CONNECTIONS 17(2):46-52

Displaying Hierarchical Clusters

Linton C. Freeman
Institute for Mathematical Behavioral Sciences, University of California, Irvine

1. Introduction

Visual images are widely used in reporting the results of scientific research (Koestler, 1964; Arnheim, 1970; Taylor, 1971; Tukey, 1972; Klovdahl, 1981). In recognition of that fact, scientists often devote a good deal of effort to the development of visual devices for the display of their models and their observations.

The display of the results of hierarchical clustering is a case in point. Proposals for displaying hierarchical clusters as trees, dendograms, castles, skylines, loops, icicles and shaded matrices have been made (Ward, 1963; Wirth *et al.* 1966, Bertin, 1967; Johnson, 1967; McCammon, 1968; Ling, 1973; Shepard, 1974; Hartigan, 1975; Engelman, 1977; Kleiner and Hartigan 1981; Kruskal and Landwehr, 1983).

Each of these proposed modes of display was developed to facilitate some particular goal. Some are simple to program on a computer. Some can be produced on a conventional line printer and require no special equipment. And some are designed to be easy to interpret; they make it possible to see at a glance which objects join which clusters at what level.

None of these standard modes of display, however, is entirely satisfactory. All but one are guilty of violating Tufte's (1983, p. 87) rule that graphics should not "miss the real news in the data." And the one that is designed to present the news--the shaded matrix--is a good example of what Tufte (1983, p. 15) called a "visual puzzle" or a "crypto-graphical mystery."

This note will propose a way of taking advantage of the power of micro computers to display the results of clustering. Such micro computer displays can provide more information than traditionally has been possible. They can make it simple to see the overall structure of the clusters and they can reveal the details of the correspondence between the data and the clustering model.

2. Clusters and Displays

Hierarchical clustering is a collection of procedures for organizing objects into a nested sequence of partitions on the basis of data on the proximities (or distances) among the objects. Strictly speaking, objects can be clustered if and only if their proximities are ultrametric. But since observed proximity data are seldom if ever strictly ultrametric, the goal of hierarchical cluster analysis is to find a collection of ultrametric proximities that are reasonably close to the observations.

A great many methods for clustering actual data have been proposed (Ward, 1963; Hartigan, 1967; Johnson, 1967; McWhitty, 1967; Lance and Williams, 1967; Gower, 1967; Sibson, 1970; Sneath and Sokal, 1973; D'Andrade, 1978; Weller and Buchholtz, 1986). They all use one or another algorithm to find an ultrametric that approximates the observed data.

		Woman																	
											1	1	1	1	1	1	1	1
		1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8
		-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	1	8	6	7	6	3	4	3	3	3	2	2	2	2	2	1	2	1	1
	2	6	7	6	6	3	4	4	2	3	2	1	1	2	2	2	1	0	0
	3	7	6	8	6	4	4	4	3	4	3	2	2	3	3	2	2	1	1
	4	6	6	6	7	4	4	4	2	3	2	1	1	2	2	2	1	0	0
	5	3	3	4	4	4	2	2	0	2	1	0	0	1	1	1	0	0	0
	6	4	4	4	4	2	4	3	2	2	1	1	1	1	1	1	1	0	0
	7	3	4	4	4	2	3	4	2	3	2	1	1	2	2	2	1	0	0
	8	3	2	3	2	0	2	2	3	2	2	2	2	2	2	1	2	1	1
Woman	9	3	3	4	3	2	2	3	2	4	3	2	2	3	2	2	2	1	1
	10	2	2	3	2	1	1	2	2	3	4	3	3	4	3	3	2	1	1
	11	2	1	2	1	0	1	1	2	2	3	4	4	4	3	3	2	1	1
	12	2	1	2	1	0	1	1	2	2	3	4	6	6	5	3	2	1	1
	13	2	2	3	2	1	1	2	2	3	4	4	6	7	6	4	2	1	1
	14	2	2	3	2	1	1	2	2	2	3	3	5	6	8	4	1	2	2
	15	1	2	2	2	1	1	2	1	2	3	3	3	4	4	5	1	1	1
	16	2	1	2	1	0	1	1	2	2	2	2	2	2	1	1	2	1	1
	17	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2
	18	1	0	1	0	0	0	0	1	1	1	1	1	1	2	1	1	2	2

Table 1. Co-attendance at 14 Social Events by 18 Southern Women

When an analyst fits an ultrametric model to proximity data, what is typically displayed is simply the sequence of partitions produced by the data. To illustrate this point, consider the data on co-attendance by Southern women at a series of social events as reported by Davis, Gardner and Gardner (1941). The data are shown in Table 1. Figure 1 shows four common forms for the display of the results of hierarchical clustering based on average proximity. From any of these displays we can see that two of women (1 and 3) cluster most closely. They are followed by the four others (2 and 4 join with 1 and 3 , and 12 and 13 join together). Then woman 14 joins the 12, 13 cluster, and so on until at the lowest level all the women are lumped together.

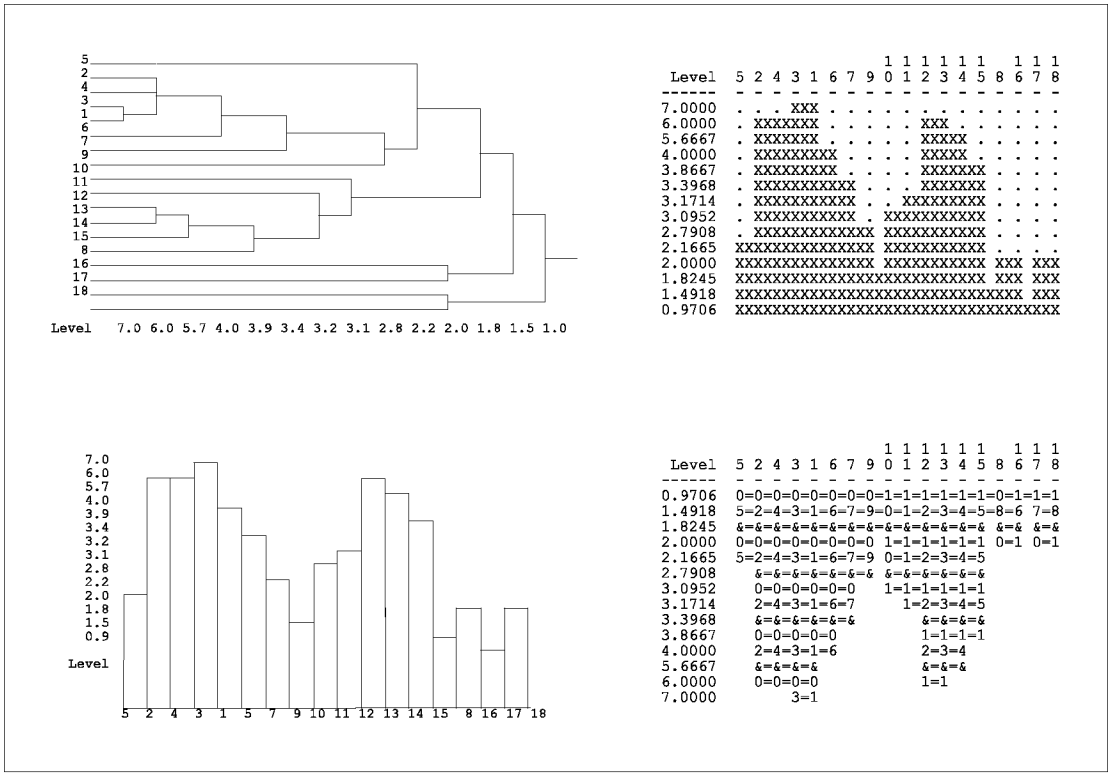

Figure 1. The clustering of the Southern Women data of Table 1 shown in four standard forms of graphic display.

But these standard modes of display reveal only the idealized cluster structure produced by the model; they provide a nested sequence of partitions. They order the women according to the clustering algorithm and show the level at which each set of clusters is merged. But they reveal absolutely nothing about the degree to which these idealized ultrametric proximities correspond to the observed proximities linking the women.

The situation is as if, when fitting a set of observed data points to an idealized curve, we displayed only the curve and suppressed all information that might show how closely the observations fit that curve. A good deal more important information is communicated when the curve is displayed along with the scatter of the points that it is intended to represent.

Ling (1973) recognized that this was a problem in displaying the results of cluster analysis. He adapted Sneath's (1957) shaded matrix as a way to show both the observed data and the idealized partitioning simultaneously. Essentially, Ling's proposal was to take half of a symmetric data matrix of proximities and to rearrange its rows and columns in terms of the order dictated by the clustering algorithm. Then a visual display is created by printing the cells representing close proximities using dark symbols and those that are more distant using lighter symbols. The Southern women example is shown in Figure 2.

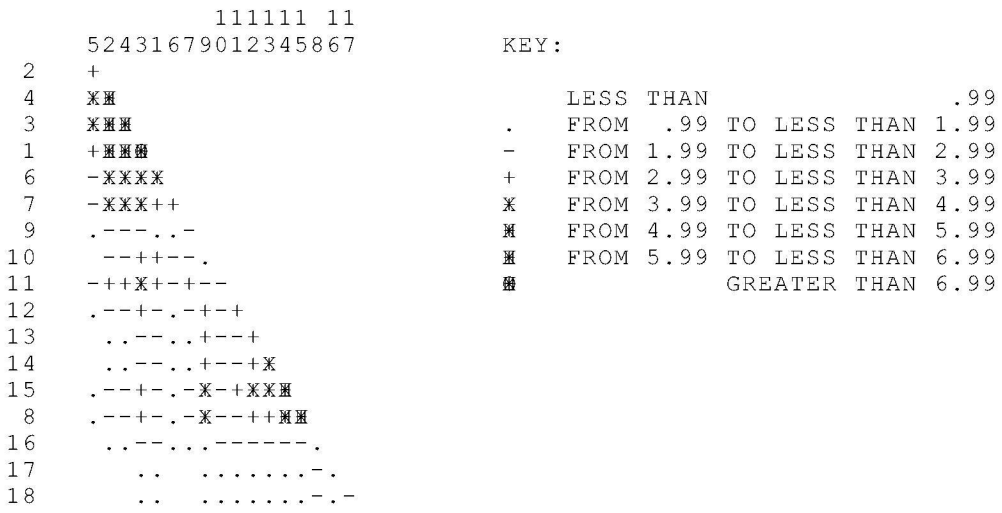

Figure 2. The clustering of the Southern Women data shown in shaded form.

As compared with other display forms, the shaded matrix makes a genuine attempt to provide the needed information on fit. It displays not only the idealized structure, but, by its arrangement of dense and light symbols, it tries to show how well the original data correspond to the model. In most cases, however, shaded matrices simply fail to do the job. They are both crude and difficult to interpret. Differences in the density of symbols cannot capture the full range of variation in most data sets. And, since the reader must constantly refer to the symbol

key, it is difficult to "see" the fit displayed in these tables. Although they are included as options in both BMDP and SYSTAT, shaded matrices are not often used to report research results.

3. Shaded Density Plots

The overstrikes shown in Figure 2 were all we could do in an era when printing was performed by a daisy wheel or a line printer chain. Printing then was limited to the use of typescript characters. But modern printers typically use dot matrices, lasers or ink jets and they are able to produce a wide range of uniform shadings. Shaded density plots simply substitute a full range of gray shades for the awkward and confusing overstrikes of the shaded matrix. Clearly, it is much easier to see differences in gray densities than to decode differences in the densities of overstruck typescript symbols.

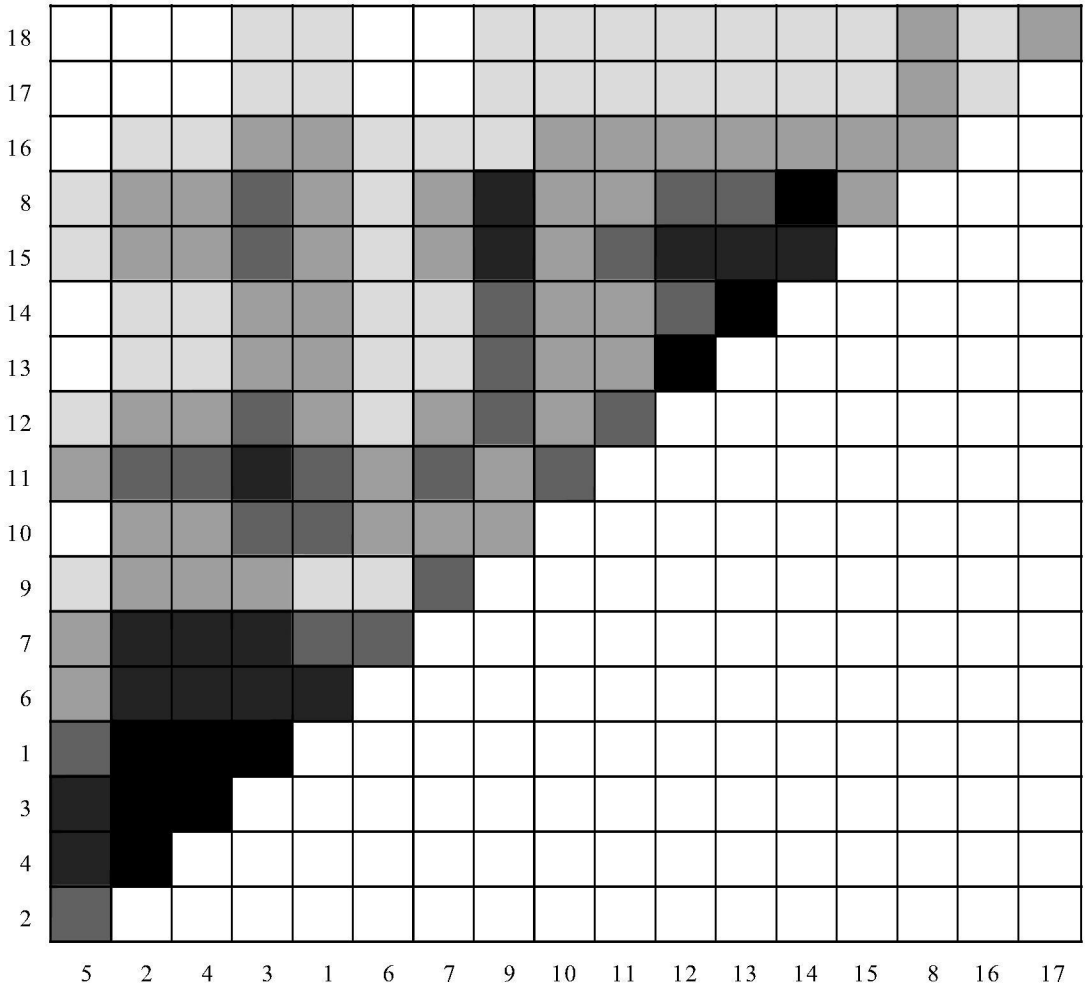

Figure 3. The clustering of the Southern Women data shown as a shaded density plot.

Consider Figure 3. It shows a shaded density plot representing the Southern women's proximities. It reveals the basic pattern of clusters along the diagonal. And, at the same time, it also captures the departures from ultrametric form that were present in the original data. When the entries in a proximity matrix are ultrametric they produce a pattern of smooth steps, peaking at the diagonal. Departures from strictly ultrametric proximities are apparent as

irregularities--dark and light spots--in the smooth appearance of the plot. Figure 4 shows what the proximities would have looked like if they had been exactly ultrametric.

A shaded density plot, then, provides an alternate way of presenting the information contained in a shaded matrix. But it is clear that in a density plot the information is both easier to decipher and provides greater detail about departures of the data from the ultrametric form. All in all, shaded density plots have the advantages of shaded matrix displays without having their limitations.

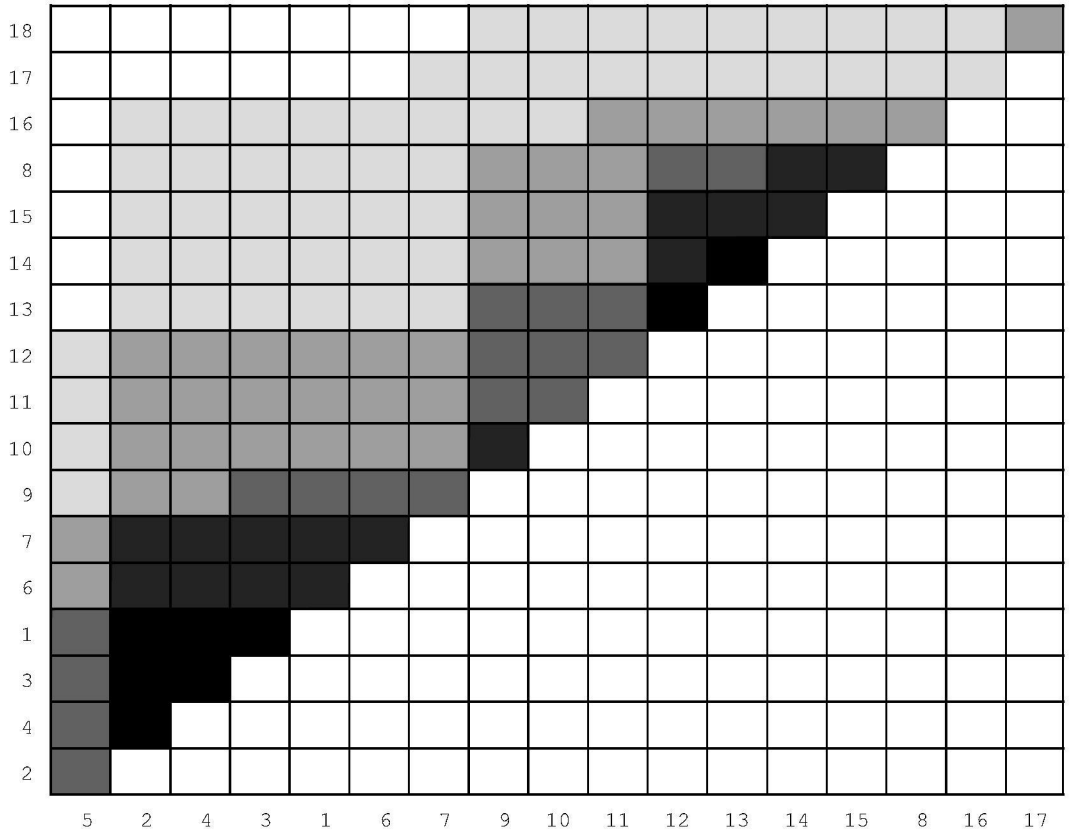

Figure 4. The clustering of the Southern Women data as they would appear if their proximities were exactly ultrametric.

Obviously these shaded density plots cannot be produced on traditional mainframe line printers, but given the ready availability of laser printers and ink jets that is no longer a problem. A great many graphic programs are available, and almost any home computer can generate graphic images. Images like those shown in Figures 3 and 4 clearly contain a great deal more

information about the ultrametric model and its fit to the data than was possible with earlier equipment.[1]

REFERENCES

Arnheim,R. (1970), *Visual Thinking*, London: Faber.

Bertin,J. (1967), *Semiologie Graphique,* Paris: Gauthier-Villars.

D'andrade,R. (1978), "U-Statistic Hierarchical Clustering," *Psychometrika,* 4 , 58-67.

Davis,A., Gardner,B.B., and Gardner,M.R. (1941) *Deep South,* Chicago: University of Chicago Press.

Engelman,L. (1977), "Cluster Analysis of Cases," *BMDP Biomedical Computer Programs*, Eds: W.J. Dixon, and et al., Berkeley: University Of California Press.

Gower,J.C. (1967), "A Comparison of Some Methods of Cluster Analysis," *Biometrics*, 2 , 62 -6 7.

Hartigan,J.A. (1967), "Representation of Similarity Matrices by Trees," *J. Amer. Statist. Assoc.,* 62, 1140-1158.

Hartigan,J.A. (1975), *Clustering Algorithms,* New York: John Wiley.

Johnson,S.C. (1967), "Hierarchical Clustering Schemes," *Psychometrika,* 2, 241-254.

Kleiner,B., and Hartigan,J.A. (1981), "Representing Points in Many Dimensions by Trees and Castles," *Journal of the American Statistical Association,* 76, 260-269.

Klovdahl,A.S. (1981), "A Note on Images of Networks," *Social Networks, 3*, 1973-214.

Koestler,A. (1964), *The Act of Creation,* New York: Macmillan.

Kruskal,J.B., and Landwehr,J.M. (1983), "Icicle Plots: Better Displays For Hierarchical Clustering," *American Statistician,* 7, 162-168.

Lance,G.N., and Williams,W.T. (1967), "A General Theory of Classificatory Sorting Strategies I. Hierarchical Systems," *The Computer Journal,* 9, 7 - 80.

Ling,R.F. (1973), "A Computer Generated Aid for Cluster Analysis," Communications of the ACM, 16, 55- 61.

McCammon,R.B. (1968), "The Dendograph: a New Tool for Correlation," *Geological Society of America Bulletin*, 79, 1663 -1670.

McWhitty,L.L. (1967), "A Mutual Development of Some Typological Theories and Pattern-Analytic Methods," *Educational and Psychological Measurement,* 17, 21-46.

Shepard,R.N. (1974), "Representation of Structure in Similarity Data: Problems and Prospects," *Psychometrika,* 9, 373-421.

Sibson,R. (1970), "A Model for Taxonomy Ii," *Math. Biosci.,* 6, 405-4 0.

Sneath,P.H.A. (1957), "The Application of Computers to Taxonomy," *J. Gen. Microbiol.,* 17, 201-226.

Sneath,P.H.A., and Sokal,R.R. (1973), *Numerical Taxonomy,* San Francisco: W. H. Freeman.

Taylor,A.M. (1971), *Imagination and the Growth of Science,* New York: Shocken.

Tufte,E.R. (1983), *The Visual Display of Quantitative Information,* Cheshire, Connecticut: Graphics Press.

Tukey, J.W. (1972), "Some Graphic and Semigraphic Displays," *Statistical Papers in Honor of George W. Snedecor,* Ed: T.a. Bancroft, Ames Iowa: Iowa State University Press.

Ward,J.H.,Jr. (1963), "Hierarchical Grouping to Optimize an Objective Function," *Journal of the American Statistical Association,* 58, 236- 244.

Weller,S.C., and Buchholtz,C.H. (1986), "When a Single Clustering Method Creates More than One Tree: a Reanalysis of the Salish Languages," *American Anthropologist,* 88, 36-43.

Wirth,M.G., Estabrook,G.F., and Rogers,D.J. (1966), "A Graph Theory Model for Systematic Biology, with an Examination for the Oncidiinae (Orchidaceae)," *Systematic Zoology,* 15, 59-69.

---

[1]Figures 3 and 4 were produced using the standard graphics in the MATHEMATICA system.