PSYCHOMETRIKA—VOL. 66, NO. 2, 249–270
JUNE 2001

A VARIABLE-SELECTION HEURISTIC FOR K-MEANS CLUSTERING

MICHAEL J. BRUSCO AND J. DENNIS CRADIT

FLORIDA STATE UNIVERSITY

One of the most vexing problems in cluster analysis is the selection and/or weighting of variables in order to include those that truly define cluster structure, while eliminating those that might mask such structure. This paper presents a variable-selection heuristic for nonhierarchical (K-means) cluster analysis based on the adjusted Rand index for measuring cluster recovery. The heuristic was subjected to Monte Carlo testing across more than 2200 datasets with known cluster structure. The results indicate the heuristic is extremely effective at eliminating masking variables. A cluster analysis of real-world financial services data revealed that using the variable-selection heuristic prior to the K-means algorithm resulted in greater cluster stability.

Key words: cluster analysis, K-means partitioning, variable selection, heuristics.

1. Introduction

The significant number of citations in the research literature easily supports the importance of cluster analysis for applications in social science, physical science, engineering, and business (see, for example, Waller, Kaiser, Illian, & Manry, 1998, for recent information regarding the volume of citations). In each of these disciplines, there is often a wide spectrum of candidate variables that can be used in the process of clustering objects. It has been frequently observed, however, that only a limited subset of variables is actually valuable in defining cluster structure (DeSarbo, Carroll, Clark, & Green, 1984; De Soete, DeSarbo, & Carroll, 1985; Gnanadesikan, Kettenring, & Tsao, 1995; Milligan, 1989). Further, the incorporation of variables that do not define true cluster structure may effectively complicate or obscure the recovery of this structure during a hierarchical or nonhierarchical cluster analysis (Milligan, 1980; Milligan, 1989). Fowlkes and Mallows (1983) referred to these complicating variables as "masking variables."

There are two broad approaches for identifying and mitigating the effect of masking variables; (a) variable weighting, and (b) variable selection. Variable-weighting methods, which attempt to differentially weight variables based on their relative ability to define cluster structure, have been developed and refined in the literature for quite some time (Anderberg, 1973; Art, Gnanadesikan, & Kettenring, 1982; Cormack, 1971; DeSarbo et al., 1984; De Soete et al., 1985; Friedman & Rubin, 1967; Gnanadesikan et al., 1995; Kruskal, 1972; Rohlf, 1970). One of the most popular of these methods for nonhierarchical (iterative K-means) clustering is SYNCLUS (DeSarbo et al., 1984), which simultaneously generates partitions and variable weights using a weighted K-means procedure. Weights are chosen to minimize a measure of stress through an iterative fitting process. Green, Carmone, and Kim (1990) evaluated SYNCLUS and observed that it worked well on one dataset, but that its performance on a second dataset was sensitive to initial seed points.

We gratefully acknowledge the constructive comments of three anonymous reviewers, the Associate Editor, and Editor, which led to considerable improvements in this article. We note that our variable-selection heuristic evolved during the review process. This evolution was attributable to a variety of factors including: (a) the publication of the HINoV procedure (Carmone et al., 1999), (b) a thoughtful comment from an anonymous reviewer regarding correlated masking variables, and (c) a helpful suggestion from the Associate Editor concerning multiple true cluster structures in a single dataset. Requests for reprints should be sent to Michael J. Brusco, Marketing Department, College of Business, Florida State University, Tallahassee, FL 32306-1110, E-Mail: mbrusco@cob.fsu.edu

0033-3123/2001-2/1999-0736-A $00.75/0

De Soete (1986) proposed an algorithm that identifies weights yielding Euclidean distances optimally suited for representation by an ultrametric tree. De Soete et al. (1985) and De Soete (1986) demonstrated that the weighting algorithm was effective in assigning near-zero weights to extraneous variables in a dataset. Milligan (1989) examined the utility of the De Soete (1986) algorithm as a variable-weighting procedure and reported that it was effective in reducing the influence of variables with little or no contribution to the true cluster structure.

Perhaps the most comprehensive investigation of variable-weighting methods was provided by Gnanadesikan et al. (1995), who tested eight variable-weighting schemes across a sample of simulated and real datasets. These weighting schemes included: equal-weight scaling; standardization of variables based on the standard deviation (autoscaling) or range; several alternatives based on the within- and/or between-sum-of-squares and cross-products matrices; SYNCLUS; and, De Soete's (1986) procedure. Their results revealed that the two latter procedures under-performed relative to schemes based on within and between sums-of-squares cross-products matrices. In fact, the De Soete (1986) algorithm frequently under-performed simpler schemes such as equal-weight scaling, autoscaling, and range-scaling.

One of the most significant findings of Gnanadesikan et al. (1995) was that variable-weighting schemes were often out-performed by a procedure based on variable selection. Variable-selection procedures attempt to define a subset of variables for use in a cluster analysis (Fowlkes, Gnanadesikan, & Kettenring, 1987; Fowlkes, Gnanadesikan, & Kettenring, 1988). In fact, such procedures can be viewed as a special case of variable weighting where all such weights are required to be 0 or 1. Variable selection has definite advantages over weighting approaches (Fowlkes et al., 1988; Gnanadesikan et al., 1995). For example, variable-selection procedures eliminate the need for future measurement of variables that do not define cluster structure. Typical applications require the inclusion of all clustering variables. Given the growing size of datasets and the use of datamining techniques in fields such as business (Berry & Linoff 1997; Blattberg, Glazer, & Little 1994), this seems particularly important. Perhaps more importantly, variable-selection procedures do not present the difficulties associated with trying to interpret the meaning of differentially weighted variables.

The Gnanadesikan et al. (1995) study evaluated a forward variable-selection procedure, developed by Fowlkes et al. (1988), which selects variables in an iterative manner based on a multivariate analysis-of-variance separation criterion. Fowlkes et al. (1988) characterize their procedure as "informal" because the selection of clustering variables is based upon graphical information. Although this approach often compared favorably to the variable-weighting procedures in the study, Gnanadesikan et al. (1995) found it "disappointing" that it was not consistently the best performing method. Another limitation they cited was the fact that the Fowlkes et al. (1988) procedure was limited to autoscaled data. Summing up the results of their testing of variable weighting and selection procedures, Gnanadesikan et al. (1995) concluded: "Considerable additional research in this general area seems, therefore, to be required" (p. 135). Milligan (1996) echoed this sentiment, noting "...more general approaches to variable weighting would make worthwhile contributions to the field of classification" (p. 352).

1.1. HINoV Method

Carmone, Kara, and Maxwell (1999) recently observed that, if the ultimate goal is the recovery of true cluster structure, then a good measure of actual recovery might be useful for guiding the selection of cluster variables to include in the analysis. Based on this principle, they developed a graphical variable-selection procedure (HINoV: heuristic identification of noisy variables) based on Hubert and Arabie's (1985) adjusted version of Rand's (1971) index for measuring the agreement of partitions (see Arabie & Hubert 1996; Helsen & Green, 1991; Krieger & Green, 1999; Milligan 1989; Milligan & Cooper 1986; Salstone & Stange 1996 for discussions of the effectiveness of the adjusted Rand index). A description of this procedure requires the following notation:

$M =$ the number of objects (observations, cases, customers, etc.) in the dataset, indexed $i = 1, \ldots, M$;

$D =$ the number of variables (or dimensions) on which the objects are measured, indexed $j = 1, \ldots, D$. Additionally $D = D_1 + D_2$, where D_1 is the number of true (structure) variables and D_2 is the number of masking (noise) variables.

$\mathbf{X} =$ an $M \times D$ matrix of measurements, where x_{ij} denotes the measurement of object i on variable j; for $i = 1, \ldots, M$ and $j = 1, \ldots, D$.

$C =$ the number of clusters in the dataset indexed $c = 1, \ldots, C$;

$\mathbf{p_j} =$ a $(1 \times M)$-dimensional row vector that defines a partition developed using only variable j. Elements of the vector, p_{ji}, designate the cluster to which object i is assigned.

$\mathbf{R} =$ A symmetric $D \times D$ matrix (with zeros on the main diagonal) of adjusted Rand indices, where $r_{jk} = r_{kj}$ is the adjusted Rand index associated with partitions $\mathbf{p_j}$ and $\mathbf{p_k}$, for $j = 1, \ldots, D-1$ and $k = j+1, \ldots, D$;

The HINoV procedure is initiated with the development of a K-means partition, $\mathbf{p_j}$, using only variable j, for each of the $j = 1, \ldots, D$ variables. The next step is to compute the adjusted Rand index, $r_{jk} = r_{kj}$, between each of the $(D(D-1)/2)$ pairs of partitions. A total pairwise adjusted Rand index ($TOPRI_j$) is then computed for each variable as follows:

$$TOPRI_j = \sum_{k=1}^{D} r_{jk}.$$

Variables are subsequently selected (in a single pass rather than iteratively) based on a scree plot of the ranked $TOPRI_j$ values. According to the authors, low-value TOPRI variables (i.e., the noisy variables) are identified and eliminated from analysis. The procedure terminates by running K-means with only the selected variables.

Limitations to HINOV. In addition to the fact that it makes use of the well-justified adjusted Rand index, HINoV is straightforward and efficient in its implementation. However, the procedure does have some serious limitations. First, like the method of Fowlkes et al. (1988), HINoV is informal and subjective because the variable-selection process requires the interpretation of scree plots. This makes it difficult for researchers in this area to replicate the findings of the Carmone et al. (1999) study, or to evaluate HINoV against other methods in a Monte Carlo comparison. Second, and much more importantly, the computation of the $TOPRI_j$ sums is predicated upon the assumption that the pairwise adjusted Rand indices will be large (close to 1.0) for a pair of true variables, and small (close to 0) for a pair of masking variables (or a pair consisting of one masking and one true variable). This may not always be true.

As we will see, HINoV is particularly prone to failure under two sets of data conditions that are clearly possible to arise in realistic datasets: (a) a high degree of correlation among the masking variables, and (b) multiple sets of true cluster structures in the same dataset. When two or more masking variables are highly correlated, they may lead to roughly the same cluster structure and thus the adjusted Rand index associated with partitions based on these variables could be quite large. This can subsequently lead to their erroneous inclusion in the set of clustering variables.

To illustrate, consider an example with $M = 9$ objects measured on $D = 4$ variables (v_1, v_2, v_3, and v_4) as shown in Table 1. The plot of v_2 versus v_1 in Figure 1 reveals $C = 3$ well-separated homogeneous clusters, which is the "true" structure in the dataset. Thus, v_1 and v_2 are the *true variables* in this example. A plot of the *masking variables* (v_4 vs. v_3) in Figure 2 shows that, although these two variables do not define as nice a structure as v_1 and v_2, they are *highly correlated*. This correlation has serious implications for HINoV. For example, if K-means

TABLE 1.
A small example data set

Object	1	2	3	4	5	6	7	8	9
v_1 measure	6	7	8	2	3	4	12	14	14
v_2 measure	14	15	13	3	1	2	3	4	2
v_3 measure	15	3	10	5	11	7	13	6	1
v_4 measure	15	4	10	6	12	8	12	7	1

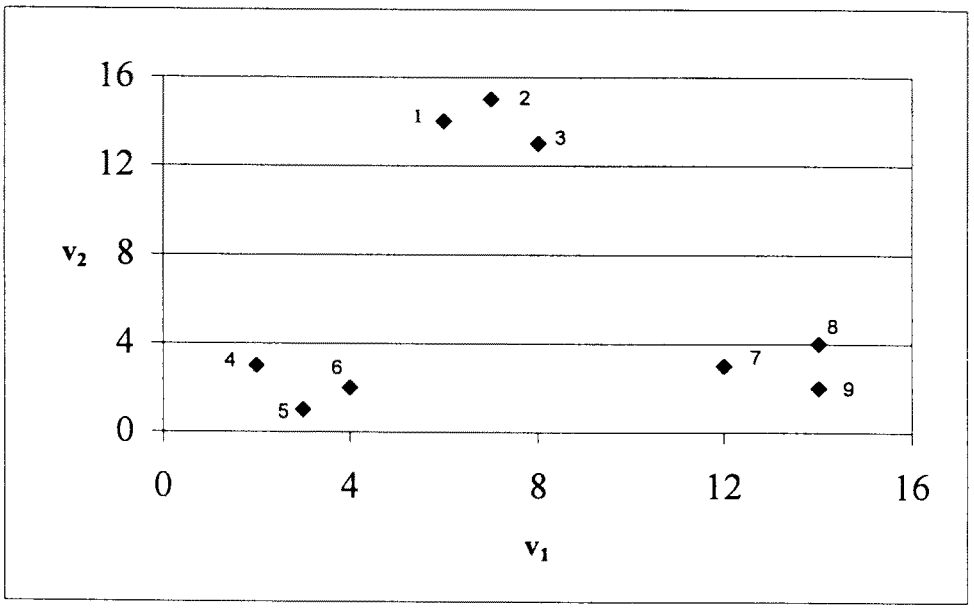

FIGURE 1.
A plot of the true variables for the example data set.

partitions based only on variable j are developed for each variable j ($j = 1, \ldots, D$) using $C = 3$ clusters, the resulting partitions are $\mathbf{p_1} = [1, 1, 1, 2, 2, 2, 3, 3, 3]$, $\mathbf{p_2} = [1, 1, 1, 3, 2, 2, 3, 3, 2]$, and $\mathbf{p_3} = \mathbf{p_4} = [1, 2, 1, 3, 1, 3, 1, 3, 2]$. The matrix of adjusted Rand indices corresponding to each pair of single variable partitions, along with corresponding $TOPRI_j$ values is shown below:

$$\mathbf{R} = \begin{bmatrix} 000 & .407 & -.071 & -.071 \\ .407 & .000 & -.071 & -.071 \\ -.071 & -.071 & .000 & 1.000 \\ -.071 & -.071 & 1.000 & .000 \end{bmatrix} \qquad \begin{bmatrix} TOPRI_1 = .265 \\ TOPRI_2 = .265 \\ TOPRI_3 = .857 \\ TOPRI_4 = .857 \end{bmatrix}.$$

Observe that the largest $TOPRI_j$ values correspond to variables v_3 and v_4. Thus, HINoV would incorrectly select the masking variables for inclusion and eliminate the true variables. This example leads us to conclude that there are at least two potential reasons why the procedure is prone to failure. First, two variables (v_1 and v_2 in this example) can define true cluster structure, yet when these variables are used independently to develop partitions, the adjusted Rand index associated with such partitions can be rather small. Indeed, Figure 3 reveals a situation where $r_{12} \approx 0$, yet variables v_1 and v_2 again define a well-separated, homogeneous structure. Second, masking variables can have a very large Rand index due to high correlation, thus leading to their erroneous inclusion.

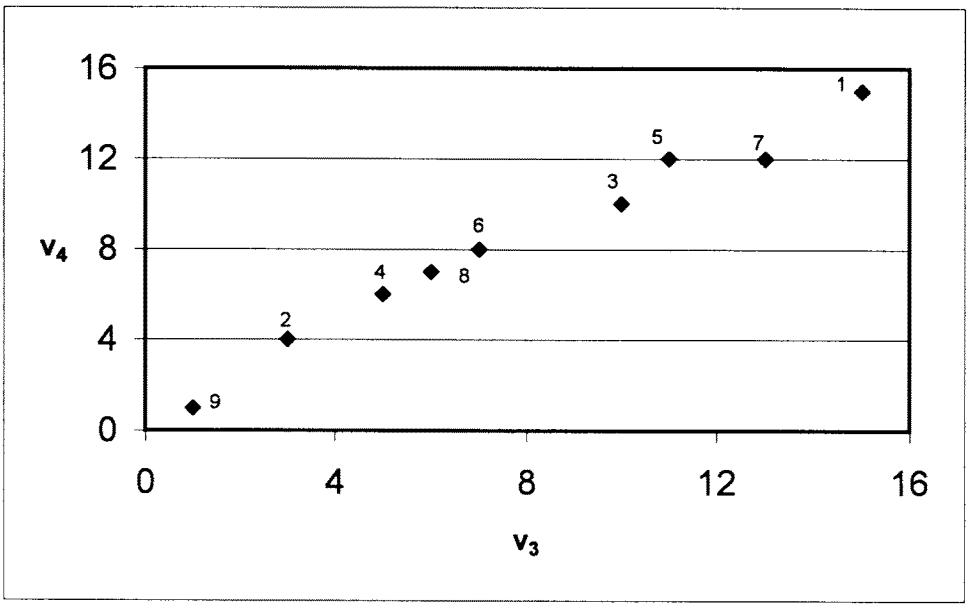

FIGURE 2.
A plot of the masking variables for the example data set.

For similar reasons, HINoV is also prone to failure when there are multiple "true" structures in the same dataset. In such cases, it is desirable to recover at least one of the true structures. Unfortunately, HINoV's use of aggregated sums of adjusted Rand indices can easily result in the selection of variables from two or more true structures, which impedes the recovery of at least one of the structures. Although it is easy to do so, for the sake of brevity we do not include a

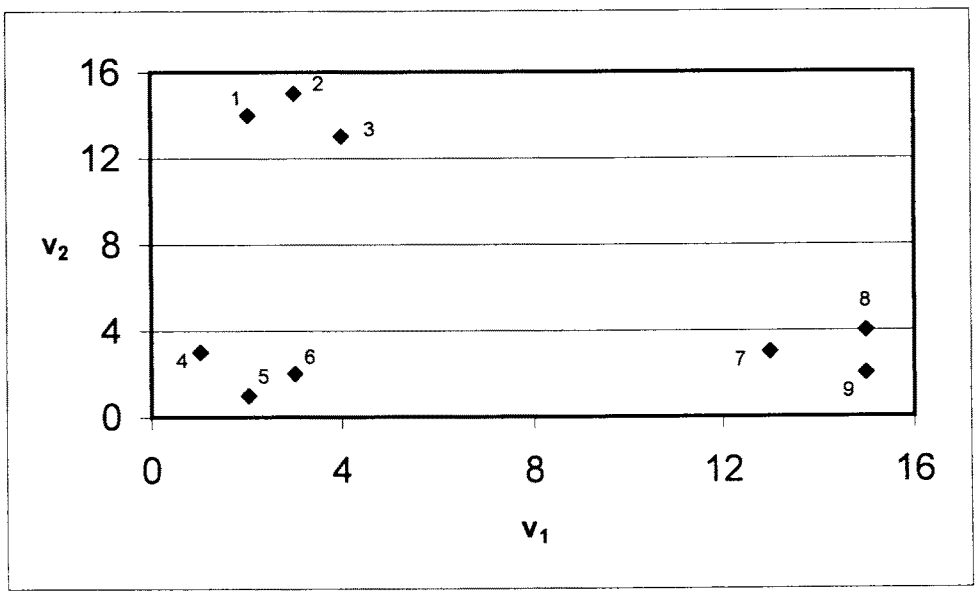

FIGURE 3.
A plot of true variables with low adjusted Rand Index for the single-variable partitions.

numerical example of this situation. However, we do provide an empirical comparison of HINoV and our new procedure across datasets with two true structures in section 6.2.

1.2. Purpose

Our objective in this paper is to develop a heuristic variable-selection procedure that can be successfully applied to large datasets. We refer to this procedure as VS-KM (variable-selection heuristic for K-means clustering). VS-KM builds on the strengths of earlier variable-selection methods (Carmone et al., 1999; Fowlkes et al., 1988), while overcoming some of their weaknesses. Like HINoV, VS-KM utilizes the adjusted Rand index. Unlike HINoV, VS-KM adds variables in a forward (stepwise) manner and also uses information about the between-cluster and total sum-of-squares, similar to the Fowlkes et al. (1988) method. VS-KM is based on the premise that the adjusted Rand index information is better used in a forward-selection process. To illustrate this logic, consider a situation where three variables have already been selected for inclusion and a fourth variable, j, is currently under consideration for inclusion. Suppose that one partition is developed using the three selected variables and a second partition is developed using only variable j. The adjusted Rand index is then used to compute the agreement between these two partitions. A large adjusted Rand index would suggest that adding variable j to the mix of variables would not mask the current cluster structure. Alternatively, if the adjusted Rand index is small, it can be concluded that variable j does mask the current structure and should not be added to the set of clustering variables.

In the next section we present the details of VS-KM. We describe an initial Monte Carlo study designed to test VS-KM in section 3, the results of which are provided in section 4. We then conduct a second study to extend our investigation to much larger datasets (up to 7000 objects). A description of this second study and the corresponding results are presented in section 5. Section 6 provides a third study that directly compares HINoV with VS-KM. Section 7 completes the testing with a real-world demonstration that clusters financial institutions in a business-to-business strategy context. The paper concludes in section 8 with a summary of the findings, an overview of potential areas of application, and a discussion of the limitations and possible extensions of our investigation.

2. Variable-selection Heuristic—VS-KM

The description of VS-KM uses the notation from section 1 in addition to the following:

Notation:

$\mathbf{S} =$ the set of variables selected for inclusion in the cluster analysis;

$\mathbf{U} =$ the set of unselected variables, $\mathbf{S} \cup \mathbf{U} = \{1, 2, \ldots, D\}$ and $\mathbf{S} \cap \mathbf{U} = \{\emptyset\}$;

$\mathbf{w_{jk}} =$ a $(1 \times M)$-dimensional row vector that defines a partition developed using variables j and k, for $j = 1, \ldots, D-1$ and $k = j+1, \ldots, D$. Elements of the vector, w_{jki}, designate the cluster to which object i is assigned;

$\mathbf{Q} =$ a $D \times D$ symmetric matrix with zeros on the main diagonal and remaining elements $q_{jk} = q_{kj}$ defining the ratio of the between cluster sum-of-squares to the total sum-of-squares for partition $\mathbf{w_{jk}}$, for $j = 1, \ldots, D-1$ and $k = j+1, \ldots, D$;

$T =$ a threshold value for the adjusted Rand index when selecting the first pair of variables;

$\mathbf{y} =$ an $(1 \times M)$-dimensional row vector that defines a partition developed using variables $j \in S$. Elements of the vector, y_i, designate the cluster to which object i is assigned.

$G_j =$ the adjusted Rand index associated with partitions $\mathbf{p_j}$ and $\mathbf{y}$, for $j = 1, \ldots, D$;

$G_{min} =$ the minimum allowable value of G_j such that variable j can be selected for inclusion in the cluster analysis;

$G_{fac} =$ a factor that when multiplied by the G_j value of the most recently selected variable serves as a threshold for the inclusion of the next variable.

Heuristic Algorithm

Step 0. Initialize. $\mathbf{y} = \mathbf{0}$; $\mathbf{p_j} = 0 \forall j = 1, \ldots, D$; $\mathbf{w_{jk}} = 0, \forall j = 1, \ldots, D-1$ and $k = j+1, \ldots, D$; $\mathbf{S} = \{\emptyset\}$, $\mathbf{U} = \{1, 2, 3, \ldots, D\}$.

Step 1. Develop a partition, $\mathbf{p_j}$, of C clusters using only variable j, for $j = 1, \ldots, D$.

Step 2. Compute the adjusted Rand index for all $D(D-1)/2$ pairs of partitions $\mathbf{p_j}$ and $\mathbf{p_k}$, for $j = 1, \ldots D-1$ and $k = j+1, \ldots, D$.

Step 3. Develop a partition, $\mathbf{w_{jk}}$, of C clusters using variables j and k and compute q_{jk}, for $j = 1, \ldots D-1$ and $k = j+1, \ldots, D$.

Step 4. If $\mathbf{Max}_{j,k}(r_{jk}) \geq T$, then let $\delta = \underset{j,k}{\mathbf{Max}}(q_{jk} | r_{jk} \geq T)$, else let $\delta = \underset{j,k}{\mathbf{Max}}(q_{jk})$. Let j' and k' denote variables such that $q_{j'k'} = \delta$, and let $\eta = r_{j'k'}$. Set $\mathbf{S} = \mathbf{S} \cup \{j', k'\}$ and $\mathbf{U} = \mathbf{U} - \{j', k'\}$.

Step 5. Develop a partition, $\mathbf{y}$, of C clusters using all selected variables $j \in \mathbf{S}$.

Step 6. For each unselected variable $j \in \mathbf{U}$, compute the adjusted Rand index, G_j, between partitions $\mathbf{p_j}$ and $\mathbf{y}$.

Step 7. Let $\lambda = \underset{j \in U}{\mathbf{Max}}(G_j)$. If $\lambda < G_{min}$, or $\lambda < \eta \bullet G_{fac}$, then go to Step 8. Otherwise, let j' denote the variable for which $G_{j'} = \lambda$, set $\eta = \lambda$, and set $\mathbf{S} = \mathbf{S} \cup \{j'\}$ and $\mathbf{U} = \mathbf{U} - \{j'\}$. If $\mathbf{U} = \{\emptyset\}$, then go to Step 8. Otherwise, return to Step 5.

Step 8. Variables in $\mathbf{S}$ are selected for inclusion and variables in $\mathbf{U}$ are discarded. A K-means cluster analysis is run using only the variables in $\mathbf{S}$.

In Step 1 of VS-KM, a partition ($\mathbf{p_j}$) of C clusters is developed for each individual variable j ($j = 1, \ldots, D$). In this study, a convergent version of MacQueen's (1967) K-means algorithm was used to generate these partitions. The algorithm, which seeks to minimize a total within-cluster sum-of-squares criterion, converges to a local (but not necessarily global) minimum. It has been observed that K-means clustering procedures are often very sensitive to the initial centroids that are used when implementing the algorithm (Green et al., 1990; Helsen & Green, 1991; Waller et al., 1998). Therefore, we used Ward's (1963) hierarchical clustering procedure to generate the initial centroids. Consistent with previous studies (Helsen & Green, 1991; Waller et al., 1998), Ward's algorithm was applied and the resulting tree was cut at the known number of clusters.

In Step 2 of VS-KM, the adjusted Rand index (r_{jk}) is computed for each of the $D(D-1)/2$ pairs of partitions. In step 3, a K-means partition ($\mathbf{w_{jk}}$) is developed using each possible pair of variables and the ratio of the between-cluster sum-of-squares to the total-sum-of-squares (q_{jk}) is computed. Going into Step 4, there are two relevant pieces of information for each possible pair of variables: (a) the pairwise adjusted Rand index, and (b) an estimate of cluster homogeneity from the sum-of-squares ratio. Both pieces of information are factored into the selection process in Step 4. Specifically, the pair of variables resulting in the largest value of q_{jk} is selected for inclusion, provided that its r_{jk} achieves a threshold value (T). The goal is to select a pair of variables with a reasonable degree of agreement yet, at the same time, a homogeneous clustering. If no pair of variables achieves the threshold value, then the pair with the largest q_{jk} value is selected.

An important distinction between VS-KM and HINoV method is that the first two variables are selected jointly (as a pair) and the remaining variables are added independently (one at a time) in Steps 5 through 8. In Step 5, all the selected variables are used to generate a partition, $\mathbf{y}$, using the same procedure described above for Step 1. In Steps 6 and 7, adjusted Rand indices are computed, for each unselected variable, between the individual variable partitions and $\mathbf{y}$. Variables are added, in an iterative manner, until no adjusted Rand index in Step 6 is greater than G_{min}, or the greatest adjusted Rand index is less than the product of G_{fac} and the adjusted Rand

index associated with the most recently selected variable. VS-KM terminates in Step 8 with a set of selected variables for a K-means cluster analysis.

2.1. Application of the VS-KM to the Example Dataset

VS-KM was applied to the example dataset in Section 1. Since Steps 1 and 2 of VS-KM are the same as those in HINoV, we obtain the same pairwise Rand index matrix, **R**, as reported above. In Step 3 of VS-KM, K-means partitions are developed using all possible pairs of variables and the sum-of-squares ratio, $q_{jk} = q_{kj}$, is computed for each pair. The resulting matrix **Q** is obtained for the example dataset:

$$\mathrm{Q} = \begin{bmatrix} - & .971 & .737 & .709 \\ .971 & - & .771 & .774 \\ .737 & .771 & - & .884 \\ .709 & .774 & .884 & - \end{bmatrix}.$$

Observe that the largest value in the matrix above is $q_{12} = q_{21}$. For any threshold value $T \leq .407$, this means that variables v_1 and v_2 would be initially selected for inclusion in Step 4. VS-KM subsequently proceeds to Step 5, where a partition **y** is constructed for these selected variables (at this stage, this is done by simply noting that $\mathbf{y} = \mathbf{w_{12}}$). In Step 6, the adjusted Rand indices are computed between partitions **y** and $\mathbf{p_3} = -.07143$, and **y** and $\mathbf{p_4} = -.07143$. Since both adjusted Rand indices are quite low, neither v_3 or v_4 would be added to the set of selected variables, and VS-KM terminates with the correct selection of v_1 and v_2.

We note that scenarios can also be constructed where VS-KM can fail. For example, consider the case of Figure 3. In this scenario, the adjusted Rand index between variables v_1 and v_2 is $-.107$, far below most threshold values. Indeed, it appears that for the data in Figure 3 an informal inspection of the graph is more effective than either HINoV or VS-KM.

Like the graphical method of Fowlkes et al. (1988) and HINoV, VS-KM could be implemented as an informal approach to variable selection. An informal implementation precludes the need to specify the T, G_{min} and G_{fac} parameters. Instead, the researcher could generate tabular information regarding the adjusted Rand indices and between-cluster-to-total sum-of-squares ratios for combinations of variables, and make subjective decisions regarding which combinations are useful for defining cluster structure. Unfortunately, such an approach is not conducive to a large computational study and, therefore, we formalize VS-KM by specifying a definitive procedure for selecting variables.

In this formalized version of VS-KM, there are three parameters that must be specified by the research analyst: T, G_{min} and G_{fac}. These parameters control the selection process and jointly determine when no more variables are to be included. The goal is to set the parameters such that true variables will meet the criteria for inclusion, but masking variables will not. For any given set of parameters, it would be possible to construct a dataset for which that particular parameter set would not meet this goal. Nevertheless, it might be useful to have a general idea of reasonable settings for T, G_{min} and G_{fac} and this is investigated in the computational study described in section 3.

3. Study I. Methodology

3.1. Data Generation

Our initial Monte Carlo study was similar to Milligan's (1989) assessment of De Soete's (1986) ultrametric algorithm for providing near-zero weights for masking variables. We wanted to ascertain the effectiveness of VS-KM for eliminating masking variables from datasets with known cluster structure. We generated 1701 datasets (each consisting of 500 objects, not including outliers) consistent with the process described by Milligan (1985), which has been utilized

in a number of previous studies (Balakrishnan, Cooper, Jacob, & Lewis, 1994; Carmone et al., 1999; Helsen & Green, 1991; Milligan, 1980, Milligan, Soon, & Sokol, 1983, Milligan & Cooper, 1986; Milligan & Cooper, 1989; Waller et al., 1998).

Five primary factors were manipulated. The first factor, the number of clusters in the datasets, was examined at three levels, $C = 3$, 4, and 5. The second factor, density of the clusters, was tested at 3 levels; (a) an equal number of objects in each cluster, (b) 10% of the objects in one cluster and an equal division of the remaining objects across the remaining clusters, and (c) 60% of the objects in one cluster and the remaining objects equally divided across the remaining clusters. The third factor, the number of outliers in the dataset, was evaluated at three levels; (a) no outliers, (b) 20% outliers, and (c) 40% outliers. The fourth factor, the number of true structure variables, was evaluated at three levels, $D_1 = 4$, 6, and 8. The fifth factor manipulated in the study, the one of particular interest, was the number of masking variables and the correlation among them. Based on the results of the Gnanadesikan et al. (1995) study, we believed that it was important to incorporate problems for which the number of masking variables equaled or exceeded the number of true variables, as well as problems with more true variables than masking variables. Further, we extend previous research in variable weighting and selection methods (Milligan, 1989; Carmone et al., 1999) by considering different levels of correlation among the masking variables. Seven levels of the masking variable factor were considered; (a) $D_2 = 0$, (b) $D_2 = 2$, low correlation, (c) $D_2 = 2$, high correlation, (d) $D_2 = 4$, low correlation, (e) $D_2 = 4$, high correlation, (f) $D_2 = 6$, low correlation, and (g) $D_2 = 6$, high correlation. This produced a $3 \times 3 \times 3 \times 3 \times 7-$ level design resulting in 567 data scenarios. In addition, three replications were made of each scenario resulting in 1701 datasets.

The Fortran source code, as well as an executable file, for generating Milligan's (1985) test datasets can be obtained from the website: www.pitt.edu/~csna/Milligan/readme.html. We developed our own generation program but used precisely the same methods for generating uniform random variates (Knuth, 1997) and normal random variates (Box & Muller, 1958) that Milligan and his colleagues have used. Our generation program accommodates very large data arrays, as well as correlation among masking variables. The masking variables in Milligan's data generation program are generated, independently, by drawing from a uniform distribution on the range of the first variable. This leads to rather low correlation (average of .031) among the masking variables and served as our "low correlation" setting for the masking variables. For the "high-correlation" problems, the first masking variable was generated using the uniform distribution on the range of the first variable, but subsequent masking variable measurements for each object were generated by perturbing the measurement value for the immediately preceding masking variable by $+/-$ 50% (again using a uniform distribution). This resulted in an average correlation among the masking variables of approximately 0.598 for the high setting.

3.2. Characteristics of the Dataset

A preliminary exploration of the datasets was conducted to assess two issues regarding the selection of variables: (a) the effect of omitting a true variable from the set of clustering variables (Type I error), and (b) the effect of including a masking variable in the set of clustering variables (a Type II error). This exploration was conducted using one of the replicates of the 567 design points. For each of the 567 datasets, the K-means algorithm was run under four sets of conditions: (a) using all of the true variables, (b) using all of the true variables except the first true variable, (c) using all of the true variables except the second true variable, and (d) using all of the true variables and one of the masking variables. The average adjusted Rand indices associated with these four conditions are reported in Table 2.

Table 2 reveals that the average adjusted Rand index is .8937 when **S** contains all of the true variables and no masking variables. If variable $j = 1$ is deleted from **S**, this average drops to .7294, whereas if $j = 2$ the average drops to .8821. This finding is easily explained by Milligan's

TABLE 2.
An exploration of Type I and Type II errors in variable selection

	Selected Variables			
	All true variables	All true variables except $j = 1$	All true variables except $j = 2$	All true variables and one masking variable
Average adjusted Rand Index*	.8937	.7294	.8821	.6037

*The average is computed across 567 data sets defined by the experimental conditions.

(1985) generation procedure, which requires complete separation of the clusters on the first true variable, but not on any of the other true variables. Thus, the probability of a Type I error appears to be negligible for all but the first true variable. In other words, as long as the first true variable is selected, other true variables can be omitted without significant degradation in cluster recovery.

Table 2 also shows that the inclusion of just one masking variable in **S** can cause serious recovery problems even when all true variables are also contained in **S**. The decrease in the average adjusted Rand index from .8937 to .6037 when just one masking variable is included suggests that Type II error is much more serious than Type I error, at least for datasets generated in this manner.

3.3. Clustering Methods and Computer Implementation

The *control procedure* for the Monte Carlo study was a convergent version of MacQueen's (1967) K-means procedure using all variables (both true and masking) with equal weights. As was described for Steps 1 and 4 of the heuristic algorithm, initial centroids for the K-means procedure were generated using Ward's (1963) hierarchical method and cutting the resulting tree at the correct number of clusters. It was not computationally practical to apply Ward's method based on all M objects. Therefore, along the lines of Helsen & Green (1991), a random sample of 100 objects was taken and Ward's procedure was used to generate initial centroids based only upon this sample. The objects in the sample were selected based on the uniform probability distribution, again using Knuth's (1997) uniform random number generation procedure. The centroids determined on the basis of this sample were subsequently used as a starting point by the K-means algorithm to cluster all M objects. Throughout the remainder of this paper, we will refer to the control procedure as ALL-KM because all variables are used for the K-means clustering process.

The *experimental procedure* was VS-KM, which used only variables selected in Steps 1–7. In addition, VS-KM, which also uses Ward's method to compute initial centroids, had to be modified to use a sample of objects. Specifically, the single variable partitions, $\mathbf{p_j}$, as well as the selected variable partitions, $\mathbf{y}$, were developed based on a random sample of 100 objects.

In the initial Monte Carlo study, we wanted to provide some evaluation as to the necessity of the threshold parameter, T. Therefore, two versions of VS-KM were tested. The first version, VS-KM1, assumes that $T = -\infty$ and thus the first pair of variables selected by the algorithm is based solely on the ratio of between cluster sum-of-squares to total-sum-of squares. In the second version, VS-KM2, we used $T = .25$. This parameter setting ensures that, in addition to the sum-of-squares ratio, the first pair of variables selected also has reasonable pairwise adjusted Rand index values.

The stopping criteria for both VS-KM1 and VS-KM2 were based on parameter values of $G_{min} = .05$ and $G_{fac} = 0.5$. The justification for these parameter settings is twofold. First, the results in Table 2 suggested that Type II errors are somewhat more serious than Type I errors and we wanted a G_{fac} setting that guarded against relatively large decreases in the adjusted Rand indices during the selection process. Second, we tested other parameter settings in this

neighborhood, including the less restrictive settings of $G_{min} = .03$ and $G_{fac} = .3$ and the more restrictive settings of $G_{min} = .07$ and $G_{fac} = .7$. We found that the sensitivity of the heuristic to changes of parameter settings in this range was not particularly severe.

The data generation program, K-means clustering procedure, VS-KM, and adjusted Rand index program were all written in Fortran. The source codes are available from the authors via email at mbrusco@cob.fsu.edu. The entire study was conducted using a 400 MHz Pentium II microcomputer running at the DOS prompt in a Windows 98 operating environment. Data were collected regarding true cluster recovery (as measured by the adjusted Rand index) and CPU time for each of the 1701 test problems, and the three solution procedures (VS-KM1, VS-KM2, and ALL-KM).

4. Results of Study I

The results of Study I are summarized in Tables 3 and 4. Table 3 reports the analysis of variance results (main effects and two-way interactions) for the experimental study. All main effects were statistically significant, with the largest main effect corresponding to the variable-selection procedure. A pairwise comparison of means using Tukey's procedure indicated that all pairs of means associated with the solution procedures were statistically different. The results in Table 3 also reveal that most two-way interactions were significant. Most notable among these interactions are all three combinations of two-way interactions associated with the variable-selection procedure, masking variable, and outlier factors. These large interactions are readily explained by the data in Table 4. For example, the large A × F interaction term can be partially explained

TABLE 3.
Results of study I. Analysis of variance

Source	DF	SS	MS	F	P
Masking variable (A)	6	201.3540	33.5590	789.83	0.000
Number of clusters (B)	2	4.8672	2.4336	57.28	0.000
Cluster density (C)	2	4.0117	2.0058	47.21	0.000
True variables (D)	2	4.4063	2.2032	51.85	0.000
Outliers (E)	2	51.3734	25.6867	604.55	0.000
Variable selection (F)	2	192.9316	96.4658	2270.39	0.000
(A × B)	12	0.8644	0.0720	1.70	0.061
(A × C)	12	0.9572	0.0798	1.88	0.032
(A × D)	12	1.2648	0.1054	2.48	0.003
(A × E)	12	19.6078	1.6340	38.46	0.000
(A × F)	12	39.7549	3.3129	77.97	0.000
(B × C)	4	5.8103	1.4526	34.19	0.000
(B × D)	4	0.6264	0.1566	3.69	0.005
(B × E)	4	2.0439	0.5110	12.03	0.000
(B × F)	4	3.0778	0.7694	18.11	0.000
(C × D)	4	4.6231	1.1558	27.20	0.000
(C × E)	4	0.3213	0.0803	1.89	0.109
(C × F)	4	1.3560	0.3390	7.98	0.000
(D × E)	4	0.4383	0.1096	2.58	0.036
(D × F)	4	1.8242	0.4560	10.73	0.000
(E × F)	4	20.6315	5.1579	121.39	0.000
Error	4986	211.8484	0.0425		
Total	5102	773.9943			

TABLE 4.
Results of Study I: True cluster recovery results*

		Adjusted Rand Index (averages)			Percentage of perfect recoveries		
Factor	Level	ALL-KM	VS-KM1	VS-KM2	ALL-KM	VS-KM1	VS-KM2
Number of clusters	3	.3628	.6465	.7563	4.41	25.57	26.63
	4	.3515	.7199	.8150	4.59	27.51	28.22
	5	.3576	.7825	.8522	4.94	31.04	31.92
Number of true variables	4	.3033	.6838	.7698	4.06	25.40	26.98
	6	.3577	.7593	.8282	4.41	28.75	29.63
	8	.4109	.7059	.8255	5.47	29.98	30.16
Density	Even	.3840	.7413	.8182	5.29	30.16	30.86
	10%	.3999	.7182	.8201	4.23	29.10	30.51
	60%	.2879	.6894	.7852	4.41	24.87	25.40
Outliers	None	.3734	.9291	.9775	13.93	84.13	86.77
	20%	.3544	.6704	.7870	0.00	0.00	0.00
	40%	.3543	.5527	.6546	0.00	0.00	0.00
Masking variables	0	.9089	.9066	.9021	30.45	29.63	29.22
	2-low	.4633	.8701	.8993	2.06	29.22	29.22
	2-high	.3156	.7530	.8657	0.00	29.22	29.22
	4-low	.3761	.8595	.8993	0.00	28.81	29.22
	4-high	.0890	.4861	.6883	0.00	28.40	28.81
	6-low	.3402	.8508	.8974	0.00	28.81	29.92
	6-high	.0079	.2882	.5029	0.00	22.22	27.57
Overall		.3573	.7163	.8078	4.64	28.04	28.92

*Adjusted Rand Index column values represent the mean adjusted Rand index at each factor level, for each factor. The "Percentage of Perfect Recoveries" column values reflect, for each factor level, the percentage of test problems for which perfect recovery was achieved.

by the fact that all three methods perform roughly the same when there are no masking variables, but exhibit some significant disparities at other levels of the masking variables.

For each of the three solution procedures and each factor level, Table 4 reports the average adjusted Rand index and percentage of test problems for which perfect cluster recovery (adjusted Rand index = 1.0) was achieved. These results clearly demonstrate the effectiveness of the variable-selection heuristic. Across all 1701 datasets, the average adjusted Rand indices for VS-KM1 and VS-KM2 were more than double the corresponding average for ALL-KM. Moreover, the percentage of datasets for which perfect cluster recovery was achieved was only 4.64% for the ALL-KM procedure, but over 25% for both of the VS-KM procedures.

VS-KM1 and VS-KM2 improved the average true cluster structure recovery regardless of the number of clusters. The average adjusted Rand indices for VS-KM2 were better than those of VS-KM1 at all levels of C, and more than twice those of ALL-KM at all levels of C. The VS-KM2 procedure also outperformed VS-KM1 and ALL-KM across all levels of the number of true variables and all levels of cluster density. Not surprisingly, outliers had a much more pronounced effect on true cluster recovery than the number of clusters, the number of true variables, or cluster density. When no outliers were present in the dataset, the VS-KM procedures performed exceptionally well. True cluster recovery declined markedly for the 20% and 40% outlier conditions. However, both VS-KM1 and VS-KM2 still maintained a sizable advantage over ALL-KM.

Perhaps the most important results in Table 4 are those pertaining to the number of masking dimensions. As expected, when there were no masking variables in the dataset, the ALL-KM procedure performed slightly better than VS-KM1 and VS-KM2. However, a pairwise comparison of means using Tukey's procedure revealed no significant difference between the procedures at the .05 significance level. Although both of the VS-KM procedures resulted in a slight decrease in the average adjusted Rand index, the differences are quite small considering that the heuristic can only fail for these test problems (by eliminating true variables). These results are also consistent with Milligan's (1989) findings concerning De Soete's (1986) variable-weighting method within the context of hierarchical clustering. Milligan (1989) observed that such minor reductions in these recovery values "... would have little, if any, practical impact in an applied clustering" (p. 59).

For datasets with two or more masking variables, the importance of using the variable-selection heuristic prior to the K-means analysis was unequivocal. For the ALL-KM procedure, the average adjusted Rand indices for $D_2 = 2$, 4, and 6 at *low* levels of masking variable correlation were .4633, .3761, and .3402, respectively. Comparable figures for the VS-KM2 procedure were .8993, .8993, and .8974, respectively. At *high* levels of masking variable correlation, the average adjusted Rand indices for ALL-KM were a dismal .3156, .0890, and .0079 for $D_2 = 2$, $D_2 = 4$, $D_2 = 6$, respectively. Comparable figures for VS-KM2 were .8657, .6883, and .5029, respectively. The VS-KM2 procedure outperformed VS-KM1 at all masking variable settings other than $D_2 = 0$. Most notably, VS-KM2 was substantially better than VS-KM1 when there was high correlation among the masking variables.

In terms of CPU time, the ALL-KM procedure required an average of 4.33 seconds. The average CPU time for the VS-KM1 and VS-KM2 procedures were 93.15 seconds (89.82 seconds for variable selection +3.33 seconds for the K-means algorithm) and 38.22 seconds (35.29 seconds for variable selection +2.92 seconds for the K-means algorithm), respectively. The VS-KM2 procedure is far more efficient that VS-KM1 because a value of $T > 0$ enables a fathoming step in the computer code. Specifically, once a pair of variables (j, k) is identified that achieves a value of $r_{jk} > T$, the development of a two-variable partition (in Step 3) for all subsequent pairs of variables is only made if the adjusted Rand index value between the two single-variable partitions (computed in Step 2) equals or exceeds the threshold level.

5. Study II

5.1. Data Generation for Study II

The vast majority of Monte Carlo analyses reported in the clustering literature have used anywhere from 50 to 300 objects (Balakrishnan et al., 1994; Helsen & Green, 1991; Milligan 1980; Milligan 1981; Milligan & Cooper 1986; Milligan, 1989; Waller et al., 1998). In most previous applications this was reasonable given the settings in which cluster analysis might be applied.

At the same time, many practical clustering applications may contain thousands of objects (see for example, the recent discussions of market segmentation applications by Balasubramanian, Gupta, Kamakura, & Wedel, 1998; Chaturvedi, Carroll, Green, & Rotondo, 1997; Wedel & Kamakura 1997). Moreover, with the advent of very large databases in marketing research and e-commerce settings (Berry & Linoff 1997; Blattberg, et al. 1994), a host of datamining applications have been applied to larger and larger datasets. What are the implications of the cluster techniques when applied to large datasets that will most certainly require efficient use of computational resources? Can the VS-KM procedure be successfully applied to large datasets within a reasonable amount of CPU time?

To answer these questions, we conducted a second Monte Carlo study to determine if the new variable-selection heuristic was still effective for larger datasets. Specifically, 567 additional datasets were generated in precisely the same manner as described in section 3, but with a much

larger number of objects. These datasets consisted of a minimum of $M = 5000$ objects (no outliers) and a maximum of $M = 7000$ objects (40% outliers). The ALL-KM, VS-KM1, and VS-KM2 procedures were applied to each of these 567 large datasets.

5.2. *Results of Study II*

The results for the $5000 \leq M \leq 7000$ test problems of Study II are summarized in Table 5. For each factor level and all three solution procedures, Table 5 reports the average adjusted Rand index and percentage of test problems for which perfect cluster recovery was achieved. It is evident that the results in Table 5 are very consistent with those in Table 4. The VS-KM2 procedure maintained its considerable advantage over ALL-KM and VS-KM1 across most factor level settings. The average adjusted Rand index for the VS-KM2 procedure was only slightly less for $5000 \leq M \leq 7000$ (.7934) than it was for $500 \leq M \leq 700$ (.8078), as was the percentage of datasets for which perfect recovery was achieved (25.93% for $5000 \leq M \leq 7000$ and 28.92% for $500 \leq M \leq 700$). These results suggest that basing the variable-selection heuristic on a sample size of 100 did not yield a substantial penalty in terms of solution quality despite the tenfold increase in the number of objects.

For the $5000 \leq M \leq 7000$ datasets, most of the results pertaining to the number of clusters, number of true variables, density of the clusters, and number of masking dimensions were comparable to those described for Study I. VS-KM1 and VS-KM2 were superior to ALL-KM at

TABLE 5.
Results of Study II: True cluster recovery results for $5000 \leq M \leq 7000$ datasets*

		Adjusted Rand Index (averages)			Percentage of perfect recoveries		
Factor	Level	ALL-KM	VS-KM1	VS-KM2	ALL-KM	VS-KM1	VS-KM2
Number of Clusters	3	.4115	.6675	.7305	3.70	21.69	22.22
	4	.3706	.6856	.7562	3.17	20.63	22.22
	5	.3861	.7986	.8937	5.29	31.75	33.33
Number of true variables	4	.3577	.6484	.7285	3.17	20.11	22.22
	6	.4136	.7308	.8156	3.17	21.69	22.22
	8	.3969	.7726	.8362	5.82	32.28	33.33
Density	Even	.4203	.7628	.8153	4.76	25.40	25.93
	10%	.4528	.7358	.8178	3.70	24.87	25.93
	60%	.2951	.6531	.7472	3.70	23.81	25.93
Outliers	None	.4115	.9369	.9806	12.17	74.07	77.78
	20%	.3849	.6987	.7514	0.00	0.00	0.00
	40%	.3719	.5251	.6483	0.00	0.00	0.00
Masking variables	0	.9043	.8976	.8979	25.93	25.93	25.93
	2-Low	.5513	.8757	.8945	2.47	25.93	25.93
	2-High	.3391	.7412	.8212	0.00	25.93	25.93
	4-Low	.4472	.8558	.8894	0.00	25.93	25.93
	4-High	.1008	.5356	.6795	0.00	24.69	25.93
	6-Low	.3786	.8486	.8823	0.00	25.93	25.93
	6-High	.0047	.2663	.4893	0.00	18.52	25.93
Overall		.3894	.7172	.7934	4.06	24.69	25.93

*Adjusted Rand Index column values represent the mean adjusted Rand index at each factor level, for each factor. The "Percentage of Perfect Recoveries" column values reflect, for each factor level, the percentage of test problems for which perfect recovery was achieved.

every factor level setting, except for the case where there are no masking variables present in the data. Further, the VS-KM2 procedure's superiority to the ALL-KM procedure increased as the number of masking dimensions was increased, and VS-KM2 was far superior to VS-KM1 when the correlation among the masking variables was high.

With respect to CPU time, the average for the ALL-KM procedure was 39.11 seconds, whereas the averages for the VS-KM1 and VS-KM2 procedures were 112.35 and 52.34 seconds, respectively. The slight increase in the average CPU times for the VS-KM1 and VS-KM2 procedures is due to the fact that even though there were ten times as many objects in the dataset, the variable-selection procedures were still only using a sample of 100 objects. The significant increase in processing time for the ALL-KM procedure was due to the fact that, for certain datasets with a large number of masking variables, the K-means algorithm required several hundred seconds to converge. When VS-KM1 and VS-KM2 were applied prior to the K-means algorithm, masking variables were eliminated and thus the CPU times were appreciably lower. For example, the average CPU time for the K-means component of the VS-KM2 procedure was only 16.78 seconds, while the variable-selection heuristic required 35.56 seconds.

The VS-KM2 procedure strongly outperformed VS-KM1 in terms of both true cluster recovery and CPU time in both experimental studies. For this reason, the remainder of the paper will focus solely on VS-KM2. Hereafter, we refer to VS-KM2 as simply VS-KM.

6. Study III

In a third study, we provide a comparison of VS-KM to a formalized version of HINoV, hereafter referred to HINoV-F. Our implementation of HINoV-F was precisely the same as that of VS-KM as described in sections 2 and 3.3, except for the criterion used to select the variables. Because it was not practical to analyze plots for a large number of datasets, we formalized HINoV by using the differences between the rank-ordered $TOPRI_j$ values. The largest difference between the ordered values was used as a surrogate for the elbow of the scree plot. This strategy tends to be conservative regarding the number of variables selected. However, based on the results in Section 3.2, we believed this conservative approach worked to the advantage of our HINoV-F implementation because it was less likely to incorporate a masking variable. We acknowledge that other formalized versions might perform differently for some datasets.

6.1. A Comparison of HINoV-F and VS-KM for the Datasets in Study I

The HINoV-F procedure was applied to the 1701 datasets described in section 3. The average total CPU time for HINoV-F (16.71 seconds) was considerably less than that of VS-KM (38.22 seconds). HINoV-F yielded an average adjusted Rand index of .7229 and perfect recoveries for 21.4% of the datasets, which compares reasonably well to corresponding values of .8078 and 28.92% for VS-KM. The average adjusted Rand index for VS-KM was larger than that of HINoV-F across all factor level settings of the number of clusters, cluster density, the number of true variables, and outliers. HINoV-F yielded slightly better average adjusted Rand indices for the 2-low and 2-high masking variable levels, whereas VS-KM was slightly better for the 0, 4-low, and 6-low levels. The average adjusted Rand indices for VS-KM were much larger than the corresponding averages for HINoV-F at the 4-high (VS-KM = .6883, HINoV-F = .5039) and 6-high (VS-KM = .5029, HINoV-F = .1211) masking variable levels.

It is also pertinent to note that, although it outperformed HINoV-F at all three levels of outliers, VS-KM was far superior to HINoV-F across datasets with no outliers. For the 567 datasets with no outliers, the average adjusted Rand indices for VS-KM and HINoV-F were .9774 and .8337, respectively. Relative to HINoV-F, VS-KM provided a better (worse) adjusted Rand index for 183 (20) datasets, and the same adjusted Rand index for 364 datasets. For the subset of outlier-free datasets associated with high masking variable correlation, the average adjusted Rand-indices for VS-KM and HINoV-F were .9610 and .6290, respectively. These results sup-

port our supposition that VS-KM outperforms HINoV when correlated masking variables are present.

The relative performance of HINoV-F improved when moving to the 20% outlier level, and further improved at the 40% outlier level. For the 567 datasets at the 40% outlier level, the average adjusted Rand indices for VS-KM and HINoV-F were .6546 and .6334, respectively. Relative to HINoV-F, VS-KM provided a better (worse) adjusted Rand index for 188 (159) datasets, and the same adjusted Rand index for 220 datasets. Thus, neither method performed especially well in the presence of a large number of outliers. However, it is difficult to imagine many realistic data sets where the 40% (or even 20%) outlier conditions would prevail.

6.2. A Comparison of HINoV-F and VS-KM for Datasets with Two True Cluster Structures

We also compared HINoV-F and VS-KM across 27 synthetic error-free datasets ($M = 500$) that contained two true cluster structures in the data. The 27 datasets corresponded to all combinations of the three levels of the number of clusters, the three levels of cluster density, and the three levels of the number of true variables. For each of these combinations, one synthetic dataset was generated using Milligan's (1985) procedure, resulting in the first true cluster structure. A second synthetic dataset was constructed in exactly the same manner using a different random number seed. This dataset provided the second true cluster structure. The two datasets were subsequently merged into a single dataset after reordering the rows (objects). For example, consider the case of a dataset with the following parameters: $C = 4$, even cluster density, and $D_1 = 8$. The generated dataset consisted of $M = 500$ objects measured on 16 (2×8) variables. One subset of eight variables defined the first true cluster structure of four groups in eight variables, whereas the second subset defined the second true cluster structure of four groups in eight variables. An effective variable-selection heuristic should select variables that reveal one of the true structures (although perhaps not always the "better" of the two). HINoV-F and VS-KM were applied to each of the 27 datasets. We also examined the 27 scree plots associated with HINoV and they were generally consistent with the HINoV-F results. The results are reported in Table 6.

Table 6 reveals that VS-KM substantially outperformed HINoV-F for the datasets with two true structures. VS-KM perfectly recovered one of the two true structures for 26 of the 27 datasets, whereas HINoV-F perfectly recovered one of the two structures for only 10 of the 27 datasets. For the one dataset where VS-KM did not provide perfect recovery, the adjusted Rand index for one of the two true structures was a "nearly perfect" .9910. Because VS-KM clearly reveals one of the two structures, it can be argued that the unrevealed structure is also more identifiable as a function of the remaining unselected variables. The results for HINoV-F indicated that, in 14 of the 27 instances, it provided adjusted Rand indices of less than .7 for both true structures. In other words, HINoV-F frequently selected variables from both of the two structures and thus was unable to clearly reveal either of the single structures. These findings generally support our supposition that VS-KM is superior to HINoV when there are multiple true structures in the data.

7. Study IV

As a final demonstration of the usefulness of the new heuristic, we applied VS-KM to a set of data taken from a business-to-business segmentation study of the credit union market. The purpose of the study was to identify relevant customer segments, within the credit union market, which might show differential demand for a member-database software product. Data were collected from financial and statistical reports collected by the National Credit Union Administration (www.ncua.gov). The final dataset included 11402 reporting credit unions providing 15 variables representing measures of institutional size (e.g., number of full-time employees, number of regular shares, total assets), performance (e.g., total operating income, total deposits, loans outstanding), and portfolio value (e.g., interest income, profit/loss).

TABLE 6.
A comparison of HINoV-F and VS-KM for datasets with two true cluster structures*

Number of clusters	Cluster density	Number of true vars	HINoV-F		VS-KM	
			True-1	True-2	True-1	True-2
3	Even	2 sets of 4	.2775	.5113	1.0000	.0032
3	Even	2 sets of 6	.0033	1.0000	.0033	1.0000
3	Even	2 sets of 8	.0032	1.0000	1.0000	.0032
3	10%	2 sets of 4	.2665	.6687	.9910	.0051
3	10%	2 sets of 6	.0005	1.0000	.0005	1.0000
3	10%	2 sets of 8	.0011	1.0000	1.0000	.0011
3	60%	2 sets of 4	.0063	.9968	1.0000	.0065
3	60%	2 sets of 6	.0091	1.0000	.0091	1.0000
3	60%	2 sets of 8	.5823	.1803	1.0000	.0005
4	Even	2 sets of 4	−.0007	1.0000	−.0007	1.0000
4	Even	2 sets of 6	.6678	.1625	1.0000	−.0010
4	Even	2 sets of 8	.0005	1.0000	1.0000	.0005
4	10%	2 sets of 4	.1084	.7963	.0054	1.0000
4	10%	2 sets of 6	.6060	.2415	1.0000	.0041
4	10%	2 sets of 8	.0037	1.0000	1.0000	.0037
4	60%	2 sets of 4	.0205	1.0000	.0205	1.0000
4	60%	2 sets of 6	.6113	.2701	1.0000	.0081
4	60%	2 sets of 8	−.0087	1.0000	−.0087	1.0000
5	Even	2 sets of 4	.2300	.3786	−.0022	1.0000
5	Even	2 sets of 6	.3516	.2558	1.0000	−.0006
5	Even	2 sets of 8	.3025	.2469	1.0000	−.0010
5	10%	2 sets of 4	.3733	.2779	.0037	1.0000
5	10%	2 sets of 6	.0003	.9542	−.0005	1.0000
5	10%	2 sets of 8	.3909	.3697	1.0000	−.0024
5	60%	2 sets of 4	.3516	.5264	.0155	1.0000
5	60%	2 sets of 6	.3568	.3568	1.0000	−.0062
5	60%	2 sets of 8	.4170	.3856	1.0000	−.0052

*The HINoV-F and VS-KM columns contain the adjusted Rand indices for each of the 2 two true structures in the data set. Values of 1.0 in either the "True-1" or "True-2" columns indicate that perfect recovery of one of the true structures was realized.

The means and variance measurements for this segmentation study varied by more than five orders of magnitude, which would cause certain variables to have an unduly large influence on the K-means partitioning results. Standardization of variables is a thorny issue because variable transformations can mask structure that is present in the original variables. However, Milligan and Cooper (1988) and Milligan (1996) identified several variable standardization alternatives that outperformed traditional standardization methods based on standard deviations. The financial institution data were standardized using a transformation procedure proven effective in these studies (see Milligan, 1996; Milligan & Cooper, 1988). The cluster analysis was subsequently conducted on this standardized dataset, $\mathbf{Z}$, with the elements of the dataset, z_{ij} defined as follows:

$$z_{ij} = \frac{x_{ij}}{\underset{i}{\text{Max}}(x_{ij}) - \underset{i}{\text{Min}}(x_{ij})} \qquad \forall i = 1, \ldots, M \text{ and } j = 1, \ldots, D. \tag{1}$$

The number of true clusters and corresponding cluster memberships associated with the financial-services data were unknown. Therefore, evaluation of the variable-selection heuristic was conducted using replication analysis (Breckenridge, 1989; McIntyre & Blashfield, 1980; Morey, Blashfield, & Skinner, 1983), the steps of which have been outlined by Milligan (1996).

The dataset was divided into two random samples each consisting of 5701 objects. The ALL-KM clustering procedure described in section 5 was used to cluster the first sample. Centroids associated with the resulting solution were subsequently used to assign objects in the second sample to clusters. In other words, a partition of the second sample was obtained by assigning each object in that sample to the closest centroid corresponding to the clustering of the first sample. Next, a second partition of the second sample was directly obtained using the ALL-KM clustering procedure. The adjusted Rand index was subsequently computed to identify the level of agreement between these two partitions of the second sample. Milligan (1996) observed that this measure reflects the degree of stability associated with the data clusters. The steps of the replication analysis were subsequently repeated, except that HINoV-F and VS-KM were each applied to the full data set in order to select a subset of variables prior to K-means clustering. Thus, a replication analysis was conducted for clusters developed without the use of the variable-selection heuristic, clusters developed using only variables selected by HINoV-F, and clusters developed using only variables selected by VS-KM. Because the results of these analyses might be sensitive to the random partitioning of the dataset, we conducted three repetitions of the replication analysis. For each repetition, we used a different random partition of the 11402 objects into two samples of 5701.

Because the number of clusters was unknown, the replication analysis was conducted using $C = 2$, 3, 4, 5, and 6 clusters. The results corresponding to each of these cluster sizes are presented in Table 7, which reports the average adjusted Rand indices corresponding to the ALL-KM, HINoV-F, and VS-KM procedures for each cluster size. For each of the procedures, average adjusted Rand indices increased over the range of $2 \leq C \leq 5$, and decreased slightly when moving from 5 to 6 clusters. This finding provided evidence that five is a reasonable number of clusters for this dataset. However, regardless of the number of clusters selected, HINoV-F and VS-KM always resulted in larger average adjusted Rand indices than ALL-KM. HINoV-F and VS-KM provided the same average adjusted Rand indices for $2 \leq C \leq 4$, however, VS-KM yielded better averages for $C = 5$ and $C = 6$. At $C = 5$, the average adjusted Rand index was .8657 for VS-KM, but only .7489 for HINoV-F. We also examined the scree plot of the ranked $TOPRI_j$ values for the five-cluster solution. Two potential elbows were evident. The first elbow suggested elimination of all variables except 1, 2, 3, 4, and 15, whereas the second elbow excluded only variables 11, 12 and 13. The average adjusted Rand indices associated with these two potential subsets were .8382 and .8582, respectively. These values are somewhat better than the HINoV-F result and nearly as good as the VS-KM result.

The VS-KM procedure eliminated two variables related to "interest income" (variables 11 and 13), which enabled greater stability in the final cluster solution. When the K-means algorithm ($C = 5$) was applied to the entire dataset upon removal of these two variables, an interpretable

TABLE 7.
Results of Study III: Replication analysis for the financial services data*

	ALL-KM	HINoV-F		VS-KM	
Number of clusters	Adjusted Rand Index	Adjusted Rand Index	Eliminated variables	Adjusted Rand Index	Eliminated variables
2	.1705	.2636	11 and 12	.2636	11 and 12
3	.5447	.6294	11 and 13	.6294	11 and 13
4	.5689	.7425	11 and 13	.7425	11 and 13
5	.6889	.7489	11	.8657	11 and 13
6	.6205	.7136	11, 12, and 13	.8290	11 and 13

*The replication analysis was conducted for three independent partitions of the 11402 objects into 2 samples of 5701 objects each. The table values contain the average (across the three replication analyses) adjusted Rand indices for both the ALL-KM and VS-KM procedures.

cluster structure was realized. This structure produced two relatively small clusters of large firms: one showing high performance and strong portfolio value, the other displaying moderate performance. The three remaining clusters were composed of small to moderate sized firms primarily differentiated based on performance measures.

8. Discussion

8.1. Summary of Major Findings

This paper has presented a heuristic algorithm for selecting subsets of variables for inclusion in a K-means cluster analysis. This heuristic, which makes use of a measure of cluster homogeneity and the popular adjusted Rand index for measuring cluster recovery, was tested across a wide range of problems in order to discern its ability to eliminate masking variables. The major findings of the study can be summarized as follows:

1. When no masking variables were present in the datasets, the use of the variable-selection heuristic resulted in little deterioration of cluster recovery. For these datasets, the performances of ALL-KM, HINoV-F, and VS-KM were equally impressive.
2. When no outliers were present in the datasets, VS-KM was effective at eliminating masking variables regardless of the number of masking variables and the correlation among them, whereas HINoV-F was only effective when the masking variable correlation was low. Our results suggest that, both theoretically and empirically, VS-KM will generally outperform HINoV-F when correlation among masking variables is high.
3. For datasets with two distinct true cluster structures, VS-KM was extremely effective at identifying one of the two structures, whereas HINoV-F often failed to identify a true structure. Our results suggest that VS-KM will generally outperform HINoV-F when multiple true structures are present in a dataset.
4. HINoV-F is computationally more efficient than VS-KM. However, like HINoV-F, VS-KM is not difficult to program and is reasonably efficient, averaging less than one minute of microcomputer CPU time even for the $5000 \leq M \leq 7000$-object datasets.
5. Another important finding of this study, often not addressed in previous Monte Carlo analyses, is the strong performance of VS-KM for large datasets. Because datasets with thousands of objects are often encountered in practice, it was encouraging to observe only a small decrease in heuristic performance for the $5000 \leq M \leq 7000$-object datasets.

8.2. A Brief Comparison of HINoV and VS-KM

Carmone et al.'s (1999) HINoV procedure has several advantages over VS-KM that are noteworthy. First, it is conceptually more straightforward because it requires only sums of adjusted Rand indices and adds variables in a single pass/observation. A second, and related, advantage is that it is computationally more efficient than VS-KM. A third advantage is that the minimum number of variables selected by HINoV is one, whereas VS-KM begins with the selection of a pair of variables. This is not a major advantage, however, because it seems reasonable to expect that two or more variables would generally be selected for clustering.

We have provided theoretical arguments that VS-KM should outperform HINoV under two conditions: (a) correlated masking variables, and (b) multiple true cluster structures in a dataset. Our empirical studies in section 6 support these arguments. Because it is reasonable to expect that correlated masking variables and multiple true cluster structures are likely to be present in many datasets, particularly market segmentation databases that might consist of hundreds of candidate variables, we feel that VS-KM is a worthy alternative to HINoV and other variable-weighting and selection methods. We recognize that no method is apt to always provide the best results.

8.3. Potential Applications

Market segmentation. It has often been noted that nonhierarchical cluster analysis techniques, particularly variants of MacQueen's K-means method, have received extensive use in market segmentation applications (Arabie & Hubert, 1994; Chaturvedi et al., 1997; DeSarbo et al., 1984; DeSarbo, Manrai, & Manrai, 1993; Helsen & Green, 1991, Wedel & Kamakura, 1997). As noted by DeSarbo et al. (1984), the number of potential variables for market segmentation cluster analyses may exceed 200 variables and thus variable weighting and selection decisions are extremely important. Any technique that can be used to eliminate variables would be of considerable value. The results presented in this paper demonstrate the effectiveness of the new variable-selection heuristic for eliminating unnecessary variables from datasets with up to the $5000 \leq M \leq 7000$ objects and 14 variables (8 true + 6 masking). Therefore, we believe that the variable-selection method presented in this paper might be a promising new tool for selecting variables in large-scale market segmentation studies.

Behavioral science applications. DeSarbo et al. (1984) described several areas of applications for nonhierarchical clustering methods in the behavioral sciences. For subject classification in clinical psychology applications, they note that variable weighting and selection methods could be appropriate for choosing test items that would be useful for such classification. DeSarbo et al. (1984) also observed that classification of students in educational psychology could be improved by using variable-selection procedures. For example our procedure could be used to select from an assortment of variables pertaining to classroom characteristics, administrative structure, faculty characteristics, student and parent demographic information, and other factors. Arabie and Hubert (1996) recently observed that cluster analysis applications in social psychology and sociometry have also become more frequent. In all cases, the VS-KM procedure should result in improved data analysis.

8.4. Limitations and Extensions

Limitations regarding the experimental testing. We selected Milligan's (1985) procedure for generating test problems because it is one of the best-documented and most frequently deployed generation procedures in the literature (Balakrishnan et al., 1994; Helsen and Green, 1991; Milligan, 1980, Milligan & Cooper, 1986, Milligan 1989, Waller et al., 1998). Further, in some respects, we have pushed the boundaries of this generation process by considering up to 6 noise variables and increasing the level of correlation among such variables. We also have expanded testing to include the case of multiple true structures in a single dataset, which might provide a very fruitful avenue for future research. Nevertheless, we recognize that other types of data structures might yield different findings and could necessitate new parameters or methods for VS-KM. For this reason, comparisons of VS-KM with other variable weighting and selection procedures across a variety of other data structures would be a worthy avenue for subsequent research.

Limitations and extensions of the variable-selection heuristic. VS-KM can be implemented in an informal manner (using tables or graphs) like HINoV or other variable-selection methods (Fowlkes et al., 1988). However, we presented a formal, parameterized version of VS-KM in order to facilitate a large computational study. One of the most serious criticisms of VS-KM is the need to identify parameter values for T, G_{min}, and G_{fac}. The parameter settings of $T = .25$, $G_{min,} = .05$, and $G_{fac} = .5$ provided good results for most of the more than 2200 synthetic data sets we considered, as well as the real-world data set, and can be considered as a guideline for subsequent studies.

We observed that the recovery of the true cluster structure deteriorated as the number of outliers increased. Two points are noteworthy in this regard. First, the VS-KM procedure performed

well when the outlier level was 40% (or 20%) and masking variable correlation was low, and when masking variable correlation was high and there were no outliers. The difficulty was encountered when both masking variable correlation was high and the outlier level was 40% (20%). The 20% and 40% outlier levels denigrate the "true" cluster structure, while at the same time a strong correlation among the "masking" variables creates a misleading structure among those variables. It seems that these extreme conditions blur the distinction between "true" and "masking" variables and thus are of little pragmatic relevance. A second point is that recent progress has been made in the detection of outliers (Cheng & Milligan, 1996) and it might be useful to examine outlier detection methods in conjunction with variable-selection procedures.

Although the above noted shortcomings of VS-KM were infrequent in our computational study, it might be interesting to examine variations of the heuristic. The crux of VS-KM is that multiple criteria should be considered in the variable-selection process, particularly when including the first couple of variables. If a bad choice of variables is made at the beginning of a forward selection process, then the inclusion of additional variables based on the adjusted Rand index might just compound the problem. In order to mitigate the chance of a poor initial selection, we consider both the adjusted Rand index and cluster homogeneity in the selection process. It might be possible to incorporate other types of homogeneity or separation information when making this decision. Indeed, a separation criterion might be very effective for the data in Figure 3. Another possible improvement would be to construct a local search method that adds or removes variables from **S** either randomly or based on one or more criteria. Although such a process overcomes some inherent problems with forward selection, it could be rather computationally intensive.

References

Anderberg, M.R. (1973). *Cluster analysis for applications*. New York, NY: Academic Press.

Arabie, P., & Hubert, L.J. (1994). Cluster analysis in marketing research. In R.P. Bagozzi (Ed), *Advanced methods in marketing research* (pp. 160–189). Oxford, England: Blackwell.

Arabie, P., & Hubert, L.J. (1996). An overview of combinatorial data analysis. In P. Arabie, L.J. Hubert, & G. De Soete (Eds), *Clustering and classification* (pp. 5–63). River Edge, NJ: World Scientific Publishing.

Art, D., Gnanadesikan, R., & Kettenring, J.R. (1982). Data-based metrics for cluster analysis. *Utilitas Mathematica, Series A, 21*, 75–99.

Balakrishnan, P.V., Cooper, M.C., Jacob, V.S., & Lewis, P.A. (1994). A study of the classification capabilities of neural networks using unsupervised learning: A comparison with K-means clustering. *Psychometrika, 59*, 509–525.

Balasubramanian, S., Gupta, S., Kamakura, W., & Wedel, M. (1998). Modelling large data sets in marketing. *Statistica Neerlandica, 52*, 303–323.

Berry, M.J.A., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. New York, NY: John Wiley & Sons.

Blattberg, R., Glazer, R., & Little, J. (1994). *The marketing information revolution*. Boston, MA: Harvard Business School Press.

Box, G.E.P., & Muller, M.E. (1958). A note on the generation of random normal deviates. *Annals of Mathematical Statistics, 29*, 610–611.

Breckenridge, J.N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research, 24*, 147–161.

Carmone, F.J., Kara, A., & Maxwell, S. (1999). HINoV: A new model to improve market segmentation by identifying noisy variables. *Journal of Marketing Research, 36*, 501–509.

Chaturvedi, A., Carroll, J.D., Green, P.E., & Rotondo, J.A. (1997). A feature-based approach to market segmentation via overlapping K-centroids clustering. *Journal of Marketing Research, 34*, 370–377.

Cheng, R., & Milligan, G.W. (1996). K-means clustering methods with influence detection. *Educational and Psychological Measurement, 56*, 833–838.

Cormack, R.M. (1971). A review of classification (with Discussion). *Journal of the Royal Statistical Society, Series A, 134*, 321–367.

DeSarbo, W.S., Carroll, J.D., Clark, L.A., & Green, P.E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with different weighting of variables. *Psychometrika, 49*, 57–78.

DeSarbo, W.S., Manrai, A.K., & Manrai, L.A. (1993). Non-spatial tree models for the assessment of competitive market structure: An integrated review of the marketing and psychometric literature. In J. Eliashberg & G. Lilien (Eds), *Handbook in operations research and management science: Marketing* (pp. 193–257), New York, NY: Elsevier.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. *Quality and Quantity, 20*, 169–180.

De Soete, G. (1988). OVWTRE: A program for optimal variable weighting for ultrametric and additive tree fitting. *Journal of Classification, 5*, 101–104.

De Soete, G., DeSarbo, W.S., & 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., & Kettenring, J.R. (1987). Variable selection in clustering and other contexts. In C.L. Mallows (Ed.), *Design, data, and analysis* (pp. 13–34). New York, NY: John Wiley & Sons.

Fowlkes, E.B., Gnanadesikan, R., & Kettenring, J.R. (1988). Variable selection in clustering. *Journal of Classification, 5*, 205–228.

Fowlkes, E.B., & Mallows, C.L. (1983). A method for comparing two hierarchical clusterings (with comments and rejoinder). *Journal of the American Statistical Association, 78*, 553–584.

Friedman, H.P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association, 62*, 1159–1178.

Gnanadesikan, R., Kettenring, J.R., & Tsao, S.L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification, 12*, 113–136.

Green, P.E., Carmone, F.J., & Kim, J. (1990). A preliminary study of optimal variable weighting in K-means clustering. *Journal of Classification, 7*, 271–285.

Helsen, K., & Green, P.E. (1991). A computational study of replicated clustering with an application to market segmentation. *Decision Sciences, 22*, 1124–1141.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2*, 193–218.

Knuth, D.E. (1997). *The art of computing: Vol. 1. Fundamental algorithms*. Reading, MA: Addison-Wesley.

Krieger, A., & Green, P.E. (1999). A generalized rand-index method for consensus clustering of separate partitions of the same data base. *Journal of Classification, 16*, 63–89.

Kruskal, J.B. (1972). Linear transformations of multivariate data to reveal clustering. In R.N. Shepard, A.K. Romney, & S.B. Nerlove (Eds), *Multidimensional scaling: Theory and applications in the behavioral sciences* (pp. 181–191). New York, NY: Seminar Press.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, I*, 231–297.

McIntyre, R.M., & Blashfield, R.K. (1980). A nearest-centroid technique for evaluating the minimum variance clustering procedure. *Multivariate Behavioral Research, 15*, 225–238.

Milligan, G.W. (1980). An examination of six types of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45*, 325–342.

Milligan, G.W. (1985). An algorithm for generating artificial test clusters. *Psychometrika, 50*, 123–127.

Milligan, G.W. (1989). A validation study of a variable-weighting algorithm for cluster analysis. *Journal of Classification, 6*, 53–71.

Milligan, G.W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L.J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 341–375). River Edge, NJ, World Scientific Publishing.

Milligan, G.W., & Cooper, M.C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21*, 441–458.

Milligan, G.W., & Cooper, M.C. (1988). A study of the standardization of variables in cluster analysis. *Journal of Classification, 5*, 181–204.

Milligan, G.W., Soon, S.C., & Sokol, L.M. (1983). The effect of cluster size, dimensionality, and the number of clusters on the recovery of true cluster structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 5*, 40–47.

Morey, L.C., Blashfield, R.K., & Skinner, H.A. (1983). A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research, 18*, 309–329.

Rand, W.M. (1971). Objective criteria for evaluating clustering methods. *Journal of the American Statistical Association, 66*, 846–850.

Rohlf, F.J. (1970). Adaptive hierarchical clustering schemes. *Systematic Zoology, 19*, 58–82.

Salstone, R., & Stange, K. (1996). A computer program to calculate Hubert and Arabie's adjusted Rand index. *Journal of Classification, 13*, 169–172.

Ward, J.H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244.

Waller, N.G., Kaiser, H.A., Illian, J.B., & Manry, M. (1998). A comparison of the classification capabilities of the 1-dimensional Kohonen neural network with two partitioning and three hierarchical cluster analysis algorithms. *Psychometrika, 63*, 5–22.

Wedel, M., & Kamakura, W.A. (1997). *Market segmentation: Conceptual and methodological foundations*. Boston, MA: Kluwer Academic Publishers.

Manuscript received 18 MAY 1999
Final version received 18 APR 2000