Journal of Classification 5:205-228 (1988)

Variable Selection in Clustering

E. B. Fowlkes
Bellcore

R. Gnanadesikan
Bellcore

J. R. Kettenring
Bellcore

Abstract: Standard clustering algorithms can completely fail to identify clear cluster structure if that structure is confined to a subset of the variables. A forward selection procedure for identifying the subset is proposed and studied in the context of complete linkage hierarchical clustering. The basic approach can be applied to other clustering methods, too.

Keywords: Variable selection; Cluster analysis of two-mode data; scaling of variables; Pillai trace statistic; Interactive data analysis

We are grateful to T. Achariyapaopan and S. H. Lin for introducing us to the challenging problem of modeling microwave fading, and to E. M. Sieff for her careful reading of the manuscript and helpful comments.

Authors' Addresses: E.B. Fowlkes, R. Gnanadesikan, and J.R. Kettenring, Bellcore, 445 South Street, Morristown, New Jersey 07960-1910, USA.

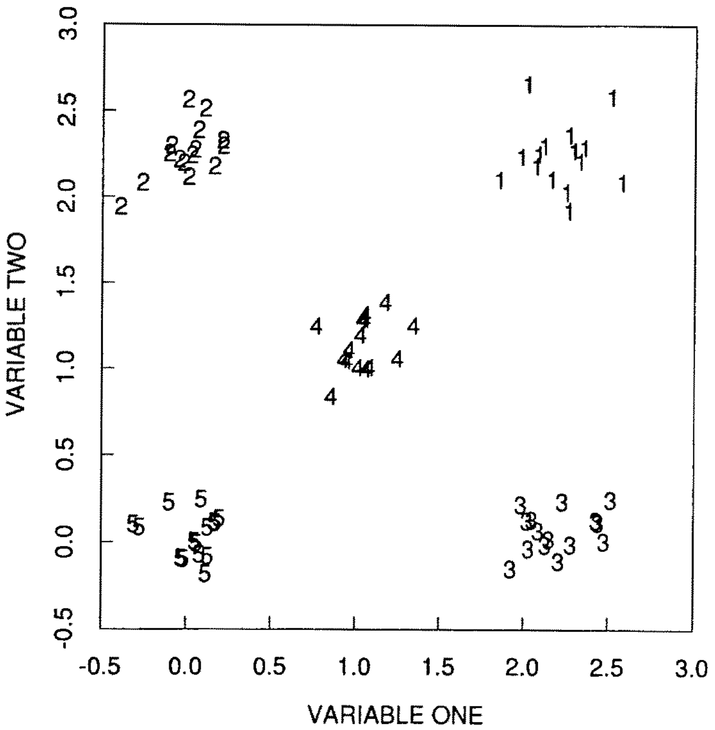

Figure 1. Scatter plot of two variables that shows five clusters.

1. Introduction

The problem addressed in this paper is best illustrated by an example. See Figure 1. The scatter plot of two variables in the figure suggests that there are five clusters present (and the points are labeled accordingly); however, it shows only two of the five variables in this data set. The other three consist simply of random noise and show no cluster structure.

A standard cluster analysis of this data set might proceed by separately standardizing each of the five variables, and then applying a convenient algorithm such as *k*-means (MacQueen 1967) or either single linkage or complete linkage hierarchical clustering based on Euclidean distances between all pairs of the objects being clustered. Each of these standard algorithms fails to identify the five group structure suggested by Figure 1.

With knowledge that the cluster structure is confined to two of the variables, one could use only these variables and ignore the others. The result of doing this is shown in Figure 2 for the complete linkage method. The five clusters, with their members labeled 1 through 5, appear as five major branches in the dendrogram. The corresponding picture when all five

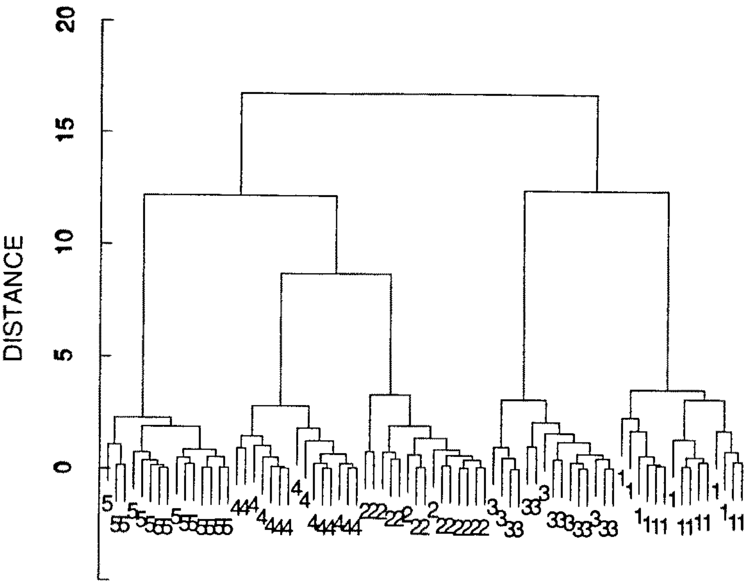

Figure 2. Dendrogram that results when clustering is done using the two variables plotted in Figure 1.

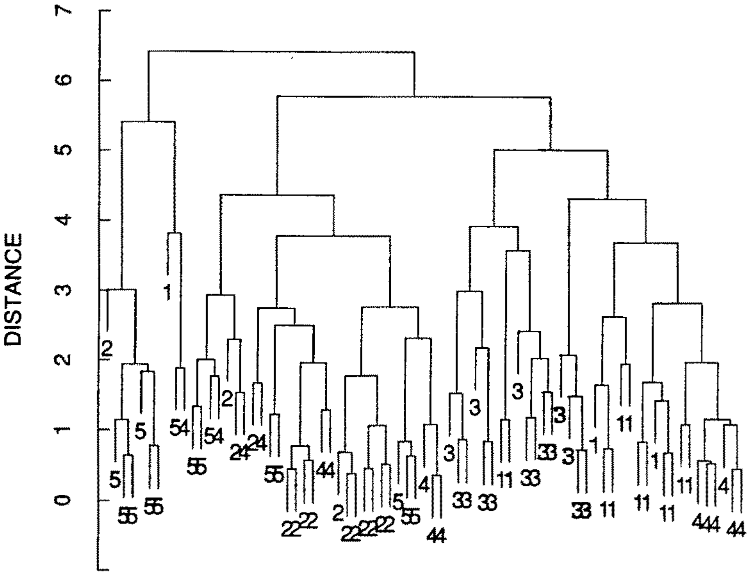

Figure 3. Dendrogram that results when clustering is done using all five variables.

variables are used for clustering is displayed in Figure 3. The "true-cluster" numbers are now mixed up because the cluster structure has been completely masked by including the three noise variables!

This type of phenomenon is no surprise to experts in clustering. In pattern recognition, the importance of "feature selection" is well recognized. The approach of variable selection developed in this paper is one answer to this problem — an *a posteriori,* data-based selection of key features rather than a choice based solely on considerations prior to analyzing the data. In practice the situation tends not to be as clear cut as in the above artificial example, and the question is what to do about it.

There is a considerable literature that touches on the problem of weighting variables to facilitate the extraction of clusters. For recent reviews, see DeSarbo, Carroll, Clark, and Green (1984) and De Soete, DeSarbo, and Carroll (1985). Much of this work, including in particular these two papers, focuses on ways of finding differential weights for variables as part of the clustering algorithm. However, in De Soete (1986), a method is proposed for finding "optimal" weights for use as input to ultrametric or hierarchical clustering. Block clustering methods (see, e.g., Hartigan 1972), which divide a data matrix of variables by observations into blocks, provide another way of tackling the problem. In the present paper, the approach is different from any of these: *subsets* of variables are extracted for use in conjunction with *standard* clustering procedures.

Consequently, the flavor here is more akin to variable selection in discriminant analysis. However, the fact that the groups are pre-specified, rather than data-based, makes the development of variable selection procedures for discriminant analysis much easier. For recent reviews of the variety of techniques available in this area, see Seber (1984, Section 6.10) and McKay and Campbell (1982). See also Fowlkes, Gnanadesikan, and Kettenring (1987) for a combined discussion of variable selection in regression, discriminant analysis, and clustering.

One popular procedure in discriminant analysis has been the use of tests for additional information as a basis for selecting variables in a forward selection manner. The next variable to be added is the one yielding the most significant value in a test of equality of group means, conditioning on the presence of the previously chosen variables. This is equivalent to picking as the next variable the one that maximizes Wilks' likelihood ratio statistic, used in testing equality of group mean vectors in multivariate analysis of variance (see, e.g., Seber 1984, p. 341).

This statistic is one of several that are functions of the eigenvalues, e_i, of $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ and $\mathbf{B}$ are the usual within and between group sums of cross products matrices. A similar approach is followed in this paper for the clustering problem: decisions are based on the relative sizes of standard eigenvalue-based statistics measuring separation among contemplated clusters. Of course, the distribution theory for these statistics is different, and much more complicated, in this situation because the clusters are data dependent. Furthermore, the type of conditioning arguments used in the discriminant analysis problem cannot be carried over because the cluster structure is evolving along with the choice of variables.

To put the problem in a broader setting, the objective can be described as trying to achieve an effective ''reduction of dimensionality,'' which is often advertised as one of the major goals of multivariate analysis. Methods for finding such reductions range from very formal ones involving significance testing to very informal ones that are guided by numerical summaries and graphical displays in an interactive data analytic environment. One can also distinguish between procedures that perform the reduction indirectly based on linear combinations of variables and those that extract subsets of the variables directly. The procedures developed in this paper are very definitely in the informal, subset selection corner.

While the specific procedure that is developed here is for use in conjunction with complete linkage hierarchical clustering, the same basic approach could be easily adapted to other hierarchical schemes and, with appropriate modifications, to non-hierarchical ones such as the k-means method. The procedure is of the forward selection variety, but other types have been considered, too. For instance, a backward elimination approach was investigated but found to be ineffective. The primary reason is suggested by the example already described: the noise variables distort the analysis based on all variables to such an extent that procedures designed to select the best ones to drop are fooled.

Experiments have also been run with a ''guided selection'' procedure that attempts to find an intermediate route between forward selection and backward elimination. The idea is to make an intelligent initial guess at a subset of variables that appears promising and then to work forward and backwards from there. A crude implementation of this idea appears to work quite well, but it is not reported here in detail because of the need for further refinement of the procedure.

A different and more direct method would be to optimize a function of the eigenvalues directly with respect to the partitioning into clusters and the choice of variables. A major drawback of such an approach is its computational demands; Gnanadesikan (1977, p. 104). The procedure developed in this paper attempts to find reasonable answers in a way that avoids such demands.

The following sections deal with the details of the forward selection algorithm, simulation experiments to check its performance, application to a real data example, and concluding remarks.

2. The Forward Selection Algorithm

In the context of complete linkage hierarchical cluster analysis, the forward selection algorithm begins by searching for the single variable out of the total, p, that shows the most evidence of clustering of the n objects. This is done by producing a hierarchical tree based upon each variable; cutting each tree at successive levels to produce partitions of the data into $k = 2, \ldots, k_{\max}$ (a user specified number) groups; measuring the amount of separation for each of the $p \cdot k_{\max}$ groupings; and selecting the variable for which the separation is the most "significant."

Additional variables are selected in a similar manner. To find a second variable, all combinations of two variables that include the one selected at the first stage are considered. The combination that produces the most "significant" partition is chosen. The process continues until there no longer appears to be a variable worth adding.

The "significance" of variables is judged informally against a null background of no cluster structure. Specifically, the null assumption is that the data are a random sample of n observations from a p-dimensional multivariate normal distribution with a diagonal covariance matrix. There would be no serious difficulty in changing this background assumption to, for example, a uniform distribution or a distribution with general covariance structure.

Prior to the start of the variable selection process, the data are standardized so that each variable has a unit sample standard deviation. Then Euclidean distances are computed between every pair of objects, based on the variable or subset of variables being considered, and these are used as the input to the hierarchical cluster analysis.

This form of standardization of the data is not entirely satisfactory because it tends to mask the presence of clusters (see, e.g., Hartigan 1975, p. 62 and Milligan and Cooper 1988). Similarly, one could standardize by the sample covariance matrix if the null background were assumed to have a general covariance structure, but this can be criticized for the same reason in the non-null case. A conceptually more attractive method that is appropriate for situations where clusters have homogeneous covariance structures is to standardize by the sample covariance matrix estimate, based on pairwise differences, developed by Art, Gnanadesikan, and Kettenring (1982). This method attempts to find a rough estimate of the within-cluster covariance matrix without knowing the cluster structure in advance.

While there are clearly many ways to standardize, the relatively primitive method of equalizing variances that is used in the present study has been the popular choice of many practitioners, even though it is now possible to do better in the sense of revealing rather than obscuring clusters, at least in some circumstances.

Many different statistics can be used to measure the amount of separation in a partition induced by cutting a tree. Perhaps, the most important of these are the traditional ones from discriminant analysis or multivariate analysis of variance. They are based on the eigenvalues, e_i, of $\mathbf{W}^{-1}\,\mathbf{B}$, where $\mathbf{B}$ and $\mathbf{W}$ are the $(p^* \times p^*)$ between and within groups sum of cross-products matrices based on the k groups and the p^* variables in the subset of variables under consideration. The most successful one tested in the present study is

$$S = S(k) = \frac{1}{t} \sum_i \frac{e_i}{(1+e_i)} \,,$$

which is a scaled version of Pillai's (1955) trace statistic. The scale factor, $t = \min(p^*, k-1)$ forces Pillai's statistic to lie in the interval (0,1). Pillai's statistic has proved to be sensitive in the multivariate analysis of variance context in terms of easily interpretable functions of the noncentrality parameters (Roy, Gnanadesikan, and Srivastava 1971, Chapter 5).

To assess the strength of clustering in a partition into k groups, one could calculate

$$S^*(k) = S(k) - E(S(k)) \,, \; -1 \le S^*(k) \le 1 \,, \qquad (1)$$

where $E(S(k))$ is the mean of $S(k)$ under the null model. For example, suppose a particular variable has been chosen at the first stage, and one is now searching for a second one that may be worth adding. Then $S(k)$ would be evaluated from the data at hand for every two-variable combination that includes the one chosen at the first stage. Similarly, $E(S(k))$ would be calculated, in principle, for a corresponding normal sample, assuming the "best" variable had been entered at the first stage. To be precise, it is presumed that "best" means picking the one with the highest S-value.

Ideally, it would be preferable to perform a full standardization of $S(k)$ by also dividing through by its standard deviation. This turns out to yield a procedure that is too delicately dependent on the null assumptions. Better results have been obtained with (1), although the fact that the variance of $S^*(k)$ does depend upon k should not be forgotten.

Because $E(S(k))$ is theoretically intractable, it is necessary to estimate it via simulation. In today's computing environment, the simulations can be done as part of the variable selection process. The specific procedure used is as follows: for a particular (n, p) combination, 100 spherical normal samples are generated, variables are selected from each sample according to the maximum value of $S(k)$, and averages of S-values across samples are obtained as estimates of $E(S(k))$ with "null" data. With these estimates in hand, the variable selection process can proceed.

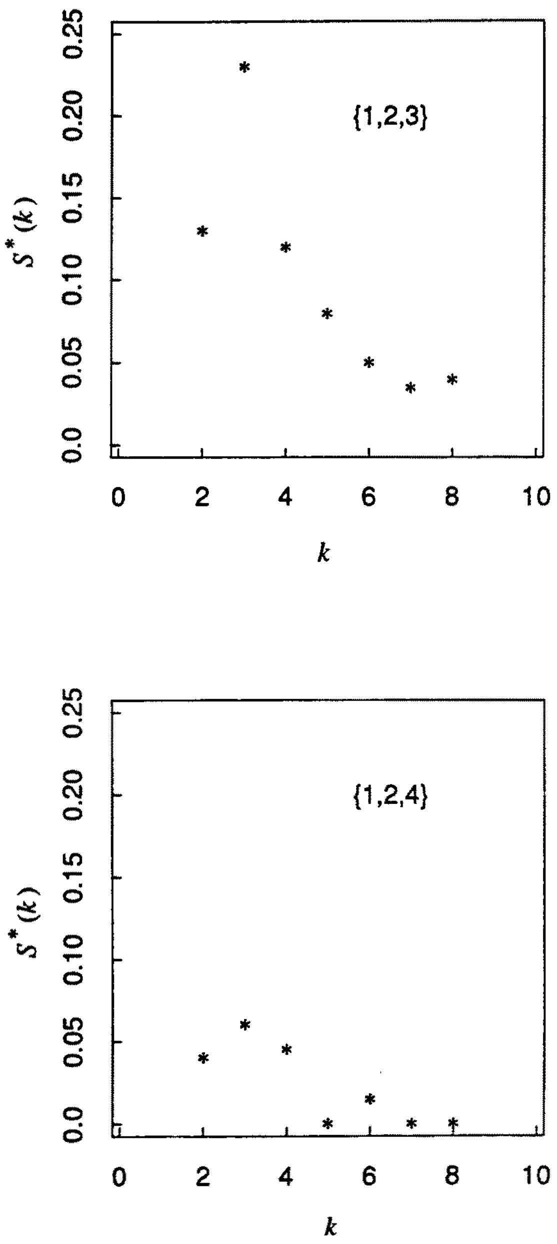

Figure 4. Plots of separation statistic, $S^*(k)$, versus number of clusters, k, for variable sets {1,2,3} and {1,2,4}.

The informal graphical procedure that is used to execute the variable selection, as well as to suggest the number of clusters, is based upon plots of $S^*(k)$ versus k. Figure 4 shows a hypothetical example. In this example, $p = 4$ and variables 1 and 2 have already been selected at the two previous stages. Now, at stage 3, the question is whether to enter variable 3, variable 4, or neither. It appears that the best choice would be to enter variable 3 because the value of $S^*(k)$ is larger in this case for every value of k and, in particular, because there is a very strong indication of three clusters associated with the $\{1,2,3\}$ combination. In practice, the choice may not be so clear cut, and other information may be needed before a decision can be made. Looking at the scatter plots of the variables in question can be very helpful.

There are two features of the graphical procedure, related to earlier comments, that need to be kept in mind while interpreting the $S^*(k)$ versus k plots. First, the variance of $S^*(k)$ is not constant under null conditions. In fact it tends to decrease as k increases. Second, because the effects of selecting ''significant'' variables at earlier stages have not been conditioned out, their impact will tend to spillover into plots at later stages.

3. Simulation Experiments

A large number of simulation experiments on computer-generated data were run to study the properties of the variable selection algorithm. They included checks on its behavior under null conditions of no clustering as well as tests of its ability to recover cluster structure consisting of various types of elliptical point clouds. In each of the non null cases, the cluster structure was confined to a subset of the p variables. Highlights of these experiments are summarized in this section.

Experiment 1 - Null Spherical Data

The forward selection algorithm was run on 100 fresh samples of spherical normal data with $(n, p) = (50,5)$ and $(n, p) = (75,5)$. The distributions of S^*-values were studied in each case using box plots. The primary issue is to what extent these distributions hover around zero since the null situation should not suggest that particular variables are needed for clustering. Figure 5 shows an example of the results at the stage of picking a second variable when $n = 75$. The distributions are centered near zero with variability that decreases as k increases. At the same time, Figure 5 does not indicate any strong bias towards picking an unneeded variable inadvertently. The other figures (not shown) provide a similar story.

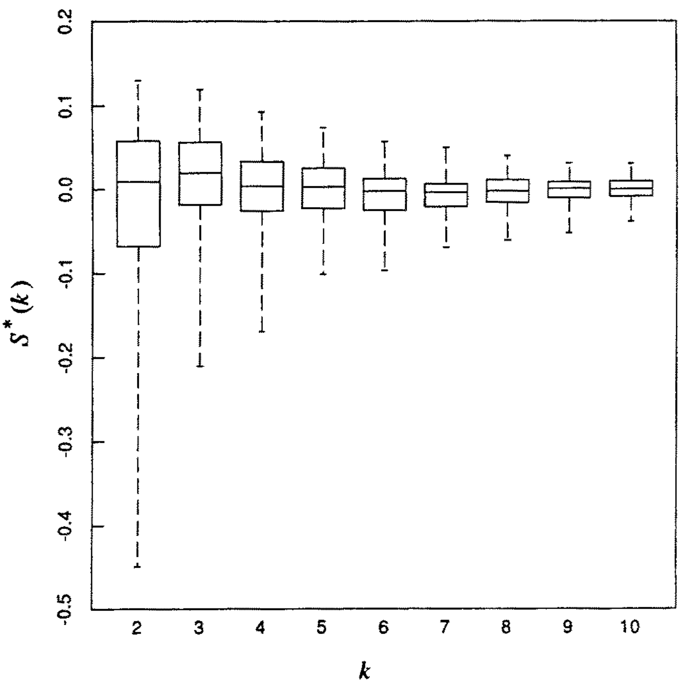

Figure 5. Box plots of distributions of $S^*(k)$ for selection of second of five variables using the null data with sample size 75.

Experiment 2 - Clusters Along Coordinate Axes

The design of this experiment is particularly favorable to the forward selection procedure because the clusters can be associated with individual variables. Specifically, unit variance spherical normal clusters of size $n(l)$ and dimension $p = 2\ l + 1$ were located at a distance $d = 5.0$ from the origin along each of the first l coordinate axes. Thus the clusters were confined to an l-dimensional subspace. The other $l+\ 1$ dimensions corresponded to independent unit normal ($N(0,1)$) noise variables with no cluster structure. Tests were run with $l = 2(1)5$ and $n(l) = [100\,/\ l]$. For each l, 100 random samples were drawn. The forward selection algorithm picked the best l-subset in all cases.

Experiment 3 - Clusters in a Plane

Four types of bivariate normal cluster structure in two variables were combined with three other independent $N(0,1)$ noise variables to check how consistently the structure variables were selected in the first two stages of the selection algorithms. For each of the four cases, the experiment was repeated 100 times. Figure 6 shows scatter plots of the variables with the cluster

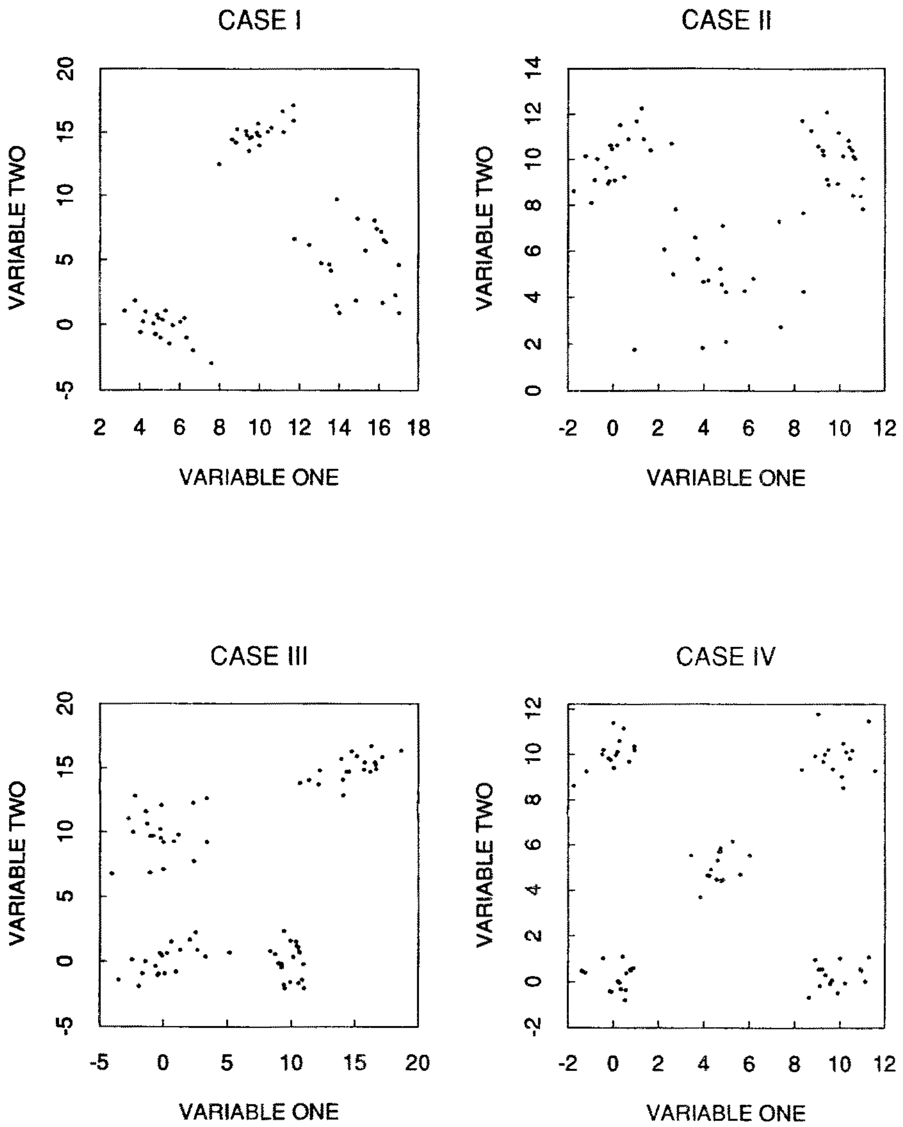

Figure 6. Scatter plots illustrating types of data used in Experiment 3.

structure for one random sample from each of the four cases. The specifications of the remaining cluster parameters are shown in Table 1. The design provides for testing the effects of differences in cluster separation and within-cluster covariance structure. The forward selection algorithm picked the cluster variables correctly in the first two steps in each of the 400 samples. Box plot summaries of the S^*-values at the second stage of variable selection are shown in Figure 7 for each of the four cases. It is clear from these plots

Table 1

Parameter Specifications

Case I:					
cluster sizes	20	20	20		
cluster means	$\begin{bmatrix}5\\0\end{bmatrix}$	$\begin{bmatrix}10\\15\end{bmatrix}$	$\begin{bmatrix}15\\5\end{bmatrix}$		
cluster variances	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}2\\2\end{bmatrix}$		
cluster correlations	-0.7	0.7	0.0		
Case II:					
cluster sizes	20	20	20		
cluster means	$\begin{bmatrix}0\\10\end{bmatrix}$	$\begin{bmatrix}5\\5\end{bmatrix}$	$\begin{bmatrix}10\\10\end{bmatrix}$		
cluster variances	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$		
cluster correlations	0.7	0.0	-0.7		
Case III:					
cluster sizes	20	20	20	20	
cluster means	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}0\\10\end{bmatrix}$	$\begin{bmatrix}10\\0\end{bmatrix}$	$\begin{bmatrix}15\\15\end{bmatrix}$	
cluster variances	$\begin{bmatrix}2\\1\end{bmatrix}$	$\begin{bmatrix}2\\2\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}2\\1\end{bmatrix}$	
cluster correlations	0.7	0.0	0.0	0.7	
Case IV:					
cluster sizes	15	15	15	15	15
cluster means	$\begin{bmatrix}0\\0\end{bmatrix}$	$\begin{bmatrix}0\\10\end{bmatrix}$	$\begin{bmatrix}5\\5\end{bmatrix}$	$\begin{bmatrix}10\\0\end{bmatrix}$	$\begin{bmatrix}10\\10\end{bmatrix}$
cluster variances	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$	$\begin{bmatrix}1\\1\end{bmatrix}$
cluster correlations	0.0	0.0	0.0	0.0	0.0

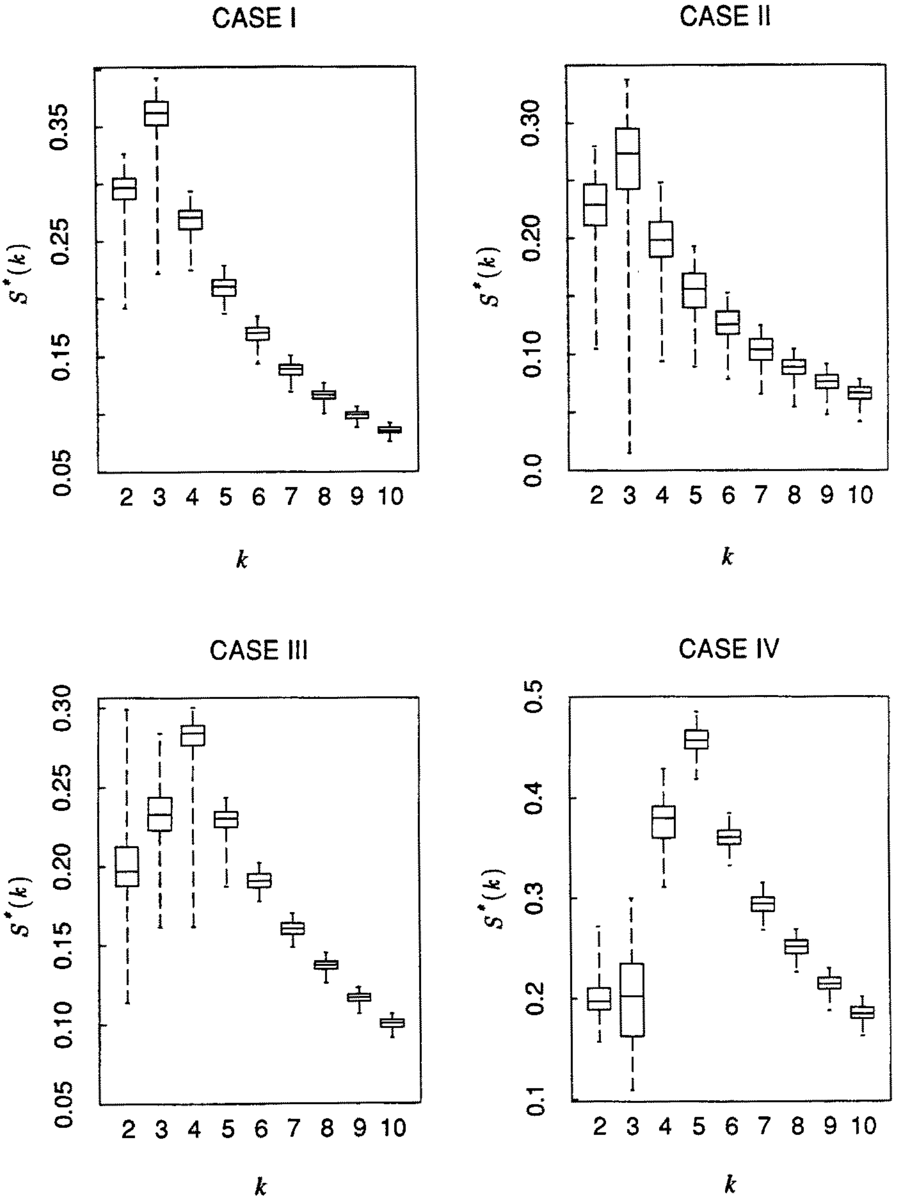

Figure 7. Box plot summaries of $S^*(k)$ for Experiment 3.

that the correct number of clusters would have been inferred in most of the samples as well. For example, note how the box heights peak in Case I at $k = 3$, which corresponds to the actual number of simulated clusters.

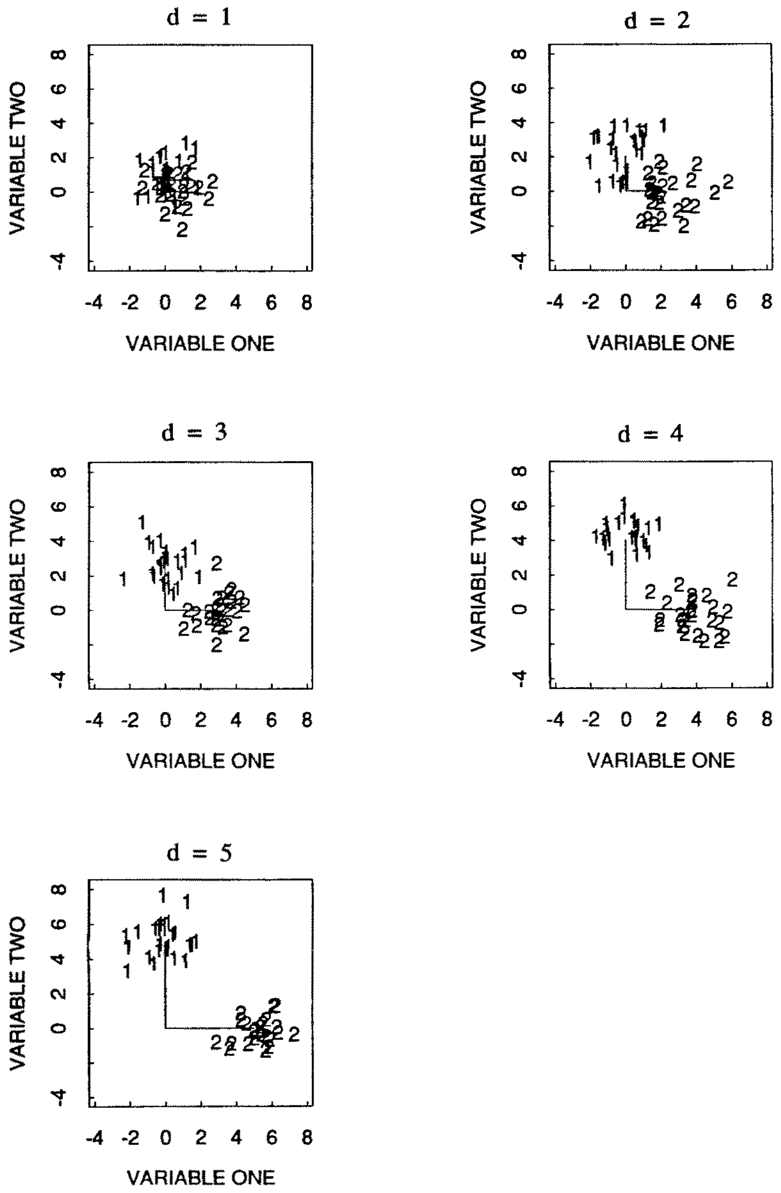

Figure 8. Scatter plots illustrating types of data used in Experiment 4.

Experiment 4 - Breakdown

The layout in this case is similar to Experiment 2 except that: $l = 2$ or 4 only, $n(l) = 25$, and $d = 1(1)5$. By reducing d, one gets an idea of when the selection procedure breaks down. Ten random samples were generated for each *(l,d)* combination. Figure 8 shows one sample realization for each d with $l = 2$. The clusters show considerable overlap when $d = 1$ or 2 but are well separated for other values. The forward selection procedure picked the correct pair of variables in the first two stages in each test sample for $d = 4$ or 5, but performed no better than one would expect by chance for $d = 1$ or 2. When $d = 3$, the pair with cluster structure was picked correctly in nine cases out of 10 when $l = 2$ and seven times when $l = 4$.

Collectively, these experiments suggest that the forward selection algorithm can successfully cull out variables with cluster structure from noise variables provided the clusters are reasonably separated and even in the presence of moderate orientation and scale differences.

4. Variable Selection on Real Data

In this section the forward selection algorithm will be used to analyze data concerning the fading of microwave signals. One purpose of analyzing these data was to predict the amount of time, measured in seconds, that the strength of microwave signals between a pair of microwave towers falls below 30 decibels (fading time) in a given year as a function of the following seven explanatory variables:

1. Distance in miles between microwave towers (hop length),
2. Absolute humidity in mg/cubic meter,
3. Terrain roughness in meters,
4. Average annual temperature in degrees Fahrenheit,
5. Annual number of hot days (temperature $\geq 90°\mathrm{F}$) ,
6. Average annual rainfall in inches,
7. Average annual number of days with thunderstorms.

Data were available on 51 pairs of towers from across the United States concentrated in a relatively small number of geographical areas. There were several instances of repeated measurements of fading time at the same site over several periods of time.

Preliminary scatter plots of the explanatory variables indicated the possible presence of cluster structure in these variables. The seven explanatory variables thus seemed natural candidates for the investigation of whether the cluster structure was present in all seven variables or largely confined to a subset. Presence of cluster structure could have an important impact on the prediction of fading time. Prior to analysis, each of the seven variables was standardized to have unit standard deviation. Figure 9 shows plots of $S^*(k)$ versus k for each subset containing one variable. (For this example, any small, negative values were rounded to zero. This does not affect the choice of variables.) The collection of plots represents the first step in a forward selection process.

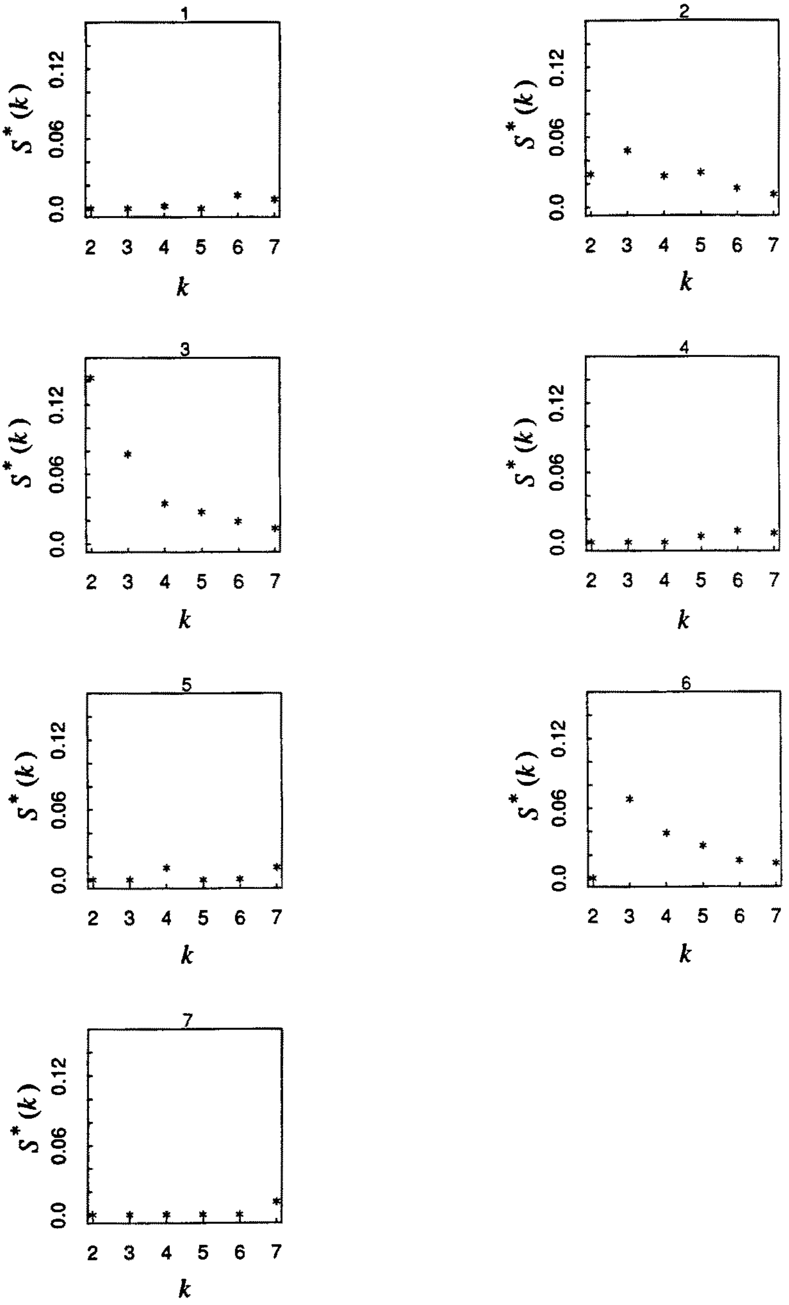

Figure 9. Plots of separation statistic versus number of clusters for selection of first variable. The number over each plot refers to the variable being considered.

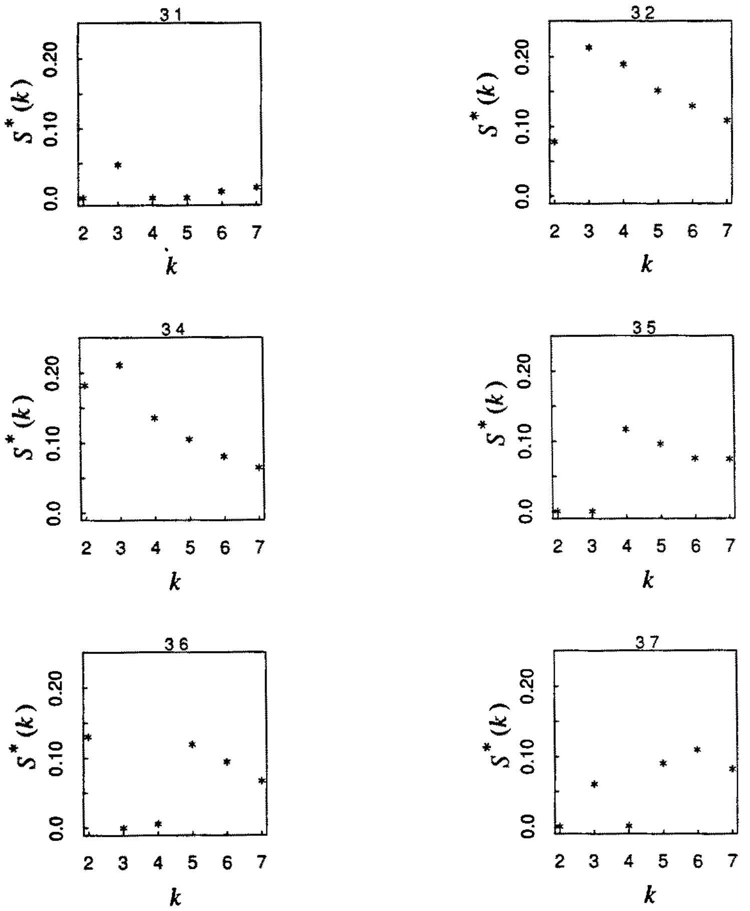

Figure 10. Plots of separation statistic versus number of clusters for selection of second variable.

The plot for variable 3, Terrain, exhibits the largest value, $S^*(2) = .14$, of those in Figure 9. (The standard deviation of $S^*(2)$ under null normal conditions is about .08 and drops to around .05 for higher values of k. Corresponding values for later stages of the selection process are about the same.) Terrain was entered at the first step in the forward selection.

Figure 10 shows plots of $S^*(k)$ versus k for all two variable subsets, given that variable 3, Terrain, was selected in the first step. The maximum value of $S^*(k)$ increased from approximately .14 at step one to approximately 22 at the second step. Given the rough estimates of standard deviation from the simulation experiment, this increase was deemed "significant."

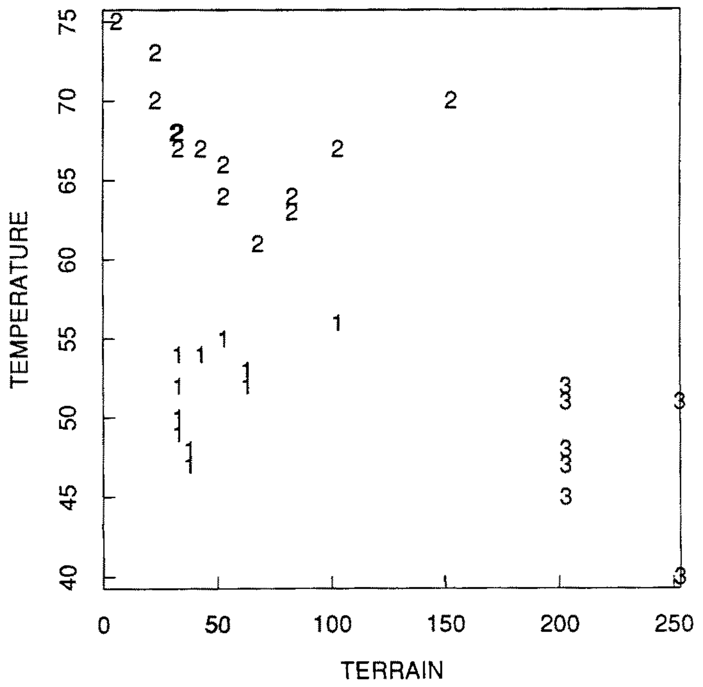

Figure 11. Scatter plot for checking reasonableness of the three clusters.

Inspection of Figure 10 shows that subsets {3,2} and {3,4} exhibit the largest values of the separation statistic, $S^*(k)$. The largest values are for $S^*(3)$. Arbitrarily, variable 4, Temperature, was chosen at the second step in preference to variable 2, Humidity. That this choice was not crucial will be evident later. Since only two variables are being considered at this step, the clustering should be clearly revealed by a scatter plot of Temperature versus Terrain. Figure 11 shows such a scatter plot with points identified by cluster number. (There appear to be fewer than 51 points because of repeated values and overstriking.) The clusters appear to be very well separated.

Having entered variables 3 and 4 (Terrain and Temperature) in the first two steps of forward selection, the selection was continued, and variable 2, Humidity, was selected at the third step. (Had Humidity been selected at the second step, then Temperature would have entered here.) Forward selection of variables was terminated after step four. Figure 12 shows plots of $S^*(k)$ versus k for each four variable subset containing the previously selected variables. The plot for variables 3,4,2,7, in which Thunderstorms is the fourth variable, exhibits the largest cluster separation so far encountered, $S^*(k) = .27$.

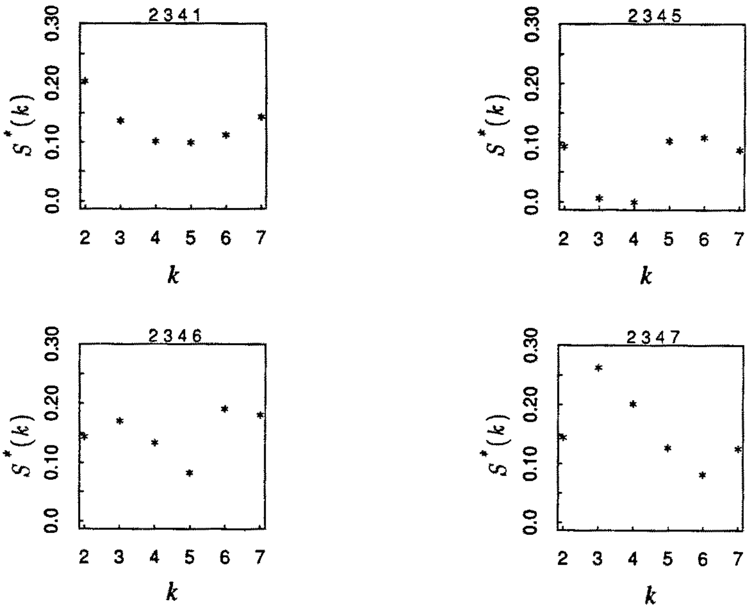

Figure 12. Plots of separation statistic versus number of clusters for selection of fourth variable.

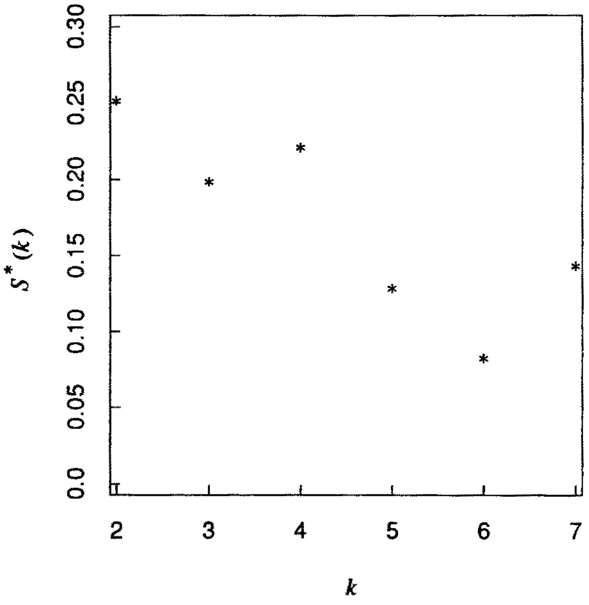

Figure 13. Plot of separation statistic versus number of clusters using all variables.

Any further addition of variables in the forward selection process decreases the value of the separation statistic. Consider Figure 13 which shows $S^*(k)$ versus k plots for all of the variables. There are sharp decreases in some of the $S^*(k)$ values from the ones for variable subset {3,4,2,7}. For example, $S^*(3)$ decreases markedly from the subset {3,4,2,7} to the full complement of seven variables. This suggests that the three cluster structure that was found for the subset {3,4,2,7} has been largely wiped out by the addition of the final three variables. In addition, the general level of $S^*(k)$ is markedly reduced from the subset {3,4,2,7} to the full complement. Overall, there appears to be weak clustering when all of the variables are used. Peaks occur at $k = 2$ and $k = 4$.

Given the subset {3,4,2,7}, attention was confined to the $k = 3$ cluster structure. The cluster memberships were determined by cutting the tree resulting from the hierarchical clustering of the four standardized variables (using Euclidean distance and the complete linkage method) to produce three clusters. In order to establish the reasonableness of the three cluster structure, the clusters were considered in more detail. In particular, the sites falling in the three clusters were studied, the separation of the clusters was assessed by methods other than the separation statistic $S^*(k)$, and characterizations of the clusters in terms of the four variables were also made. Figure 14 shows a map of the United States with the individual sites identified by the three cluster numbers. Cluster #2 comprised hot, rainy sites in the southeastern United States largely from the states of Florida, Georgia, and Alabama. Cluster #3 comprised mountainous regions of New York, Pennsylvania, New Mexico, and Wyoming, while Cluster #1 contained the remaining sites, largely cool and flat, scattered across the midwest, Texas, and New Jersey. There is thus a strong geographical-climatological component in the clustering since places that cluster together either tend to be close in terms of map distance or close in terms of terrain, temperature, humidity, and number of thunderstorms.

Figure 15 shows a cluster profile in which this geographical clustering is amplified. Values of the cluster means minus the grand mean for the standardized data are plotted, using their cluster numbers, on the X-axis while the variables Humidity, Terrain, Temperature, and Thunderstorms are positioned on the Y-axis. The plot is really a series of four unidimensional plots juxtaposed for comparison purposes. For example, the plot shows that Cluster #3 which includes sites in Pennsylvania, New York, Wyoming, and New Mexico has the roughest terrain, the lowest annual temperature, and the lowest humidity. Cluster #2, which comprises sites in Florida and other southeastern states, has the highest humidity, highest temperature, and flat terrain.

Figure 16 gives an overall assessment of cluster separation. It shows a plot of the Euclidean distances of each point to the centroid of the cluster in which it falls and to all other cluster centroids. The distances have been

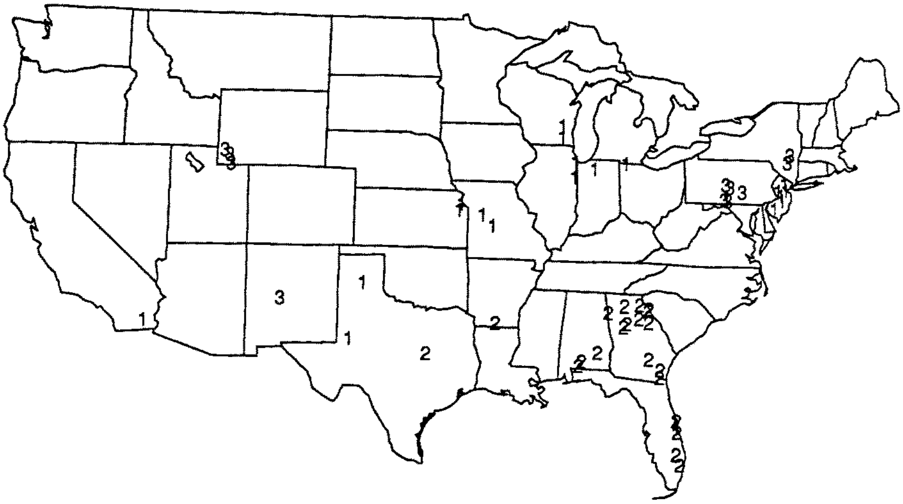

Figure 14. Geographical location of three clusters obtained using "best" four variables.

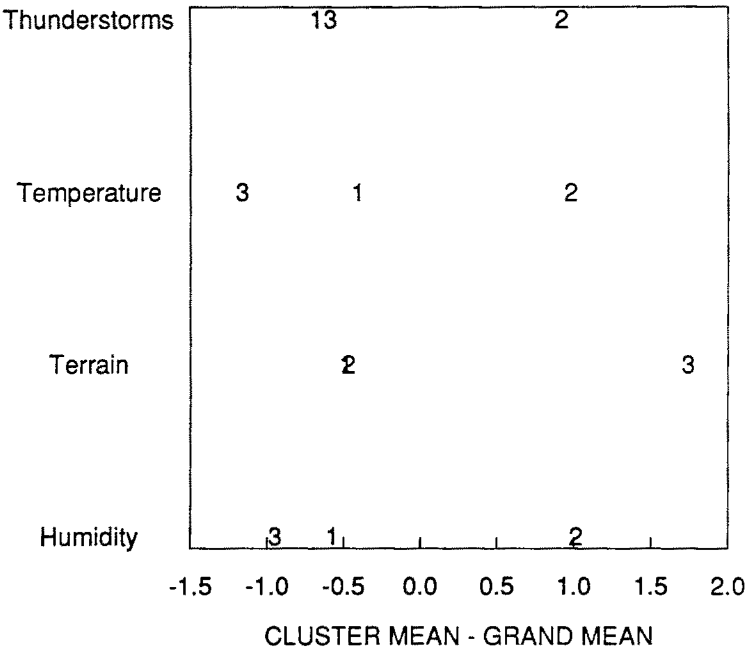

Figure 15. Profile plot that shows separation of clusters on individual variables.

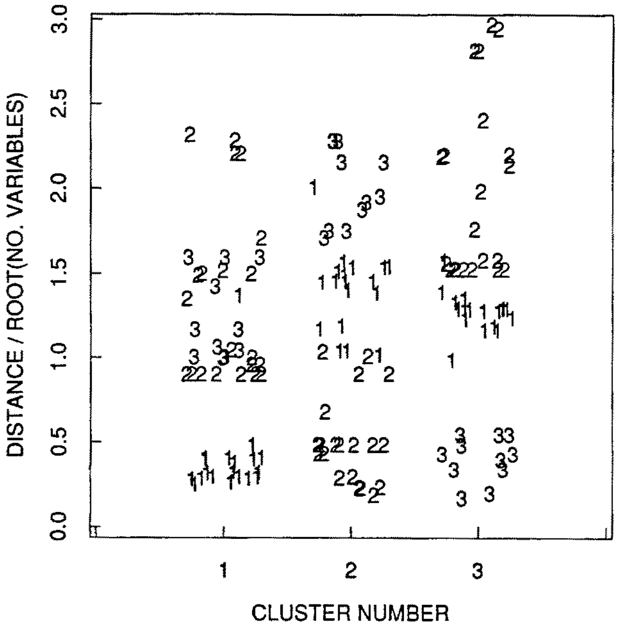

Figure 16. Distance plot of individual points to cluster centroids using "best" four variables.

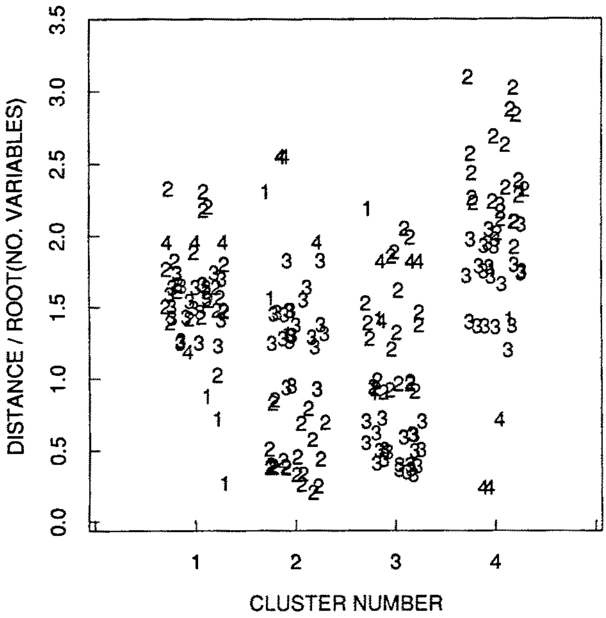

Figure 17. Distance plot of individual points to cluster centroids using all variables.

scaled by dividing by the square root of the number of variables used in the distance calculation. This facilitates the comparison of plots for differing numbers of variables. Points are again identified by cluster number, and a small uniformly distributed random variable is added to the X-coordinate so that the points may be more easily seen. The basic plot is repeated for each of the three clusters and the results are shown side-by-side. This plot is a variation of one proposed by Gnanadesikan, Kettenring, and Landwehr (1977). Beginning at the left in Figure 16 the plot shows that, for Cluster #1, the distances of the sites in Cluster #1 to their own centroid are much less than the distances of the other sites to this same centroid. The exception is one point in Cluster #1 which represents the single California site. Furthermore, there is a pronounced gap between the points for Cluster #1 and the others in this column. Cluster #2, comprising sites in the southeast, shows some slight overlap with Cluster #1, but for most points it is quite separated. Cluster #3, containing the mountainous sites, perhaps shows the largest separation of all. Its points are all much closer to the Cluster #3 centroid than to the centroids of the other two clusters.

Finally, a study was made to show how the three clusters that were found to be reasonable for the subset {3,4,2,7} are altered when all of the variables are included. From the plot of the separation statistic $S^*(k)$ versus the number of clusters k in Figure 13 there is a suggestion that there might be two or four clusters when all variables are considered. These clusterings ought to be somewhat weaker than the clustering for the subset {3,4,2,7} since the general level of $S^*(k)$ is reduced when all of the variables are entered. Limiting attention now to the four cluster case, the contents of these clusters can be compared to the three clusters found when the subset {3,4,2,7} was considered. The four component clustering derived from all of the variables left the cluster comprising sites in the southeastern U. S. (again Cluster #2) intact but removed the mountainous regions of New York and Pennsylvania from the mountainous cluster and combined them with flat regions in New Jersey and the midwest (new Cluster #3). This does not seem to be reasonable. Also a new cluster (Cluster #4) was formed containing two of the three Texas sites and the single California site. The separation of the four clusters can be assessed by constructing a distance plot, Figure 17, like that of Figure 16. The pronounced gaps found in Figure 16 for the three component clustering using subset {3,4,2,7} have now largely disappeared. Only Cluster #4 appears well separated.

This example has shown that the subset of variables, Humidity, Terrain, Temperature, and Thunderstorms contain well-separated and meaningful clusters. The remaining variables in part wipe out structure found for the subset and dilute the cluster separation.

5. Conclusion

The inclusion of unnecessary variables in a cluster analysis can cause more damage than in such other statistical procedures as regression analysis. However, constructing sound statistical procedures for variable selection in clustering appears to be particularly tricky. The contribution of this paper has been to propose a straightforward but computationally intensive procedure for use in conjunction with standard clustering algorithms. Experience with simulated and real data shows that it is effective.

References

ART, D., GNANADESIKAN, R., and KETTENRING, J. R. (1982), "Data-Based Metrics for Cluster Analysis," *Utilitas Mathematica, 21A*, 75-99.

DE SARBO, W. S., CARROLL, J. D., CLARK, L. A., and GREEN, P.E. (1984), "Synthesized Clustering: A Method for Amalgamating Alternative Clustering Bases with Differential Weighting of Variables," *Psychometrika, 49*, 57-78.

DE SOETE, G. (1986), "Optimal Variable Weighting for Ultrametric and Additive Tree Clustering," *Quality and Quantity, 20*, 169-180.

DE SOETE, G., DE SARBO, W. S. and CARROLL, J. D. (1985), "Optimal Variable Weighting for Hierarchical Clustering: An Alternating Least-Squares Algorithm," *Journal of Classification, 2*, 173-192.

FOWLKES, E. B., GNANADESIKAN, R., and KETTENRING, J. R. (1987), "Variable Selection in Clustering and Other Contexts," in *Design, Data, and Analysis,* ed. C. L. Mallows, New York: Wiley, pp. 13-34.

GNANADESIKAN, R. (1977), *Methods for Statistical Data Analysis of Multivariate Observations,* New York: Wiley.

GNANADESIKAN, R., KETTENRING, J. R., and LANDWEHR, J. M. (1977), "Interpreting and Assessing the Results of Cluster Analyses," *Bulletin of the International Statistical Institute, 47*, 451-463.

HARTIGAN, J. A. (1972), "Direct Clustering of a Data Matrix," *Journal of the American Statistical Association, 67*, 123-129.

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

MAC QUEEN, J. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1), eds. L. LeCam and J. Neyman, Berkeley: University of California Press, pp. 355-372.

MC KAY, R. J. and CAMPBELL, N. A. (1982), "Variable Selection Techniques in Discriminant Analysis. I. Description," *British Journal of Mathematical and Statistical Psychology, 35*, 1-29.

MILLIGAN, G. W. and COOPER, M. C. (1988), "A Study of Variable Standardization," *Journal of Classification, 5*, to appear.

PILLAI, K. C. S. (1955), "Some New Test Criteria in Multivariate Analysis," *Annals of Mathematical Statistics, 26*, 117-121.

ROY, S. N., GNANADESIKAN, R., and SRIVASTAVA, J. N. (1971), *Analysis and Design of Certain Quantitative Multiresponse Experiments,* Oxford: Pergamon Press.

SEBER, G. A. F. (1984), *Multivariate Observations,* New York: Wiley.