An Optimising Approach to Alternative Clustering Schemes

Alan Jessop
Durham Business School, Durham University, Mill Hill Lane, Durham, DH1 3LB, UK (e-mail: a.t.jessop@durham.ac.uk)

Abstract

Clustering objects into groups is usually done using a statistical heuristic or an optimisation. The method depends on the size of the problem and its purpose. There may exist a number of partitions which do not differ significantly but some of which may be preferable (or equally good) when aspects of the problem not formally contained in the model are considered in the interpretation of the result. To decide between a number of good partitions they must first be enumerated and this may be done by using a number of different heuristics. In this paper an alternative method is described which uses an integer linear programming model having the number and size distribution of groups as objectives and the criteria for group membership as constraints.

The model is applied to three problems each having a different measure of dissimilarity between objects and so different membership criteria. In each case a number of optimal solutions are found and expressed in two parts: a core of groups, the membership of which does not change, and the remaining objects which augment the core. The core is found to contain over three quarters of the objects and so provides a stable base for cluster definition.

.

JEL Classification: C61

Keywords: ILP; Multicriteria; Statistics; Cluster

1 Introduction

There are many circumstances in which objects are partitioned into groups: task partitioning, the analysis of social relations, the definition of taxa, and others. Objects which are similar or close will be in the same group. The necessary pairwise relations may be given directly, as when pairs of people in a society are defined as being linked in some way, or each object may be described according to a number of attributes and from these data a measure of the dissimilarity between each pair is calculated. In these latter problems the groups are usually called clusters. For the development of our model we consider that a grouping scheme will have four constituents:

d: a measure of pairwise dissimilarity
c: a criterion for group membership (based on *d*)
n: the number of groups
u: the unevenness of the size distribution of groups

There are a number of measures of dissimilarity (eg. Everitt 1993). Some measure a distance between a pair of objects based on a number of attributes possessed by each; Euclidean distance, for instance. Other measures describe the distance between pairs of groups, the mean of the pairwise Euclidean distances. The purpose is to have some metric such that the larger the value the less similar are the pair and so the less the justification for including them in the same group. This allows object pairs to be defined as either too dissimilar to be grouped or sufficiently similar to form candidate groups from which a partition may be formed.

It is common that in cluster formation a hierarchical heuristic is used in which, for instance, objects join the nearest cluster until all have been allocated. The process of cluster formation is

displayed as a tree showing the level of dissimilarity at which clusters are formed. In the light of contextual or theoretical considerations this tree is inspected and a decision made as to the most appropriate partition. Once *d* has been chosen *c*, *n* and *u* are considered together in interpreting the result. This may be a cognitively difficult. The result is a single partition. Some form of sensitivity testing, by choosing different measures for *d* say, or by using more than one method to form clusters (Kaufman and Rousseeuw 1990) is recommended as a way of exploring alternatives. Cluster formation is, in this broad sense, interactive.

Task variety may be reduced by pre-setting the number of groups to $n=k$ (the *k*-means method). There may be no strong reason for preferring a particular *k* so that while this method reduces the number of factors to be considered in interpreting results it requires a corresponding increase in sensitivity testing to see the effect of different values of *k*. Nonetheless, structuring the analysis in this way offers the prospect of easier interpretation of results.

Optimization has been intermittently proposed for clustering problems using aggregate dissimilarity as an objective. For example, we may minimise the sum of squares of intra-cluster dissimilarities given appropriate constraints on the number of groups or their size. Clusters are formed by the use of mathematical programming methods (Rao 1971; Hansen and Jaumard 1997), notably linear programming (Vinod 1969; Joseph and Bryson 1998) and dynamic programming (Jensen 1969; van Osulman 2004). A considerable benefit of optimising approaches is that they have a clear criterion which allows for an assessment of how good optimal and other partitions are (Li 2006). This helps the evaluation of alternative partitions. These optimising methods work well for small and medium sized problems but they become infeasible for very large problems for which other methods such as the application of genetic algorithms are available (Murthy and Chowdhury 1996; Cowgill, Harvey and Watson 1999; Maulik and Bandyopadhyay 2000; Chiou and Lan 2001). However, most published applications of cluster analysis are of moderate size (Kettenring 2006) so that optimization methods may find wide application.

In this paper it is proposed that a nested set of integer linear programmes (ILPs) having as their objectives the number and size distribution of groups (*n* and *u* above) offers a method of aiding cluster definition which is easily comprehended and so forms a good basis for interaction. First, the method is described in outline and a measure characteristic of the size distribution is chosen. Next, the optimisation models are set out and applied to three illustrative cases. Finally, the results are discussed.

2 Method

When the number of groups is fixed in advance (as with the *k*-means method) the distribution of objects into groups, and so the distribution of group sizes, is found as a result. If there is no compelling reason for specifying some particular number of groups then parsimony of description requires that the minimum number of groups be chosen. When the number of groups satisfactory for a particular problem is not obvious using the minimum number of groups is likely to be a good start.

To identify and distinguish between the resulting partitions some judgement about the relative sizes of groups may be made: are groups of similar size to be preferred or is it better to identify large groups, if they exist, with the corresponding reduction in size of the remainder? Given a measure of the unevenness of the distribution of group sizes an optimum profile can be found.

It is probable that there is more than one allocation of objects to groups which has this optimal profile and these must be found. The ability of an optimising approach easily to generate a number of optimal and near-optimal solutions allows the enumeration of alternatives with known properties.

Combining these ideas, the outline method is:

step 1: set criteria for group membership
step 2: minimise the number of groups given membership constraints
step 3: optimise the size distribution given constraints of group membership and the optimal number of groups
step 4: find several optimal (and near-optimal) partitions
step 5: interpret and return to step 1 if necessary

To implement this scheme requires a measure of the unevenness of a distribution of group sizes for use in step 3.

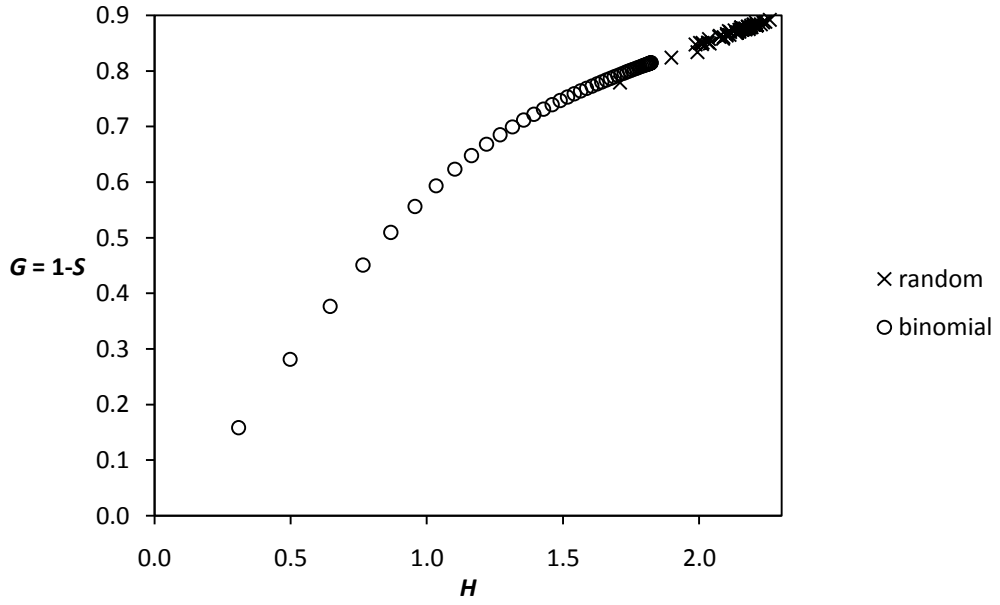
3 Measures of Evenness

In a partition m objects are allocated to n groups, the proportion in each group being p_i (i = 1,2…n). The evenness of this distribution may be measured in a number of ways, for instance by Shannon entropy $H = -\sum_i p_i \log(p_i)$ (Shannon 1948) or the Simpson index $S = \sum_i p_i^2$ (Simpson 1949). The measure $G = 1\text{-}S$ is the Gini index, popular as a descriptor of income inequality. Both H and G have maxima ($H_{max} = \log(n)$ and $G_{max} = (1 - 1/n)$) when the distribution is even ($p_i = 1/n$) and a minimum of zero when some p_i is 1. Though there are other measures of unevenness (Hill 1973) we restrict comparison to these two. Hill (1973) gives a general index $I_a = [\sum_i p_i^a]^{1/(1-a)}$ of which three cases are: $I_0 = n$; $I_1 = \exp(H)$; $I_2 = 1/S$. In this sense there is a formal relation between H and S. The logarithm of I_a is the entropy measure due to Renyi (1961). Other entropy measures and their relations to H and S are given in Mayoral (1998), Patil and Taillie (1982), Rao (1982) and Peet (1974).

There are two considerations which are often used in differentiating H from S: disaggregation properties (Theil 1967) and sampling effects (Magurran 1988; Pielou 1977; Lande 1996; Lande et al 1996; Keylock 2005). Neither is relevant in this application but it may still be the case that solutions depend on whether H or S is used to measure evenness.

In considering the construction of classification trees Breiman et al (1998) used both H and S among other measures but found the result to be relatively insensitive to which was used. To illustrate this we generate two sets of data for n = 10. First, fifty randomly selected probability distributions and, second, binomial distributions with parameter values from 0.01. The graph of $G = 1\text{-}S$ against H (Figure 1) shows that the correspondence between these two measures of unevenness is very high, certainly one is an excellent proxy for the other. We use the simpler measure S in what follows.

Figure 1. Relation between alternative measures of evenness.



4 Optimal Groups

It will be easier to deal with the number in each group, m_i , rather than the proportion and so $S = \sum_i m_i^2$ and $m = \sum_i m_i$ will be used in the optimisation. S may be maximised (lumpy size distribution preferred) or minimised (even distribution preferred). In what follows S is maximised to identify large groups if they exist.

Following Jessop (2003) and Proll (2007), Jessop et al (2007) discuss approaches for this sort of optimization and describe two methods useful for moderate sized problems. It was found that if the network density (the proportion of all inter-object relations classed as similar rather than dissimilar) was not too high then an approach based on the enumeration of groups is feasible. Problems with between 140 and 350 objects and with network densities up to 0.14 were solved in this way. In the networks tested size and density were negatively correlated indicating that for smaller problems higher density networks could be solved. An alternative formulation was successful for higher density networks of 100 objects (see also Jessop 2009).

In this paper the method based on enumeration is used. If all possible groups are enumerated it is simple to find a set which optimises S. The usual difficulty with enumeration is that the number of groups is too large for this strategy to be practicable, though this depends on network density. However, in Step 1 criteria for group membership are set, so that only feasible groups need be enumerated. For a wide class of problems this renders enumeration a viable method. During the enumeration the size of each candidate group and so of m_i^2 is retained.

The enumeration ensures that membership criteria are met and that the resulting candidate groups are listed. These are described by a binary matrix, $\mathbf{X}$, in which $x_{ij} = 1$ if object j is in group i.

Candidate groups are selected via the binary vector $\mathbf{\Lambda}$ in which $\lambda_i = 1$ if group i is chosen.

The first ILP (Step 2 above) selects the smallest number of groups such that each object appears in exactly one group:

$$\begin{aligned} \textit{Model 1}: \quad & \min \ \sum_i \lambda_i = n_{min} \\ & \text{s.t.} \ \ \sum_i x_{ij}\lambda_i = 1 \quad ; \forall j \end{aligned}$$

Since values of m_i^2 will have already been found in the enumeration selecting an optimal configuration is also an ILP:

$$\begin{aligned} \textit{Model 2}: \quad & \max S = \sum_i m_i^2 \\ & \text{s.t.} \ \ \sum_i x_{ij}\lambda_i = 1 \quad ; \forall j \\ & \qquad \sum_i \lambda_i = n_{min} \end{aligned}$$

Should there be any other requirements – minimum or maximum group sizes, for instance – these can easily be incorporated either as constraints in the ILPs or in the enumeration.

Other solutions are easily found (Step 4) by ensuring that the number of candidate groups shared between the current solution and any previous solution is less than n_{min}. If $\mathbf{\Lambda}^1$ is the first solution found, $\mathbf{\Lambda}^2$ the second, and so on, then the kth solution is found by augmenting *Model 2*:

$$\begin{aligned} \textit{Model 3}: \quad & \max S = \sum_i m_i^2 \\ & \text{s.t.} \ \ \sum_i x_{ij}\lambda^k{}_i = 1 \quad ; \forall j \\ & \qquad \sum_i \lambda^k{}_i = n_{min} \\ & \qquad \sum_i \lambda^k{}_i \lambda^a{}_i < n_{min} \ ; \ \text{a} = 1 \ldots \text{k-1} \end{aligned}$$

The first few solutions are likely to be alternative optima with increasingly sub-optimal solutions following.

The method is now applied to three cases, each illustrating a different idea of group membership.

5 Illustrative Applications

5.1 Design Network

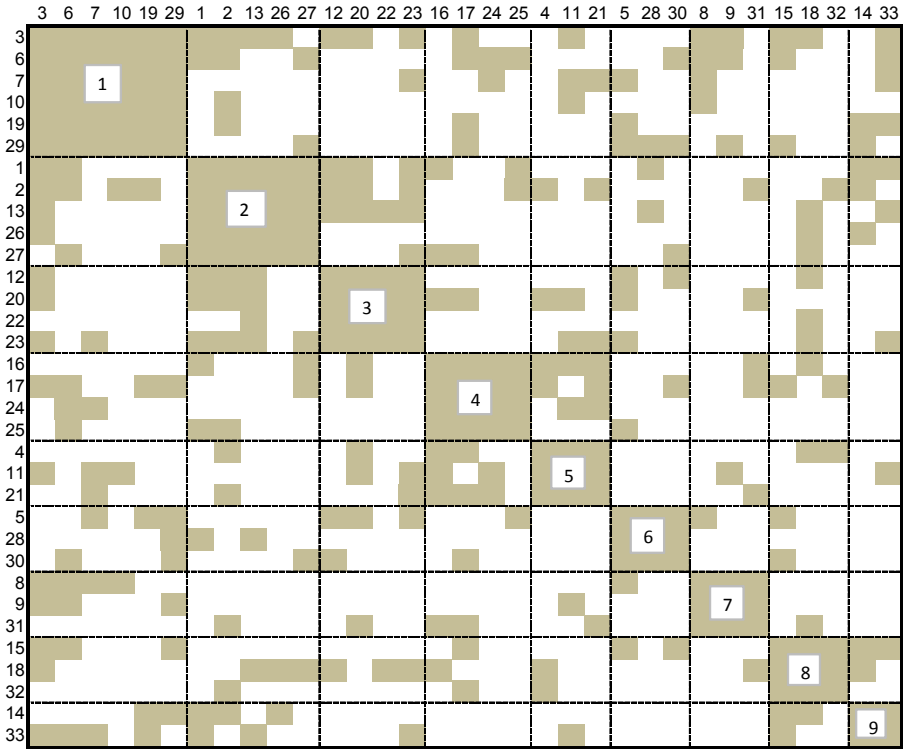
A large number of problems can be described by the binary relations between objects. Chermayeff and Alexander (1966) described a design problem in which thirty three design requirements for an urban area are defined. Requirements were such as *arrangements to protect the dwelling from local noise* and *safe and pleasant walking and wheeling surfaces*. Two requirements interact if adopting different solutions for meeting one affects the solutions used in meeting another. The pair are dissimilar if they do not interact for we wish to form groups (design tasks) of interacting requirements both to give coherent sub-problems and to minimise inter-task dependency.

The pairwise interactions form arcs in a network where $l_{ij} = 1$ if nodes are linked and 0 otherwise and $l_{ii} = 1$. This Chermayeff and Alexander network has $L = \sum_i \sum_j l_{ij} = 373$ links and so a network density $D = L/m^2 = 0.34$.

The criterion for group membership is that all groups are maximally connected, ie. all group members are connected to each other. This means that inter-group interactions are minimised so that sub-problems (groups) are as independent as possible.

There are 639 candidate groups ranging in size from 1 to 7. Using models 1 and 2 gives $n_{min} = 9$, $S_{max} = 133$ and $S_{min} = 123$. For S_{max} the distribution of group sizes is [6,5,4,4,3,3,3,3,2] and for S_{min} is [4,4,4,4,4,4,3,3,3]. A measure of the adequacy of the grouping is the proportion of links contained in groups (Gershenson et al 2004): $A = S_{max}/L = 0.36$.

Figure 2. Design. An optimal partition: $m = 33$; $L = 373$; $S_{max} = 133$; $D = 0.34$; $A = 0.36$. (This solution is no. 3 in Table 1.)



Using model 3 gives 12 S_{max} groups. Figure 2 shows one of these S_{max} groupings. Rows and columns are indexed in accord with the original presentation of the problem. Each shaded square represents a link. The nine maximally connected groups are shown on the diagonal in order of decreasing size. The figure confirms the relatively low value of A in that much of the interaction is between groups rather than being contained within them, an unsatisfactory partition. However, some further aggregation is possible either by relaxing the maximum density constraint in the enumeration of groups or by joining some of the groups which have a high inter-group density, groups 4 and 5 in Figure 2, for instance.

Table 1. Optimal partitions for the design problem.

		solutions											
	core	1	2	3	4	5	6	7	8	9	10	11	12
a	[3,6,7,10,19,29]												
b	[5,28,30]												
c	[8,9,31]												
d	[12,20,22,23]							13	13	13	13	13	13
e	[1,2,26,27]	13	13	13	13	13	13						
f	[17,25]	16,24	16,24	16,24	24	24		16,24	16,24	16,24	24	24	
g	[14,33]	15	15		15		15	15	15		15		15
h	[11,21]		4	4	4,16	4,16	16,24		4	4	4,16	4,16	16,24
i	[18,32]	4		15		15	4	4		15		15	4

28 of the 33 objects appear in the same groups in all twelve solutions. In Table 1 these groupings are shown as the core at the left of the table. Each subsequent column shows a solution. A blank entry means that that the core group appears as shown. If numbers are entered then the core group is augmented. For example, solution 1 consists of these groups:

a	[3,6,7,10,19,29]
b	[5,28,30]
c	[8,9,31]
d	[12,20,22,23]
e & 13	[1,2,13,26,27]
f&16,24	[16,17,24,25]
g&15	[14,15,33]
h	[11,21]
i&4	[4,18,32]

The structure of the set of optimal solutions is clear. Object 13 may be attached either to core d or e and this divides the 12 solutions in two. Each of the sets of six solutions are subdivided in the same way depending on the allocation of the remaining four objects 4, 15, 16, 24.

Identifying the core, and so the non-core, aids the final decision as to which partition should be adopted. In this case it is clear that the decision is about the disposition of the 5 non-core tasks, a simpler problem than considering all 33. The larger the core the more similar are the alternative optima. The proportion of objects in the core $C = 28/33 = 0.85$ indicates this.

5.2 Airport Performance

Data on the operating and financial performance of twenty five UK airports were taken from Cruicksahnk et al (2004). The airports were:

1	Heathrow	14	Belfast International
2	Gatwick	15	Cardiff International
3	Stansted	16	London Luton
4	Southampton	17	Blackpool
5	Glasgow	18	Bristol
6	Edinburgh	19	Exeter
7	Aberdeen	20	Liverpool
8	Manchester	21	London Biggin Hill
9	Bournemouth	22	London City
10	Humberside	23	Norwich
11	Nottingham East Midlands	24	Southend
12	Birmingham International	25	Teesside

13 Newcastle

From the data seven performance measures were calculated for each airport:

1. proportion of international passengers
2. proportion of charter passengers
3. number of passengers / employee
4. number of passengers / air traffic movement
5. commercial revenue / total revenue
6. commercial revenue / passenger
7. profit (after interest and tax) / revenue

These are measures of the type used to assess airport performance (Graham 2003). In benchmarking and performance analysis it is helpful to group these airports according to their performance profile, the extent to which values on the seven measures are similar.

In dealing with multicriteria problems it is common to scale incommensurable variables to some common metric. Scaling each variable to have a mean 0 and variance 1 is popular for the preparation of tables of performance measures (for instance, the rankings made by the *Financial Times* of MBA and other programmes) and is also used in cluster analysis (Hair et al 2006). While any such transformation will be to some extent a matter of convenience we choose this because it is representative of practice.

The similarity between pairs of airports is measured by the correlation, r, of the transformed variables and so 1-r is the measure of dissimilarity between them.

We use a simple criterion for group membership, that all objects in a group should be similar to some minimum extent: $1-r < \alpha$. For illustration we use $\alpha = 0.5$ and find that $n_{min} = 10$, $S_{max} = 81$ and $S_{min} = 69$. For S_{max} the distribution of group sizes is [5,4,4,3,2,2,2,1,1,1] and for S_{min} is [4,3,3,3,3,2,2,2,2,1].

Figure 3. Airports. An optimal partition: $m = 25$; $L = 121$; $S_{max} = 81$; $D = 0.19$; $A = 0.67$. (Solution 2 in Table 2.)

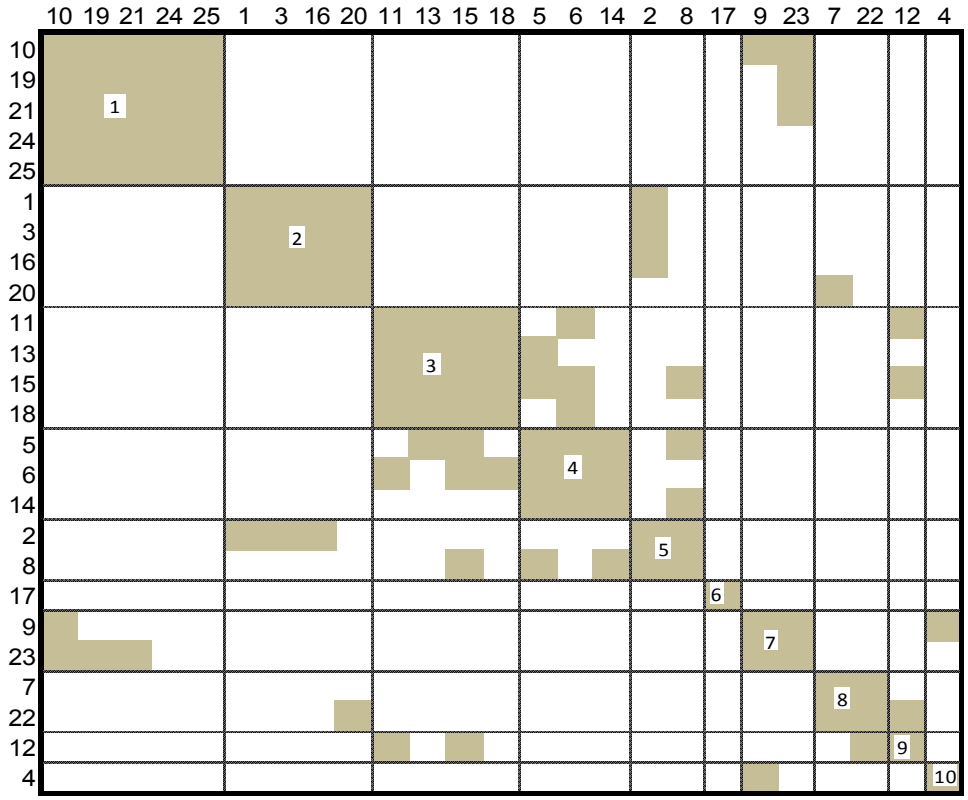


Table 2. Optimal partitions for the airport problem.

	core	solutions 1	2	3	4	max 1-*r*
a	[10,19,21,24,25]					0.48
b	[1,3,16,20]					0.25
c	[11,13,15,18]					0.40
d	[5,6,14]					0.32
e	[2,8]					0.38
f	[17]					n/a
g		9,23	9,23	4,9	4,9	
h		12,22	7,22	12,22	7,22	
i		4	4	7	12	
j		7	12	23	23	
	(1-*r*) for row g	0.17	0.17	0.48	0.48	
	(1-*r*) for row h	0.36	0.23	0.36	0.23	

An optimal result is shown in Figure 3 and all four optimal solutions in Table 2. Unlike the design case the core groups appear unaugmented in all four optimal partitions and account for 19 airports. The remaining six airports appear in various combinations (two pairs and two singletons) as shown in rows g to j. In the design case all that could be done was to present alternative solutions as an aid to a final partitioning. In this case a further guide is possible by examining the maximum values of 1-*r* for each group. The values for the core are shown at the right, but since the core groups are unaugmented these maxima do not help to differentiate between solutions. The last two rows show 1-*r* values for the non-core groups. If there are no other considerations it makes sense to prefer the solution for which these values are smallest, in this case solution 2.

5.3 MBA Programmes

Paucar-Caceres and Thorpe (2005) analysed the structure of 32 full-time UK MBA programmes according to the subjects covered in core and elective courses. The MBAs were offered at the following business schools:

1	Aberdeen	12	De Montfort	23	London
2	Ashridge	13	Durham	24	Manchester
3	Aston	14	Edinburgh	25	Middlesex
4	Bath	15	Exeter	26	Newcastle
5	Birmingham	16	Glasgow	27	Nottingham
6	Bradford	17	Henley	28	Oxford
7	Bristol	18	Imperial	29	Strathclyde
8	Brunel	19	Kingston	30	Wales
9	Cambridge	20	Lancaster	31	Warwick
10	City	21	Leeds	32	Westminster
11	Cranfield	22	Leicester		

and the core modules were

1	e-business	11	management development
2	entrepreneurship	12	microeconomics
3	ethics	13	management information systems
4	finance	14	marketing management
5	financial accounting	15	management science

6	general management	16	operations management
7	human resource management	17	organisational behaviour
8	international business	18	project management
9	macroeconomics	19	statistics
10	management accounting	20	strategy

A features matrix **Y** encodes which of the 20 core subjects were offered at each programme: y_{ij} = 1 if programme i has core subject j and zero otherwise. Programmes varied, offering between 3 and 15 of the twenty subjects. Programmes are similar to the extent that they offer the same modules, specifically that group members should all share some minimum number, θ, of core subjects.

With $\theta = 5$ the results of the optimisation are: $n_{min} = 15$, $S_{max} = 170$ and $S_{min} = 82$; for S_{max} the distribution of group sizes is [11,5,3,2,1,1,1,1,1,1,1,1,1,1,1] and for S_{min} is [4,3,3,3,3,3,2,2,2,2,1,1,1,1,1]. 29 of the 32 programmes are in the core, as shown in Table 3.

The result is also shown in Figure 4. Figures 2 and 3 showed links between similar objects. Figure 4, on the other hand, shows which modules are offered by each programme. The rows (programmes) are ordered in the blocks shown in Table 3. The columns (modules) are ordered to emphasise the extent to which programmes have modules in common.

It is no surprise that a large fraction of programmes are clustered together because they share some core modules, the largest group having a focus on operations and strategy and the next largest a focus on softer topics such as management development and ethics. The large number of singletons is at first more surprising but the diagram shows the sparseness of shared features. Of the ten singletons (rows c to m in Table 3) five possess fewer than $\theta = 5$ of the core modules, two have exactly 5, two have 6 and the remaining one has 7. The presence of the ten singletons is a result of setting $\theta = 5$; different values could be tried.

Figure 4. Features of MBA core programmes.

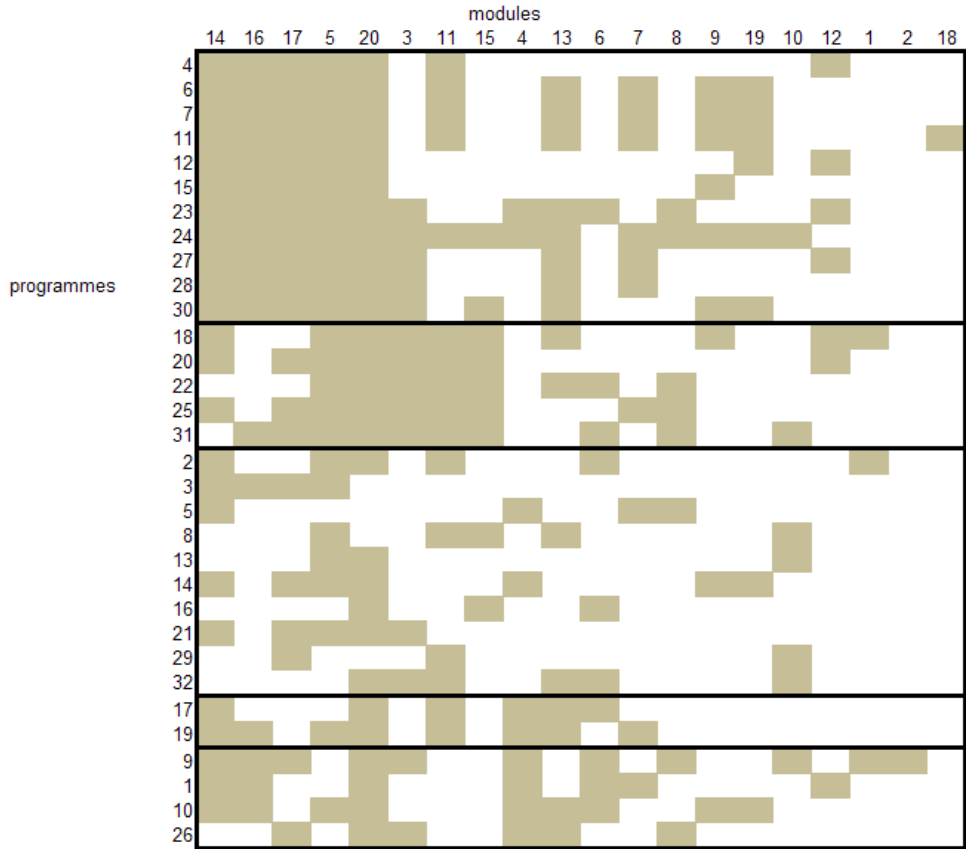


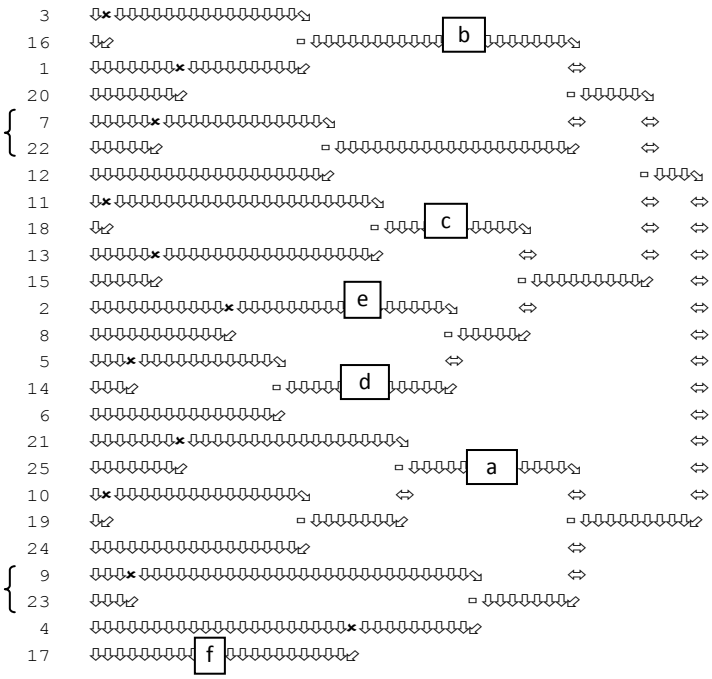
Table 3. Optimal partitions for the MBA problem.

	core	solutions 1	2	3
a	[4,6,7,11,12,15,23,24,27,28,30]			
b	[18,20,22,25,31]			
c	[2]			
d	[3]			
e	[5]			
f	[8]			
h	[13]			
i	[14]			
j	[16]			
k	[21]			
l	[29]			
m	[32]			
n	[17,19]		10	10
o	[9]	1,10	26	1
p		26	1	26

6 Comparison with Other Clustering Algorithms

While the method used here is different from the usual heuristic clustering algorithms it is of interest to compare the results found by both. For this the Airport data are used with *r* as the distance measure. Six algorithms were used. One of these, centroid clustering, gave the same results as that recommended above, Figure 3 and solution 2 in Table 2. The dendrogram for the centroid clustering is shown in Figure 5. The core groups, a – e, are indicated on the tree and the non-core pairs, (9,23) and (7,22) shown by braces to the left of the diagram.

Figure 5. Dendrogram for centroid clustering



Three of the other methods – complete linkage, single linkage and average linkage (between groups) – gave similar results, one of which is shown as Figure 6. The differences are that group e splits with airport 2 joining b and airport 8 joining d. Airport 6 is detached from the original group d. The result is clusters (c+d+8) and (b+2) together with a and f. Group e disappears.

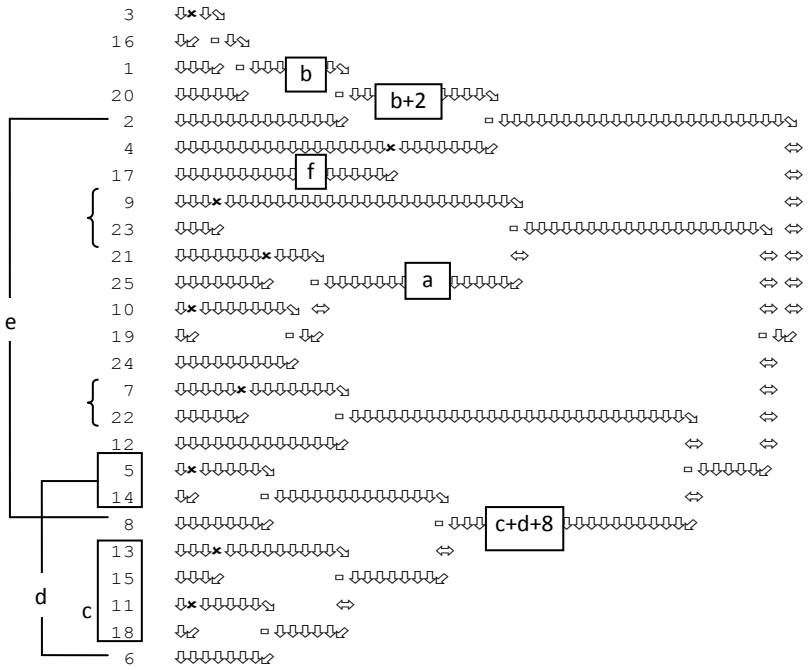
This rearrangement is not surprising given the links shown in Figure 3: groups c and d have a high inter-group density of connection; airport 8 is linked to two of the three members of d and one of the four of c; airport 2 is linked to three of the four members of b.

The results from the other two methods – median and average distance (within groups) – have the additional aggregation of airports 9 and 23 with group a earlier in the aggregation process than shown in Figure 6. Again, this is consistent with the links shown in Figure 3.

Figure 3 shows not just the groups but also the inter-group links. For common problems of moderate size, such as those shown here, high overlaps can easily be seen and decisions on reallocation taken. For example, airport 2 is Gatwick which is linked with the three other London airports in b as well as with the largest airport outside London, Manchester, in e. Similarly, group d has the main Scottish airports, Glasgow and Edinburgh, plus Belfast. These might be seen as similar to the larger regional airports in group c or, perhaps, with Cardiff airport (number 15) might be part of d, a group of British but not English airports.

Resolution and final allocation depend on the context and purpose of the analysis. The dendrograms associated with different clustering heuristics give alternative allocations, as discussed above. Diagrams such as Figure 3 show the full set of similarities and are readily suggestive of ambiguities and possible regrouping.

Figure 6. Dendrogram for complete linkage clustering



7 Discussion

A method is given for generating alternative partitions which uses a dissimilarity metric to impose constraints on group membership, and then uses the number and size distribution of groups as the objectives in ILP optimisations.

It is a criticism sometimes levelled at cluster analysis (eg. Barney and Hoskisson 1990) that it will always produce clusters whatever the data and that because of this does not provide results with much meaning. Specifying group membership criteria as constraints when enumerating candidate groups overcomes this objection, for if the criterion is not met then no groups will be formed. This is seen in the large number of singletons in the MBA problem. This resolution of the inevitability of cluster formation comes at a price, that of having explicitly to specify the criteria of group membership *a priori* rather than making some *post hoc* judgement. The initial specification can, of course, be revised and the model rerun. The contention is that it is more straightforward to use the number and size distribution of groups as objectives in the ILPs and so leave the criteria for group membership as parameters to be altered in sensitivity testing runs of the models. Membership criteria are likely to be more dependent on views taken of the context and purpose of the analysis and so forcing an *a priori* specification (just what do we mean by "similar"?) will encourage a thoughtful engagement with the problem.

Table 4. Summary of the three illustrative cases.

	Illustrative application		
	1. Design network	2. Airports	3. MBAs
no. of objects - m	33	25	32
no. of links - L	373	121	n/a
network density - D	0.34	0.19	n/a
no. of candidate groups	639	118	3861
n_{min}	9	10	15
S_{max}	133	81	170
S_{min}	123	69	82
no. of optimal groups	12	4	3
proportion of links in groups - A	0.36	0.67	n/a
proportion of objects in core - C	0.85	0.76	0.91

Table 4 summarises the three cases. The descriptors permit an appreciation of the feasibility and usefulness of the partition. First, A shows how satisfactory the partition is as a system description. The differences between the design and airport illustrations show this. The usefulness of the partition in the design case is poor but that for the airports is good. Second, the relative size of the core, C, shows the stability or robustness of the optimal partitions. In the three cases C is quite high (at least 0.76) showing that decisions between alternative optimal partitions is a matter of considering only the disposition of at most 24% of the objects. Third, the difference between S_{max} and S_{min} is a subsidiary indicator of the extent to which the optimal partitions are superior.

Because the number of objects was not large the optimisation problems were easily solved by enumeration even though the network density was as high as 0.34. This confirms the speculation made in Section 4 of the negative correlation between size and density for problems solvable by this method.

Finding an acceptable partition is a decision problem. The method proposed in this paper emphasises this by enumerating a number of alternative partitions and providing indicators to help comparison between them. In addition, the identification of core groups means that judgement is exercised on the relatively smaller problem of the disposition only of non-core objects. The larger the core the easier this will be and the more robust the result.

References

[1] Barney JB, Hoskisson RE (1990) Strategic groups: untested assertions and research proposals. Managerial and Decision Economics 11, 187–198.

[2] Breiman L, Friedman JH, Olshen RA, Stone CJ (1998) Classification and Regression Trees. CRC, Boca Raton, Ch 4.

[3] Chermayeff S, Alexander C (1966) Community and Privacy: Towards a New Architecture of Humanism. Penguin, Harmondsworth.
[4] Chiou Y-C,Lan LW (2001) Genetic clustering algorithms. European Journal of Operational Research 135, 413–427.
[5] Cowgill MC, Harvey RJ, Watson LT (1999) A genetic algorithm approach to cluster analysis. Computers and Mathematics with Applications 37, 99–108.
[6] Cruicksahank A, Flanagan P, Marchant J (2004) Airport Statistics 2003/2004. University of Bath Centre for the Study of Regulated Industries, Bath.
[7] Everitt BS (1993) Cluster Analysis (Third edn.). Arnold, London, Ch 3.
[8] Gershenson JK, Prasad GJ, Zhang Y (2004) Product modularity: measures and design methods. Journal of Engineering Design 15, 33-51.
[9] Graham A (2003) Managing Airports; An International Perspective, 2nd edn. Elsevier, Oxford.
[10] Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL (2006) Multivariate Data Analysis (Sixth edn.). Pearson Prentice Hall, Upper Saddle River NJ, p579.
[11] Hansen P, Jaumard B (1997) Cluster analysis and mathematical programming. Mathematical Programming 79, 191–215.
[12] Hill MO (1973) Diversity and evenness: a unifying notation and its consequences. Ecology 54, 427–432.
[13] Jensen RE (1969) A dynamic programming algorithm for cluster analysis. Operations Research 17, 1034–1057.
[14] Jessop A (2003) Blockmodels with maximum concentration. European Journal of Operational Research 148, 53–64.
[15] Jessop A (2009) A multicriteria blockmodel for performance assessment. Omega 37, 204–214.
[16] Jessop A, Proll L, Smith BM (2007) Optimal Cliques: applications and solutions. University of Leeds, School of computing, Research Report 2007.03, 2007.
(available via http://www.comp.leeds.ac.uk/research/pubs/reports/2007/2007_03.pdf)
[17] Joseph A, Bryson N (1998) Parametric linear programming and cluster analysis. European Journal of Operational Research 111, 582–588.
[18] Kaufman L, Rousseeuw PJ (1990) Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York, p37.
[19] Kettenring JR (2006) The practice of cluster analysis. Journal of Classification 23, 3–30.
[20] Keylock CJ (2005) Simpson diversity and the Shannon-Weiner index as special cases of generalised entropy. Oikos 109, 203–207.
[21] Lande R (1996) Statistics and partitioning of species diversity, and similarity among multiple communities. Oikos 76, 5–13.
[22] Lande R, Engen S, Sæther B-E (2003) Stochastic Population Dynamics in Ecology and Conservation. Oxford University Press, Oxford, Ch 7.
[23] Li B (2006) A new approach to cluster analysis: the clustering-function-based method. Journal of the Royal Statistical Society B 68, 457–476.
[24] Magurran AE (1988) Ecological Diversity and its Measurement. Princeton University Press, Princeton NJ., Table 4.5.
[25] Maulik U, Bandyopadhyay S (2000) Genetic algorithm-based clustering technique. Pattern Recognition 33, 1455–1465.
[26] Mayoral MM (1998) Renyi's entropy as an index of diversity in simple-stage cluster sampling. Journal of Information Sciences 105, 101–114.
[27] Murthy CA, Chowdhury N (1996) In search of optimal clusters using genetic algorithms. Pattern Recognition Letters 17, 825–832.
[28] Patil GP, Taillie C (1982) Diversity as a concept and its measurement. Journal of the American Statistical Association 77, 548–561.
[29] Paucar-Caceres A, Thorpe R (2005) Mapping the structure of MBA programmes: a comparative study of the structure of accredited AMBA programmes in the United Kingdom. Journal of the Operational Research Society 56, 25–38.
[30] Peet RK (1974) The measurement of species diversity. Annual Review of Ecology and Systematics 5, 285–307.
[31] Pielou EC (1977) Mathematical Ecology. Wiley, New York.
[32] Proll L (2007) ILP Approaches to the blockmodel problem. European Journal of Operational Research 177, 840–850.
[33] Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. Theoretical and Population Biology 21, 24–43.
[34] Rao MR (1971) Cluster analysis and mathematical programming. Journal of the American Statistical Association 66, 622–626.
[35] Renyi A (1961) On the measures of entropy and information. In: Proceeding of the Fourth Berkely Symposium on Mathematics, Statistics and Probability, Volume 1, pp 547–561.

[36] Shannon CE (1948) A mathematical theory of communication. Bell Systems Technical Journal 27, 379–423, 623–656. [Reprinted in Shannon CE, Weaver W (1949) The Mathematical Theory of Communication. University of Illinois Press, Urbana.]
[37] Simpson EH (1949) Measurement of diversity. Nature 163, 688.
[38] Theil H (1967) Economics and Information Theory. North-Holland, Amsterdam, appendices to Chapters 4 and 8.
[39] van Os BJ, Meulman JJ (2004) Improving dynamic programming strategies for partitioning. Journal of Classification 21, 207–230.
[40] Vinod HD (1969) Integer programming and the theory of grouping. Journal of the American Statistical Association 64, 506–519.