Morse, J. M., Swanson, J. M., & Kuzel, A. J. (Eds.). (2001). *The nature of qualitative evidence.* Thousand Oaks, CA: Sage.
Pelusi, J. (1997). The lived experience of surviving breast cancer. *Oncology Nursing Forum, 24*(8), 1343–1353.

CLOSED QUESTION. *See* CLOSED-ENDED QUESTIONS

CLOSED-ENDED QUESTIONS

Survey questions come in two varieties: OPEN-ENDED QUESTIONS, in which the respondents provide their own answers, and closed-ended questions, in which specific response categories are provided in the question itself. Although there has been considerable research on the relative merits of the two types of questions, the substantial preponderance of questions that appear in any survey are closed-ended questions.

All survey research involves asking questions for which the responses are categorized to facilitate analysis. In closed-ended questions, the researcher makes prior judgments about what the appropriate categories might be and offers them immediately to the respondent in the question wording. There are a number of potential problems with this format, not the least of which is that it typically describes the "world" in dichotomous terms that sometimes reflect an oversimplification of possibilities. This kind of constraint does not exist with an open-ended question, and for this reason Schuman and Presser (1996) conclude that they often provide more valid data. However, the cost saving of not having to CODE verbatim responses and the quicker access to analysis make close-ended questions the preferred form.

When designing or using close-ended questions, a number of standard suggestions reflect considerable research on the matter. Using forced choices is preferable to asking a respondent to agree or disagree with a single statement. A middle alternative should generally be offered, except when measuring intensity, and an explicit "no opinion" option should be offered as well. Researchers should use multiple questions to assess the same topic, remaining sensitive to the effects of question order. And when in doubt about the possible effects of question wording or response options, research should include split-sample versions in their questionnaires so that comparative analyses can be run.

—Michael W. Traugott

REFERENCES

Converse, J. M., & Presser, S. (1986). *Survey questions.* Thousand Oaks, CA: Sage.
Fowler, F. J., Jr. (1995). *Improving survey questions: Design and evaluation.* Thousand Oaks, CA: Sage.
Schuman, H., & Presser, S. (1996). *Questions & answers in attitude surveys.* Thousand Oaks, CA: Sage.

CLUSTER ANALYSIS

Social science DATA sets usually take the form of observations on UNITS OF ANALYSIS for a set of VARIABLES. The goal of cluster analysis is to produce a simple classification of units into subgroups based on information contained in some variables. The vagueness of this statement is not accidental. Although there may be no formal definition of cluster analysis, a slightly more precise statement is possible. The *clustering problem* requires solutions to the task of establishing clusterings of the n units into r clusters (where r is much smaller than n) so that units in a cluster are similar, whereas units in distinct clusters are dissimilar. Put differently, these clusterings have homogeneous clusters that are well separated. *Cluster analysis* is a label for the diverse set of tools for solving the clustering problem (see Everitt, Landau, & Leese, 2001). Most often, these tools are used for INDUCTIVE explorations of data. The hope is that the clusterings provide insight into the structure of the data, the nature of the units, and the processes generating the variables. For example, cities can be clustered in terms of their social, economic, and demographic characteristics. People can be clustered in terms of their psychological profiles or other attributes they possess.

DEVELOPMENT OF CLUSTER ANALYSIS

Prior to 1960, many clustering problems were solved separately in different disciplines. Progress was fragmented. The early 1960s saw attempts to provide general treatments of cluster analysis, given these many developments. Sokal and Sneath (1963) provided an extensive discussion and helped set the framework for the development of cluster analysis as a data-analytic field. Specifying clustering problems is not difficult. Nor are the mathematical foundations for expressing and creating most solutions to the clustering problem. The difficulty of cluster analysis comes

from the *computational complexities* in *establishing* solutions to the clustering problem. As a result, the field has been driven primarily by the evolution of computing technology. Generally, this has been beneficial, with substantive interpretations being enriched by useful clusterings. In addition, many technical developments have stemmed from exploring substantive applications in new domains. There are now many national societies of cluster analysts that are linked through the International Federation of Classification Societies.

SOLVING CLUSTERING PROBLEMS

In general, the clustering problem can be stated as establishing one (or more) clustering(s) with r clusters that have the minimized value of a well-defined criterion function over all feasible clusterings. The criterion function provides a measure of fit for all clusterings. In practice, however, the criterion function often is left implicit or ignored. In most applications, the clustering is a partition, but "fuzzy clustering" with overlapping clusters is possible. Once the units of analysis have been selected, there are five broad steps in conducting cluster analyses:

1. measuring the relevant variables (both QUANTITATIVE VARIABLES and CATEGORICAL variables can be included, and some form of standardization may be necessary),
2. creating a (dis)similarity MATRIX for an appropriate measure of (dis)similarity,
3. creating one or more clusterings via a clustering algorithm,
4. providing some assessment of the obtained clustering(s), and
5. interpreting the clustering(s) in substantive terms.

Although all steps are fraught with hazard, Steps 2 and 3 are the most hazardous, and Step 4 is ignored often. In Step 2, dissimilarity measures (e.g., Euclidean, Manhattan, and Minkowsky distances) or similarity measures (e.g., CORRELATION and matching COEFFICIENTS) can be used. The choice of a measure is critical: Different measures can lead to different clusterings. In Step 3, there are many ALGORITHMS for establishing clusterings. Each pair of choices (of measures and algorithms), in principle, can lead to different clusterings of the units.

Hierarchical clustering can take an agglomerative or a divisive form. An agglomerative clustering starts with each unit in its own cluster and systematically merges units and clusters until all units form a single cluster. Divisive hierarchical clustering proceeds in the reverse direction. Within these categories, there are multiple options. Three of the most popular are single-link, average-link, and complete-link clustering. In single-link clustering, the algorithm computes the (dis)similarity between groups as the (dis)similarity between the closest two units in the two groups. For complete-linkage clustering, the farthest pair of units in the two groups is used, and in average-linkage clustering, the algorithm uses the average (dis)similarity of units between the two groups. Ward's method is popular also for computing ways of combining clusters. The choice between these methods is critical, as shown in Figure 1.

The first panel of Figure 1 shows a BIVARIATE scattergram whose 25 units can be classified. Squared Euclidean distance was used for all three clusterings and dendrograms shown below. For this example, if (x_i, y_i) and (x_j, y_j) are the coordinates of two units, the squared Euclidian distance between them is $(x_i - x_j)^2 + (y_i - y_j)^2$. The top right panel shows the single-linkage clustering dendrogram. Those in the bottom row of Figure 1 are for the average- and complete-linkage methods. The scale on the left of the dendrogram shows the measure of dissimilarity, and the horizontal lines show when units or clusters are merged. The vertical lines keep track of the clusters. The average- and complete-linkage clusterings are the closest to each other, and both suggest clusterings with three clusters. However, the clusters differ in their details, and the single-link clustering differs markedly from the other two. Which clustering is "better" can be judged by examining the clusterings and the scattergram—with some idea of how and why units can be grouped. For real analyses, substance guides this judgment.

Because clustering tools are exploratory, there are few tools for STATISTICAL INFERENCE concerning the VALIDITY of a clustering (Step 4), and their utility is limited because of the prior exploration. Choosing the number of clusters from a dendrogram often is viewed as a matter of judgment, and ad hoc justifications for any clustering are easy to reach given a clustering. This, too, is illustrated in Figure 1.

When the number of clusters is known or assumed, some nonhierarchical clustering methods are available.

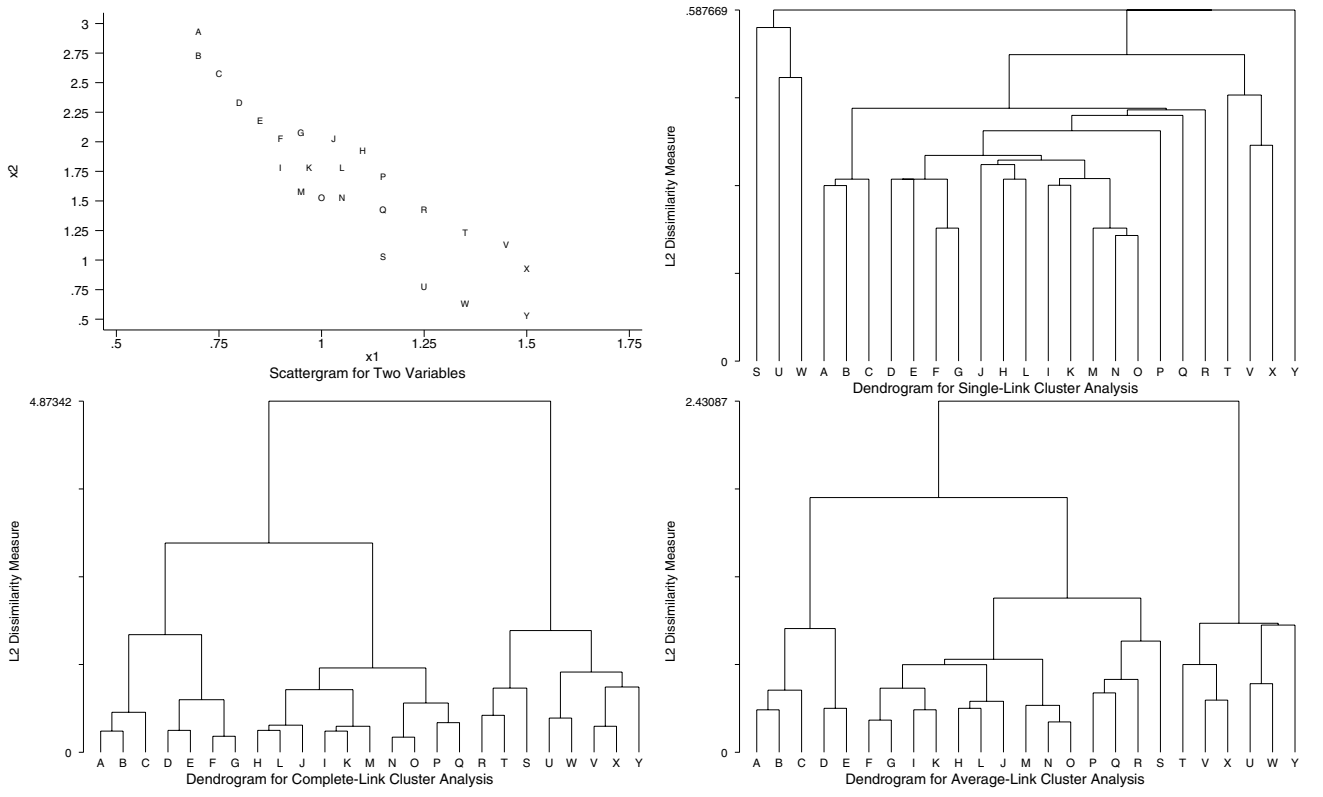

Figure 1 Three Clusterings of a Common Data Set

These include *k*-means and *k*-medians clustering. Two additional methods are the leader and relocation algorithms. Both are local optimization methods and have to be repeated many times to avoid reaching a local minimum. The relocation algorithm is at the heart of the direct clustering approach to block modeling social networks developed by Doreian, Batagelj, and Ferligoj (1994). Both network actors *and* the network ties are clustered, whereby a criterion function is defined *explicitly* in terms of substantive concerns and the NETWORK ties. As a result, the resulting solutions to the clustering problem are not ad hoc.

—Patrick Doreian

REFERENCES

Doreian, P., Batagelj, V., & Ferligoj, A. (1994). Partitioning networks based on generalized concepts of equivalence. *Journal of Mathematical Sociology, 19,* 1–27.

Everitt, B., Landau, S., & Leese, M. (2001). *Cluster analysis* (4th ed.). Oxford, UK: Oxford University Press.

Sokal, R., & Sneath, P. (1963). *Principles of taxonomy.* San Francisco: Freeman.

CLUSTER SAMPLING

Cluster sampling involves sorting the units in the study population into groups and selecting a number of groups. All the units in those groups are then studied. It is a special case of MULTISTAGE SAMPLING.

—Peter Lynn

COCHRAN'S *Q* TEST

W. G. Cochran (1950) developed the *Q* statistic for matched or WITHIN-SUBJECT DESIGNS in which each subject (*r*) provides a dichotomous response for each experimental condition (*c*). Cochran's *Q* tests whether the probability of a target response is equal across