponding table may be derived for all other glasses.

Note that most of the information used in compiling this table was already embodied in the original Q-analysis. This can be seen in that all except levels 4 and 5 were determined by a straight count of the number of common descriptors. Only for levels 4 and 5 (glasses 6 and 5) were the weights used. Whether this holds for other examples remains to be seen.

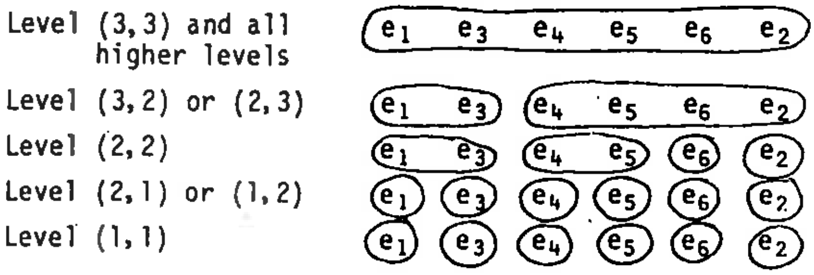
Thirdly, the information given in Table 4, and corresponding tables for other glasses is summarised in the matrix (U_{ij}) where U_{ij} is the level of similarity of glass i with respect to glass j. The first column in this table thus reproduces the information from Table 4 ($U_{11} = 1$, $U_{31} = 2$, $U_{21} = 3$, $U_{61} = 4$, $U_{51} = 5$, $U_{41} = 6$); other columns would arise from analogous tables.

Table 5. (U_{ij}) giving the level of similarity of glass i to glass j

	1	2	3	4	5	6
1	1	3	2	5	5	5
2	3	1	5	6	6	3
3	2	4	1	3	4	6
4	6	5	3	1	2	4
5	5	4	4	2	1	2
6	4	2	6	4	3	1

Table 5 offers a means of deciding whether the groupings given in Figure 1 should be refined. This is because Table 5 suggests its own way of grouping the glasses on the basis of the information in Table 1 and this grouping may be amalgamated with the original Q-analysis grouping from Figure 1. The grouping suggested by Table 5 is found by identifying how the glasses group at successively coarse levels of similarity. In other words, for any pair of integers (I,J), groups defined at level (I,J) are those for which Uij $\leq$ I and Uji $\leq$ J. (Level (I,J) is the same as level (J,I).) The resulting grouping is given in Figure 2.

Figure 2. The grouping suggested by Table 5


Level (3,3) and all higher levels
e_1 e_3 e_4 e_5 e_6 e_2
Level (3,2) or (2,3)
e_1 e_3 e_4 e_5 e_6 e_2
Level (2,2)
e_1 e_3 e_4 e_5 e_6 e_2
Level (2,1) or (1,2)
e_1 e_3 e_4 e_5 e_6 e_2
Level (1,1)
e_1 e_3 e_4 e_5 e_6 e_2

Reading upwards from the foot of the table we see that there are no groups at the first two levels, but that (e_1 and e_3) and (e_4 and e_5) are grouped together at level (2,2). This is because glasses e_1 and e_3 lie at the second level of similarity to each other ($U_{13} = 2$, $U_{31} = 2$) as do e_4 and e_5 ($U_{45} = 2$, $U_{54} = 2$). There are no other groupings at this level because all otherordered pairs $(U_{ji}, U_{ij}) > (2,2)$. At level (3,2) or (2,3), the next level up, e_1 and e_3 remain as a pair, but the group e_4 and e_5 is expanded to include e_6 and e_2 as well. This is because for instance $U_{65} = 3$, $U_{56} = 2$ (ie 5 and 6 lie at level (2,3)), and $U_{62} = 2$, $U_{26} = 3$ (ie 2 and 6 lie at level (2,3)). At level (3,3) and all higher levels it can be seen from inspection of Table 5 that all glasses can be grouped together.

The grouping of glasses given in Figure 2, can be compared with that given in Figure 1, and the latter can be used to refine the former. The refined grouping is given in Figure 3(c)*, parts (a) and (b) of this figure repeat, for convenience, the information from Figures 1 and 2. The extra discrimination of Figure 3(c) over Figure 3(a) or 3(b) can be seen.

The full Q-discrimination analysis can thus be seen to involve three steps (i) standard ordinal slicing; (ii) a standard Q-analysis on the basis of (i); and (iii) inspection of discrimination weights.

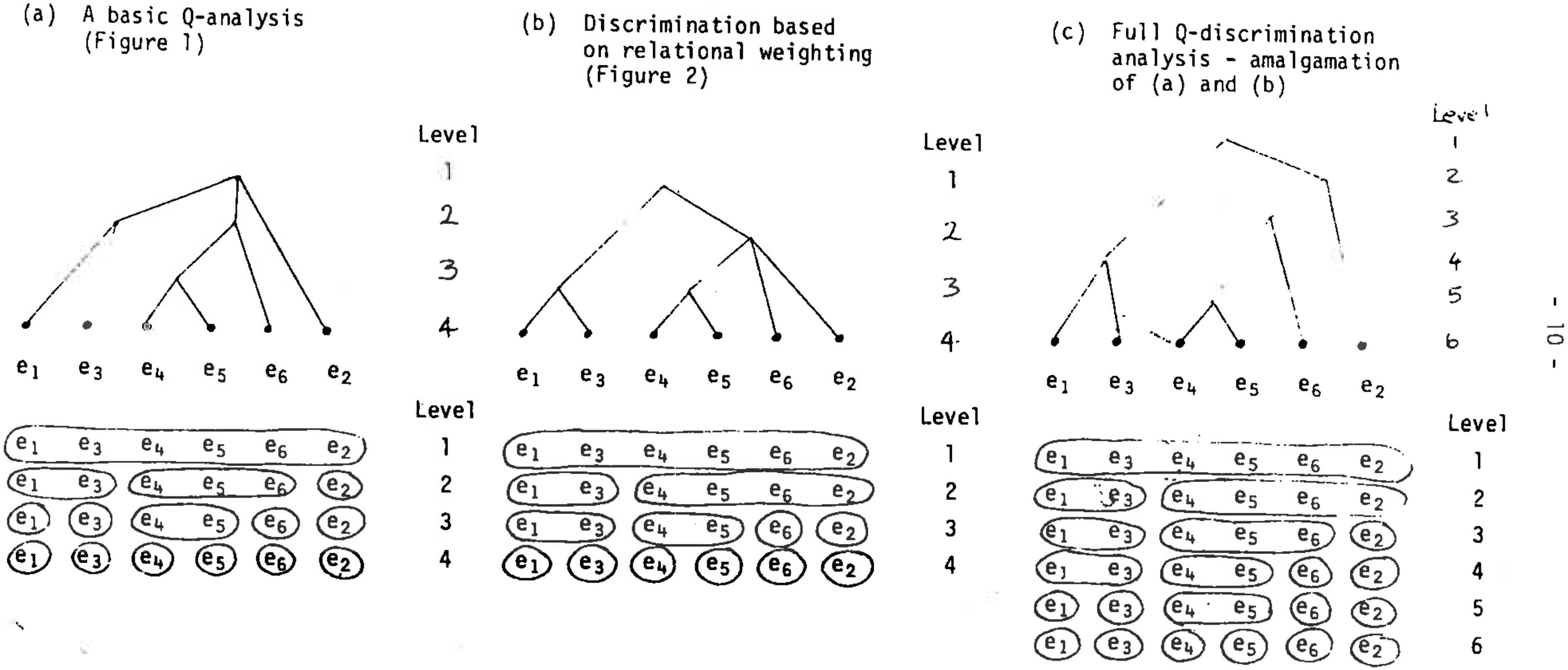
MERITS OF Q-DISCRIMINATION ANALYSIS

Q-discrimination analysis may appear at first to be a clumsier clustering method than others that are available. However, as with other seemingly complex methods, the complexity decreases with increased familiarity and computer programmes eliminate the chore of hand-worked manipulations. More importantly, Johnson (1982) argues that a number of properties of the method can more than compensate for its apparent messiness, and thus put it as a superior clustering method. The main properties are:-

1. The method is unique among clustering methods in assuming only an ordinal order on the scale of weights M_{ij}. (Thus in the above example

* Figure 3(c) in this paper is different to Johnson (1982) Fig. 7(d); the latter appears to have misrepresented the full results of the Q-discrimination analysis.

Figure 3. Results of full Q-discrimination analysis



(a) A basic Q-analysis (Figure 1)

(b) Discrimination based on relational weighting (Figure 2)

(c) Full Q-discrimination analysis - amalgamation of (a) and (b)

Q-values are: 3,2,1,0 at level 1
4 " " 2
5 " " 3
6 " " 4

Discrimination pairs are:
(3,3) and higher at level 1
(3,2) or (2,3) at level 2
(2,2) at level 3
(2,1) or (1,2) and (1,1) at level 4

Q-values and discrimination pairs are:

At level	Q	Discrimination pairs
1	Q = 3,2,1,0	(3,3) and above
2	Q = 3,2,1,0	(3,2) or (2,3)
3	Q = 4	
4	Q = 4	(2,2)
5	Q = 5	
6	Q = 6	

it is assumed only that 1<2<3<4<5 and therefore, for instance, that 2<5, 3<5 and so on. However no significance is read in to the "difference" between 3 and 5 or between 1 and 4. It cannot be said, for instance, that 3 is more similar to 5 than 1 is to 4. The latter can only be said if the data are interval, because in this case, it is legitimate to examine the difference between pairs of values. It is not surprising that results of clustering exercises are different depending on whether the data are assumed to be ordinal or interval. With most of the traditional methods, the interval assumption generates richer clustering. However, Johnson (1982) prefers the weaker ordinal assumption because the interval assumption may introduce illegal "structure" into the data.)

2. The method has been found to be relatively more discriminating than many other methods, ie. identifies a richer variety of groupings.*

3. The method exploits the individual similarities of each element to each other, on the basis of each individual descriptor, not as in the case of other methods, some broader aggregate.

To these we may add a couple of basic qualities of Q-analysis.

4. It is a data-friendly approach in that throughout it uses the data in its original form, and does not "work" on it (for example by calculating correlation coefficieints or other proxy representations).

5. The groupings are natural in that they exploit only the inherent connectivities between all pairs of elements.

Properties 2-5 are all closely interlinked and would appear to make a strong case for choosing Q-discrimination analysis in preference to other possible methods in many contexts. Notwithstanding Johnson's view of the importance of the first property the present author's judgement about the significance of property will be reserved until various further explorations of the method have been reported below. These consider non-integer values for M_{ij} and integer values that extend over a wider range than 1-5. A more general statement of the Q-discrimination step in the above method analysis is given first.

* The existing methods that generate richer groupings typically do so by introducing richness that was not inherent in the original data.

Q-DISCRIMINATION ANALYSIS : A MORE GENERAL STATEMENT

In his original paper Johnson (1982) presented a more detailed procedure for assessing the relative level of similarity U_{ij} of element j with respect to i in step (iii) above (ie. a more detailed procedure for deriving Table 5). Johnson's more general statement amounts to the following: Pairs of scale values for descriptors shared between element i and all other elements are first written down (corresponding to Table 3). Suppose, for example, that e_i and e_j share descriptors k_1, k_2 and k_4 and e_i and e_l share descriptors k_1, k_3 and k_4, we get:

For e_i and e_j: (M_{ik_1}, M_{jk_1}) (M_{ik_2}, M_{jk_2}) (M_{ik_4}, M_{jk_4})

For e_i and e_l: (M_{ik_1}, M_{lk_1}) (M_{ik_3}, M_{lk_3}) (M_{ik_4}, M_{lk_4})

Since the values of the elements in the matrix (M_{rs}) are all taken from the same scale (S_i), where $S_1 < S_2 < S_3$ ...... $< S_n <$ ..., the partial order relation defined in Figure 4 may be used to determine which of j or l is more similar to i.

Figure 4. Partial order relation on (S_i, S_j)*

$$\begin{array}{ccccccccc}
(S_1,S_1) & > & (S_1,S_2) & > & (S_1,S_3) & > & \cdots\cdots & > & (S_1,S_r) \\
\vee & \rotatebox{-45}{$>$} & \wedge & & & & & & \wedge \\
(S_2,S_1) & < & (S_2,S_2) & > & (S_2,S_3) & & \cdots\cdots & > & (S_2,S_r) \\
\vee & & & \rotatebox{-45}{$>$} & & & & & \wedge \\
\vdots & & & & \ddots & & & & \vdots \\
\vee & & & & & & \rotatebox{-45}{$>$} & & \wedge \\
(S_r,S_1) & < & (S_r,S_2) & < & (S_r,S_3) & < & \cdots\cdots & & (S_r,S_r)
\end{array}$$

This method of comparing similarity can be used to find all levels of similarity to i for all elements (and hence column i in the matrix U_{ki}). The remainder of the matrix (U_{ij}) can be found in an analogous way.

* > in this figure means "more similar than".

FURTHER EXPLORATIONS

Both of the worked examples given by Johnson (1982) involved (only) integer values for M_{ij} on a scale between 1 and 5. The desire of the present author to explore the applicability of the method to non-integer data over a much wider range of values (Macgill 1982b) led to the following findings. They all turn on one or both of two specific stages in the above description of Q-discrimination analysis, namely (a) the definition of the complementary vertices in the standard ordinal slicing, and more importantly, (b) the use of Figure 4 in identifying relative levels of similarity when undertaking the further discrimination in step (iii). Four classes will be discussed below.

1. Integer scale, ordinal data. Consider first the case when the range of the scale is 1 to 5, as above, values 1 and 2 relating to the descriptor, 4 and 5 to the anti-descriptor and 3 being the complement. In this case the identification of relative levels of similarity (ie derivation of Table 5) can be considerably simplified. This is because it can be based on a simple count of the total number of common descriptors between pairs of elements, specifying only the number which have the same weights, and the number which have different weights. Thus rather than using information in the form of Table 3 it is more convenient to re-specify it as in Table 6, with, of course, corresponding tables for other e_i's.

Table 6. Alternative means of comparing weights w.r.t e_i

	Total number of descriptors in common with e_i	Number with the same weights	Number with different weights
e_1e_1	7	7	0
e_1e_2	4	2	2
e_1e_3	5	3	2
e_1e_4	2	1	1
e_1e_5	3	1	2
e_1e_6	3	2	1

Table 6 (and the other corresponding tables.) can be derived directly from Table 1 thus greatly reducing the required workings and simplifying the method considerably. It is readily seen that ranking the similarity on the basis of Table 6 reproduces column 1 in Table 5 as required.

Consider next the case of an integer scale whose values range between 1 and 9, say. In this case, the most obvious division of the scale into three parts as between descriptors, anti-descriptors and complement would seem to be (1,2,3,4), (5) and (6,7,8,9). In this case if we assume no more than an ordinal order on the scale, then when comparing pairs of weights in order to assess relative levels of similarity (c.f. Figure 4) we cannot say, for instance, for the hypothetical data given in Table 7 and expression (1)

Table 7. Hypothetical data

	d_1	d_2	d_3	d_4
e_i	6	1	6	7
e_j	6	2	0	0
e_k	6	0	9	0
e_l	6	0	0	9

$$\begin{aligned} e_i\ e_j &= (6,6)\quad(1,2) \\ e_i\ e_k &= (6,6)\quad(6,9) \\ e_i\ e_l &= (6,6)\quad(7,9) \end{aligned} \qquad (1)$$

that j is more similar to i than k is to i. This is because the ordinal order will not let us say that 2 is more similar to 1 than 6 is to 9. Thus, the apparent increase in information made available by a range of 9 rather than 5 cannot be exploited. We may, however, be able to say that l is more similar to i than k is to i (because we can compare (6,9) with (7,9)).* Thus we set the following levels of similarity (either (a) or (b))

* Whether or not this is strictly legitimate depends on whether we may compare scale values for different descriptors (note that (6,9) relates to d_3, but (7,9) to d_4. We assume here that it <u>is</u> legitimate, but see Johnson (1982) p.425.

	a	b
Level 1	element e_i	element e_i
Level 2	element e_l	element e_l and e_j
Level 3	element e_j and e_k	element e_k

From this we can deduce that the greater the range of scale values, the less likely it is to be able to exploit the apparent increase in information. This restriction is due to the ordinal condition of the data. We consider next what happens when the ordinal assumption is relaxed.

2. Integer scale, interval data. In this case we can positively discriminate between the relative levels of similarity between j, k and i in Table 7 and expression (1), because the more relaxed interval assumption means that 1 is more similar to 2 (a difference of 1) than 6 is to 9 (a difference of 3). Thus in this case, we get

Level 1	element e_i
Level 2	element e_j
Level 3	element e_l
Level 4	element e_k

a somewhat different order than before.

The ability to infer levels of similarity by taking the difference between scale values for shared descriptors allows the procedure to be readily computed. In this case, instead of a Table of the form of 3 or 6 for each element e_i. we simply need to identify the number of shared descriptors and compute a set of terms giving the sum of the differences between i and k, d_{ik}, say,

where

$$d_{ik} = \sum_{\substack{j \text{ shared} \\ \text{by } e_i \text{ and } e_k}} |M_{ij} - M_{kj}| \tag{2}$$

The relative levels of similarity may then be found by identifying the number of shared descriptors and ranking the terms d_{ik} in ascending order, lowest first, for each given number of shared descriptors. In the case of expression (1) we get $d_{ij} = 1$, $d_{ik} = 3$, $d_{il} = 2$, $d_{ii} = 0$, and thus the levels already given above. Suppose in another context we have $d_{i_1} = 0$, $d_{i_2} = 15$, $d_{i_3} = 7$, $d_{i_4} = 10$, $d_{ii} = 0$ and element i shared 3 vertices with elements 3 and 2, and 2 vertices with element 4. Then we would find the following levels of similarity with respect to element i

Level 1	element i and 1
Level 2	element 3
Level 3	element 2
Level 4	element 4

The procedure used in this case has some affinity with the calculation of matching scores in traditional clustering methods, but with the crucial additional condition that in the case of Q-discrimination analysis the matching scores are only computed in relation to descriptors that are found to be held in common between pairs of elements after the standard ordinal slicing procedure.

As a final observation in this class, it is interesting to note that for an integer scale ranging between 1 and 5, with 3 as the descriptor-anti-descriptor cut off, whether or not the dataare interval or ordinal has no bearing on the results of a Q-discrimination analysis. This is because the properties that distinguish an ordinal from an interval scale do not come into play in this case.

3. Non-integer (ie continuous) scale, ordinal data. It is likely to be more difficult to be able to define the complement in this case. However, in cases where non-integer scales are warranted, it should be possible either to define some context-specific rule for defining the complement or to manage satisfactorily without it (ie use only descriptors and anti-descriptors). In the latter case the cut-off between descriptor and anti-descriptor could either be the mid-point of the scale, or, where justified, some other value.

A more serious problem in the case of non-integer ordinal data is a magnified version of the problem discussed in class 1 above. To illustrate this, assume a weighted data matrix as follows:

Table 8. More hypothetical data

	d_1	d_2	d_3	d_4
e_1	5.3	4.2	1.1	3.3
e_2	5.4	4.3	0	0
e_3	0	0	2.7	3.2

This gives pairs of weights as follows:

$$\begin{array}{lll} e_1e_2 & (5.3,\ 5.4) & (4.2,\ 4.3) \\ e_1e_3 & (1.1,\ 2.7) & (3.3,\ 3.2) \end{array} \qquad (3)$$

If the data in Table 8 and expression (3) is only ordinal, we cannot discriminate between the relative levels of similarity of e_2 and e_3 with respect to e_1. We have nothing to change the "null hypothesis" that they are at the same level of similarity.

To state this argument more generally, when non-integer ordinal data are used, we may get very little, if any, extra discrimination to add to that already obtainable from the Q-analysis under standard ordinal slicing. To illustrate this in the context of the glasses example worked through above, suppose we marginally change the basic data in Table 1, that is, perform a Q-discrimination analysis on the data given in Table 9.

Table 9. Modified data for describing six glasses

	d_1	d_2	d_3	d_4	d_5	d_6	d_7
e_1	4.9	1.1	3.9	2.1	2.0	2.1	4.9
e_2	4.8	1.2	4.85	4.87	1.2	4.99	1.3
e_3	1.05	1.03	1.01	1.07	1.9	0.99	4.98
e_4	0.98	4.89	2.03	4.01	4.02	0.98	5.1
e_5	0.97	5.2	5.15	4.97	4.87	0.97	5.01
e_6	5.02	3.95	4.95	3.96	4.96	4.97	4.94

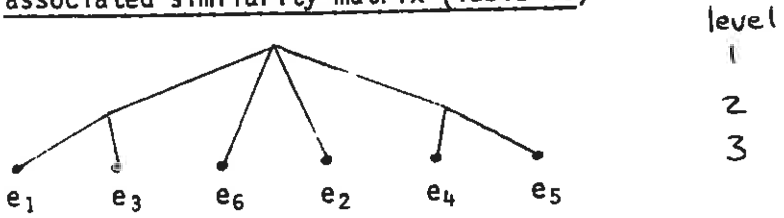
The standard ordinal slicing applied to Table 9, taking 3 as the cut-off between vertex and anti-vertex again gives rise to Table 2 and hence the Q-analysis given in Figure 1. We may now explore (step (iii) of the Q-discrimination analysis) whether the pattern of weights in Table 9 provides grounds for any further discrimination. The matrix (U_{ij}) giving the relative level of similarity of i to j in this case is given in Table 10. **

** Note that in column 1 of Table 10 there is a "tie" between elements 5 and 6 at level 4. It has been decided to specify the next level down as level 5, though it could be argued that level 6 would also be appropriate. A decision on such a matter is best made in the context of individual applications. The decision turns out to be not particularly critical in the present example; it could however, be crucial if a "tie" arose at level 1 or 2.

Table 10. (U_{ij}) for the modified data in Table 7

	1	2	3	4	5	6
1	1	2	2	4	4	4
2	3	1	5	5	5	3
3	2	3	1	3	4	5
4	5	4	3	1	2	3
5	4	3	4	2	1	2
6	4	2	6	3	3	1

Figure 5. Discrimination based on modified data (Table 7) and associated similarity matrix (Table 10)



Discrimination pairs are (2,3) and higher at level 1
(2,2) at level 2
(1,1) at level 3

Its associated pattern of clustering is given in Figure 5. A number of comments are in order. First, it has relatively little structure when compared with Figure 3c, and perhaps more different from that earlier figure than might have been expected from the marginal change in data. Second, whereas in the case of the original integer data the further discrimination in step (iii) led to a refinement of the Q-analysis partitions at two levels (Q = 3 and Q = 4), in the case of the modified data (and Figure 5) only the further discrimination at Q = 4 is suggested.

Thus in the case of ordinal non-integer data there may be little and in some case even no further discrimination over and above that achieved by a Q-analysis based on standard ordinal slicing (ie steps (i) and (ii) only). This raises a question over the significance of non-integer ordinal data. It has been shown above, somewhat paradoxically, that a simple integer scale (1-5) yields a richer structure than a wider continuous scale. While this does not imply that continuous data ought to be forced into integer format in order to end up with a richer structure, it does warrant an awareness of the significance or ordinality, and of the significance of integer/non-integer data.

4. Non-integer (ie. continuous) scale, interval data.** We again encounter the problem over the complement which was described for the previous class, and some similar backstop positions would seem relevant.

However, unlike the previous case, the weights from the original matrix (Table 9) can now be inspected as a more productive means of finding grounds for further discrimination in step (ii). Expression (2) can be used to find the relative differences between elements i and k. The results of these calculations are given in Table 11. Also listed in that table are the number of descriptors held in common by each pair of elements. From Table 11, the relative levels of similarity U_{ij} are readily found. These are given in Table 12.

Table 11. (d_{ik}) The relative differences between elements i and k (with number of shared descriptors in brackets)

	1	2	3	4	5	6
1	0(8)	1.95(5)	2.39(6)	1.32(3)	2.49(4)	1.21(4)
2	1.95(5)	0(8)	8.7(3)	8.6(2)	4.0(3)	1.25(5)
3	2.39(6)	8.7(3)	0(8)	1.22(5)	1.3(4)	4.0(2)
4	1.32(3)	8.6(2)	1.22(5)	0(8)	2.23(7)	2.09(5)
5	2.49(4)	4.0(3)	1.3(4)	2.23(7)	0(8)	2.62(6)
6	1.21(4)	1.25(5)	4.0(2)	2.09(5)	2.62(6)	0(8)

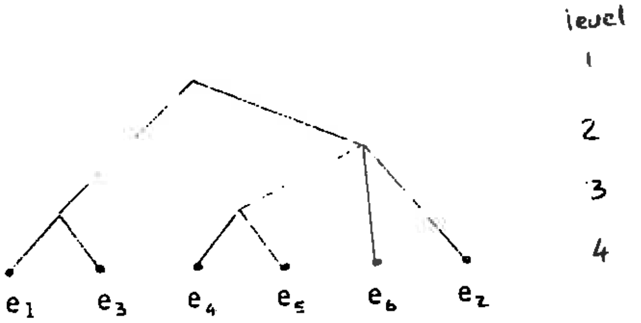
Table 12. (U_{ij}) The relative level of similarity of j to i

	1	2	3	4	5	6
1	1	3	2	5	5	5
2	3	1	5	6	6	3
3	2	5	1	3	4	6
4	6	6	3	1	2	4
5	5	4	4	2	1	2
6	4	2	6	4	3	1

On the basis of the information given in Table 12, we may further refine the partitions of Figure 1. This is given in Figure 6. Although not the same as Table 5, the data in Table 12 lead to an identical pattern of clusters (compare Figure 2 and Figure 6.

** There is a sense in which the distinction being made in this paper between integer and non-integer data is false, since a non-integer scale can generally

Figure 6.



Discrimination pairs are (3,3) and above at level 1
(2,3) at level 2
(2,2) at level 3
(1,1) at level 4

As a final comment in this section it may be useful to point out that there are a number of alternatives to the "matching scores" - type calculation given in expression (2) on which the extra discrimination in step (iii) could be based. Three alternatives are given in Table 13, following Duran and Odell (1974) p.3, expression (2) above corresponding to what those authors call the l_1 norm method.

Table 13. Some alternative "distance" functions

	Name	
1.	Euclidian	$d_{ik} = [\sum_j (M_{ij} - M_{kj})^2]^{\frac{1}{2}}$
2.	l_1 norm	$d_{ik} = [\sum_j \lvert M_{ij} - M_{kj} \rvert]$
3.	Sup-norm	$d_{ik} = \sup_j \{ \lvert M_{ij} - M_{kj} \rvert \}$
4.	l_p norm	$d_{ik} = [\sum_{j=1}^{p} \lvert M_{ij} - M_{kj} \rvert^p]^{1/p}$

In a broader view still, it would now seem to be necessary to relate more closely the powers both of Q-analysis and of Q-discrimination analysis to the existing range of clustering methods already in the literature.

Footnote continued from page 19

be converted to an integer scale by multiplying by 10^n, for some value of n. The distinction is made in the present paper in order to focus on comparisons between integer and non-integer scales of a similar length to each other.

CONCLUSIONS

Although Q-analysis is not primarily a clustering device Johnson (1982) has provided a useful refinement of the basic Q-analysis method for use in clustering and discrimination work. As shown above the three steps involved are: (i) standard ordinal slicing;(ii) Q-analysis on the basis of (i); (iii) further discrimination on the basis of known weights. A number of advantages over traditional clustering methods have been indicated. Although applicable in principle to any weighted matrix (M_{ij}) relating descriptors d_j to elements e_i with weight M_{ij} (as long as the values of (M_{ij}) can be ranked relative to one another, ie have an order), the performance of the method (specifically, the additional refinement possible in step (iii)) is likely to vary depending on the broad nature of (M_{ij}). In particular, the following properties, which were not apparent in Johnson's (1982) initial exposition, have been found.

1. Integer scale with values between 1 and 5: (a) in this case, step (iii) can be considerably simplified, as shown. (b) any distinction between ordinal and interval data will not affect the results of the algorithm.*

2. Integer scale with values between 1 and n: given ordinal data much of the apparent increase in information cannot be productively used; we may, therefore, query whether long scales are warranted if ordinality is assumed. A shorter scale between, say, 1-5 may do just as well.

3. Integer scale, interval data: the procedure for getting extra discrimination can be readily computed, and Q-discrimination analysis turns out to be an enhanced "matching scores" method (see also Table 13).

4. Non-integer scale, ordinal data: there are very few grounds for further discrimination over a Q-analysis based on standard ordinal slicing, much fewer than in the case of a 1-5 integer scale. Again, the apparent increase in information given in the richer scale of weights paradoxically cannot be exploited. This again raises a question of the fundamental compatibility of continuity with ordinality.

5. Non-integer scale, interval data: extra discrimination in step (iii) is now encountered, and the procedure is readily computed. In

* if 3 is the cut-off between descriptors and anti-descriptors.

relation to the previous point, it is seen that whether the data are assumed to be interval or ordinal is crucial to the results.

6. The definition of the complement (and hence any descriptor - anti-descriptor cut-off) may be more problematic with non-integer than with integer data. However, there may be ready solutions to this problem, defined essentially by the meaning of the original scale values.

Finally, in looking into the above explorations, three additional questions have come to mind. First, is the finer discrimination offered in step (iii) always compatible with the clustering that arises from step (ii)? If so, can this be proved? If not, a case may arise where the results of step (iii) were incompatible with those of step (ii), and could not be juxtaposed. Second, given that steps (ii) and (iii) each produce well-defined clustering patterns, why not accept one or other of them in its own right, rather than amalgamating them providing the choice can be justified in particular contexts? Third, is it necessarily desirable to alter the basic Q-analysis partitions? The aim of the basic Q-analysis algorithm is to identify a connectivity structure in data at various dimensional levels: tunnels and spaces whose breadth and configuration constrain the movement of so-called traffic. The standard ordinal slicing (step (i)) is fully compatible with such an aim. However, the finer discrimination in step (iii) dissects the natural tunnels that the Q-analysis has found. In some cases this dissection may be artificial because it separates or groups elements according to whether the weightings on the descriptors they share are different or similar. However, the existence of tunnels (via q-connected components) should depend on whether or not elements are strongly or weakly linked. Thus in step (iii) elements which are linked relatively strongly but with different weights might be separated, whereas other elements weakly linked with the same weights could stay together. This appears to be counter to some aspects of the theory of Q-analysis in the sense that the latter (remaining) linkage is in a sense weaker and therefore less able to support certain types of traffic than the former (now removed) linkage. The extent to which this argument about traffic applies must depend, of course, on what is meant by traffic. If it is taken as some (relatively) dynamic entity that can only exist within the spaces of latent structures, then dissection of these structures may not be warranted. If however, it is interpreted more generally as any graded pattern that can be supported by some given structure (Johnson, 1981) then the weights $\{M_{ij}\}$ become what Johnson (1982) appropriately calls discrimination traffic, and may dissect the original structure.

In some cases, then, there may be grounds for performing only steps (i) and (ii), and not the full discrimination analysis. If so, as in the case of basic Q-analysis, it is likely to be worth exploring the different structures (pattern of clusters) that are revealed from choosing different slicing parameters. The richness of results from such explorations is likely to depend on the richness (or otherwise) of the original set of scale values.

ACKNOWLEDGEMENT

I would like to thank Jeff Johnson for discussing this paper with me.

REFERENCES

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

Atkin, R.H. (1977) *Combinataural Connectivities in Social Systems,* Birkhauser, Basel.

Atkin, R.H. (1981) *Multi-dimensional man,* Penguin.

Beaumont, J.R. and Gattrel, A.C. (1982) *CATMOG 34 : An introduction to Q-analysis,* Geo. Abstracts, Norwich, England.

Chapman, G.P. (1981) Q-analysis, in N.Wrigley and R.J. Bennett (eds.) *Quantitative Geography : a British view,* Routledge and Kegan Paul.

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

Everitt, B. (1974) *Cluster analysis,* Heineman.

Gould, P. (1980) Q-analysis : an introduction for social scientists, geographers, and planners, *International Journal of Man Machine Studies 12,* pp. 169-199.

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

Johnson, J.H. (1982) Q-discrimination analysis, *Environment and Planning B, 8*(4), pp. 419-434.

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

Macgill, S.M. (1982b) Exploring the similarities of different risks, Working Paper 344, School of Geography, University of Leeds.

APPENDIX : PERFORMING THE Q-ANALYSIS

The standard Q-analysis algorithm clusters elements according to the number of descriptors they have in common with each other. The starting point is a binary matrix such as that given in Table 2, from which a so-called shared face matrix can be derived depicting how many descriptors element e_i has in common with element e_j. The shared face matrix (X_{ij}), say, for the present example is given as follows:

$$\begin{array}{ccccccc}
 & e_1 & e_2 & e_3 & e_4 & e_5 & e_6 \\
e_1 & 7 & 4 & 5 & 2 & 3 & 3 \\
e_2 & 4 & 7 & 2 & 1 & 2 & 4 \\
e_3 & 5 & 2 & 7 & 4 & 3 & 1 \\
e_4 & 2 & 1 & 4 & 7 & 6 & 4 \\
e_5 & 3 & 2 & 3 & 6 & 7 & 5 \\
e_6 & 3 & 4 & 1 & 4 & 5 & 7
\end{array}$$

Note that it is (necessarily) symmetric about the leading diagonal, and that the value of X_{ij} is in fact given by the number of descriptors shared by element i and element j, minus one. The reason for subtracting one is due to the fundamental pre-occupation of Q-analysis with the dimension of latent structures; a two dimensional structure being represented by three vertices, a three dimensional structure by four vertices, an n-dimensional structure by n + 1 vertices, and so on.

Now if glass or element e_i is linked to glass e_j through sharing m descriptors and glass e_j is linked to glass e_k through sharing n descriptors (and $m \leqslant n$), then i is linked to k by m descriptors. The purpose of the basic Q-analysis algorithm is to find clusters within which such "chains of connection" at each successive dimensional level arise. This can be done by examining the shared face matrix, and reading from the diagonal, either upwards and to the left or to the right and downwards, looking on each "round" for successive integer values of q in descending order and grouping elements together at a given level of whenever they share at least q descriptors. The results in this case turn out to be:

$q = 6$	$(e_1)(e_2)(e_3)(e_4)(e_5)(e_6)$
$q = 5$	$(e_1)(e_2)(e_3)(e_4e_5)(e_6)$
$q = 4$	$(e_1e_3)(e_4e_5e_6)(e_2)$
$q = 3$	$(e_1e_2e_3e_4e_5e_6)$

16

Produced by
School of Geography
University of Leeds
Leeds LS2 9JT
From whom copies may be obtained