Journal of Classification 5:181-204 (1988)

A Study of Standardization of Variables in Cluster Analysis

Glenn W. Milligan

The Ohio State University

Martha C. Cooper

The Ohio State University

Abstract: A methodological problem in applied clustering involves the decision of whether or not to standardize the input variables prior to the computation of a Euclidean distance dissimilarity measure. Existing results have been mixed with some studies recommending standardization and others suggesting that it may not be desirable. The existence of numerous approaches to standardization complicates the decision process. The present simulation study examined the standardization problem. A variety of data structures were generated which varied the inter-cluster spacing and the scales for the variables. The data sets were examined in four different types of error environments. These involved error free data, error perturbed distances, inclusion of outliers, and the addition of random noise dimensions. Recovery of true cluster structure as found by four clustering methods was measured at the correct partition level and at reduced levels of coverage. Results for eight standardization strategies are presented. It was found that those approaches which standardize by division by the range of the variable gave consistently superior recovery of the underlying cluster structure. The result held over different error conditions, separation distances, clustering methods, and coverage levels. The traditional *z*-score transformation was found to be less effective in several situations.

Keywords: Standard scores; Cluster analysis.

Authors' Addresses: Glenn W. Milligan, Faculty of Management Sciences, 301 Hagerty Hall, The Ohio State University, Columbus, Ohio 43210, USA. Martha C. Cooper, Faculty of Marketing, 421 Hagerty Hall, The Ohio State University, Columbus, Ohio 43210, USA.

1. Introduction

Although the issue of standardization of variables in a cluster analytic study is important to an applied researcher, little is known about the impact of this procedural step. Standardization of variables would seem to be necessary in those cases where the dissimilarity measure, such as Euclidean distance, is sensitive to differences in the magnitudes or scales of the input variables. Sneath and Sokal (1973) indicate that standardization is accomplished with the use of translation and expansion. The purpose is to equalize the size or magnitude and the variability of the input variables. Similarly, Anderberg (1973) states that the purpose is to adjust the magnitude of the scores and the relative weighting of the variables. Romesburg (1984) makes the same arguments, particularly with respect to the idea of equal weighting of variables. The concept of achieving equal weight for each of the input variables has received much support. Cormack (1971) indicates that equal weighting is obtained by using an adjustment factor which is inversely proportional to the variability of the measure.

Several authors have argued that the best strategy is not to standardize across all elements on the variable. Rather, standardization should occur within clusters on each variable. This position has been taken by Cormack (1971), Everitt (1980), Hartigan (1975), Fleiss and Zubin (1969), and Overall and Klett (1972). A number of authors have provided short derivations or graphical illustrations which show that the total variability is not only a function of the within-cluster variances, but also of the distance between cluster means (Bayne, Beauchamp, Begovich, and Kane 1980; Fleiss and Zubin 1969; Hartigan 1975; Lorr 1983; and Späth 1980). The inclusion of the latter distance confounds the standardization process and serves to reduce the apparent separation between clusters.

However, it is impossible to directly standardize within-clusters in an applied analysis. To find the clusters, one must know the assignments of the elements to the clusters beforehand in order to perform the standardization. That is, one must know what the clusters are before finding the clusters. Overall and Klett (1972) proposed iterating the cluster analysis in an attempt to overcome this circularity problem. One first obtains clusters based on overall estimates. Next, these initial clusters are used to help determine within group variances for standardization in a second cluster analysis. The process can continue until no changes in cluster membership occur.

Other logical difficulties face the uncritical use of standardization of variables. First, consider the possibility that the clusters which exist in the data are embedded in the unstandardized variable space. This situation seems at least as likely to occur as the existence of the clusters in the rescaled space. Sawery, Keller, and Conger (1960) were early advocates of the use of the unscaled input data for direct clustering. Second, although the arguments in favor of equal weighting of the input variables may seem appealing, there is no compelling reason to practice democracy while performing all cluster analyses. In some cases, the differential weighting of the variables before standardization may represent information that defines the clusters. In a study not addressing the issue of standardization, the concept of differential weighting of variables was approached by De Soete, DeSarbo, and Carroll (1985). The authors reported substantial improvements in cluster recovery with the use of a differential weighting algorithm for variables. A different strategy based on applying weights to the variables after standardization was introduced by Hohenegger (1986). Support for differential weighting also can be found in the early clustering literature. Both Hall (1965) and Williams, Dale, and MacNaughton-Smith (1964) advocated scaling by a measure of the importance of the variable. Hence, there has not been universal agreement that equal weighting is necessary or optimal.

1.1 Forms of Standardization

Numerous approaches to standardization of variables exist. The present study considers only the case involving numerical variables. Categorical variables or the mixture of categorical and numerical variables are not considered. Researchers from social science backgrounds usually assume that a standardized variable has been transformed to have zero mean and unity variance as found with the typical "z-score" formula. However, other proposals for the standardization or scaling of variables can be found in the classification literature and these are reviewed in this section. For convenience, the term standardization will be used in a generic sense in this article.

The first form of standardization is the *z*-score formula used for transforming normal variates to standard score form:

$$Z_1 = (X - \bar{X}) / s \,, \tag{1}$$

where X is the original data value, and $\bar{X}$ and s are the sample mean and standard deviation, respectively. The transformed variable will have a mean of 0.0 and a variance of 1.00. The transformation has been proposed by numerous authors including Dubes and Jain (1980), Everitt (1980), Lorr (1983), Romesburg (1984), SAS (*SAS User's Guide: Statistics*, 1985), Sokal (1961), Späth (1980), and Williams, Lambert, and Lance (1966). Späth warns that Z_1 may not perform properly if there are substantial differences among the within-cluster standard deviations.

It is important to note that standardization Z_1 must be applied in a global manner (across all items on the variable) and not within individual clusters. To understand this restriction, consider the case where three well-separated clusters exist in the data. Further, assume that a sample point is located at each of the three cluster centroids. Standardization within clusters would lead to a vector of scores for each of the three centroid points which contains zeros for all entries. Most any clustering procedure using the standardized data would place the three centroid points in the same cluster. More generally, the same result occurs if the three points are located at the same standardized coordinate vector relative to each of the three cluster centroids. For example, there will be three different locations in a two variable space that would have coordinate values (1.0, -1.0). Data points found near these three different locations would have near zero interpoint distances and likely would be clustered together. Thus, such results would lead to an erroneous and very misleading clustering solution. Thus, Z_1 must not be used when standardization is computed within-cluster.

The next standardization is similar to Z_1 and is computed as:

$$Z_2 = X / s \ . \tag{2}$$

Formula Z_2 will result in a transformed variable with a variance of 1.00 and a transformed mean equal to $\bar{X} / s$. However, since the scores have not been centered by subtracting the mean, the information as to the comparative location between scores remains. As such, Z_2 will not suffer from the problem of the loss of information about the cluster centroids as is the case for Z_1 as noted earlier. Several authors have proposed the use of Z_2 including Anderberg (1973), Cormack (1971), Fleiss and Zubin (1969), Hartigan (1975), and Overall and Klett (1972).

The reader should note that Z_1 and Z_2 are linear functions of each other. As such, Euclidean distances computed using the two formulas lead to identical dissimilarity values when global means and variances are used.

The third procedure involves standardization with the use of the maximum score on the variable:

$$Z_3 = X / \mathrm{Max}\,(X) \ . \tag{3}$$

If all values are greater than or equal to zero, then the transformed variable acts as a proportional measure with all scores falling between 0.0 and 1.0. (If some X's are negative, a sufficiently large positive constant can be added to all values to obtain the proportionality property.) The transformed mean and standard deviation are X / Max (X) and s / Max (X), respectively. Although the upper limit of 1.00 is obtained in each data set, the lowest observed value may be greater than 0.0. Note that Z_3 will not result in constant variances across transformed variables. In fact, Z_3 leaves the relative variability (the range divided by the maximum value) unchanged before and after transformation (Sneath and Sokal 1973). Further, Z_3 is susceptible to the presence of outliers. A single large observation on a variable can have the effect of compressing the remaining values near 0.0. It would seem that Z_3 is meaningful only when the variable constitutes a ratio scale. Standardization Z_3 has been proposed by Cain and Harrison (1958), Hall (1969), and Romesburg (1984).

The fourth and fifth standardizations involve using the range of the variable as the divisor:

$$Z_4 = X \,/\, (\text{Max}\,(X) - \text{Min}\,(X)) \ , \qquad (4)$$

$$Z_5 = (X - \text{Min}\,(X)) \,/\, (\text{Max}\,(X) - \text{Min}\,(X)) \ . \qquad (5)$$

Assuming nonnegative values, standardization Z_5 is bounded by 0.0 and 1.0 with at least one observed value at each of these end points. Formula Z_4 is not bounded in this manner and will not usually behave as a proportion. The transformed mean will be $\bar{X} \,/\, (\text{Max}\,(X) - \text{Min}\,(X))$ for Z_4 and $(\bar{X} - \text{Min}\,(X)) \,/\, (\text{Max}\,(X) - \text{Min}\,(X))$ for Z_5. Both procedures result in a transformed standard deviation of $s \,/\, (\text{Max}(X) - \text{Min}(X))$. The transformed mean or variance will not be constant across variables for either Z_4 or Z_5. However, an upper limit for the variance exists and is equal to .25. As with Z_3, both Z_4 and Z_5 may be adversely affected by the presence of outliers on the variable. Standardization Z_4 has been mentioned by Anderberg (1973), Carmichael, George, and Julius (1968), Cormack (1971), and Lance and Williams (1967). Formula Z_5 has been proposed by Gower (1971), Romesburg (1984), Sneath and Sokal (1973), and Späth (1980). In particular, Sneath and Sokal prefer this transformation for many applications.

As with the pair Z_1 and Z_2, Z_4 is a linear function of Z_5. Hence, Euclidean distances based on Z_4 will be identical in value to those computed from Z_5 for the same data set.

On occasion, a standardization based on normalizing to the sum of the observations has been suggested:

$$Z_6 = X \,/\, \Sigma X \ . \qquad (6)$$

Formula Z_6 will normalize the sum of the transformed values to 1.00 and the transformed mean will equal $1 \,/\, n$. As such, the mean will be constant across variables, but the variances will differ. Procedure Z_6 was proposed by Romesburg (1984) and is similar to a formula given by Anderberg (1973) which used division by the sample mean.

A different approach to standardization involves converting the values to ranks:

$$Z_7 = \text{Rank}\,(X)\ . \qquad (7)$$

Formula Z_7 results in a transformed variable with mean $(n + 1) / 2$, range $n - 1$, and variance $(n{+}1)\,[((2n{+}1) / 6) - ((n{+}1) / 4)]$ for all variables. The transformation reduces the impact of outliers in the data and is consistent with arguments made in favor of non-parametric procedures as advocated by Conover and Iman (1981). The rank transform in the clustering context was proposed by Sneath and Sokal (1973).

A few other standardization methods exist in the literature. For example, Burr (1968) proposed taking the logarithms of values. This strategy is similar to the frequent use of "re-expressions" in Tukey's (1977) approach to exploratory data analysis. Certainly, the log transform will reduce the impact of outliers in a fashion similar to the ranking procedure Z_7. However, log transforms will not exhibit constant means, ranges, or variances across variables.

Sokal and Rohlf (1969) presented a transformation called "Rankits." A rankit is computed as the average deviate of the r-th largest value in a sample of n observations drawn at random from a standard normal distribution. The transformation could be adapted for use in a clustering context. However, the procedure would seem to be needed only in those cases where the analysis procedure is sensitive to the departure of data values from a normal population. Similarly, Stoddard (1979) presented a specialized method for scaling measurements when a reference standard is available. Reference standards of the sort required by Stoddard's method are usually found only in laboratory experiments conducted in the chemical or the biological sciences.

Morrison (1967) raised the scaling issue, then rejected simple Euclidean distance in favor of generalized or Mahalanobis distance as a simultaneous solution to several problems. However, Gordon (1981) has argued against the use of Mahalanobis distance in a clustering context because of the difficulty of the specification of an appropriate variance-covariance matrix. Furthermore, we object to the idea that, in most cases, the variables selected for a cluster analysis represent a random sample of those available to the researcher. As the results reviewed by Milligan and Cooper (1987) indicate, selection of variables for a cluster analysis must be done with great care.

Arguing from a different perspective, Jardine and Sibson (1971) rejected standardization of variables and Euclidean distance in favor of a measure called D-dissimilarity, derived from information theory.

1.2 Previous Monte Carlo Research

Little experimental information is available concerning the usefulness of the various forms of standardization. Few studies have provided comparative information among the forms or even addressed the issue of whether or not standardization is necessary or desirable. Edelbrock (1979) compared the performance of non-standardized data and Z_1. A slight advantage was found for Z_1 when all elements were required to be clustered. When some elements were allowed to remain unassigned at reduced coverage levels, Z_1 produced no improvement in recovery. The result suggests that standardization may reduce the impact of observations which are intermediates or outliers to the clusters.

On the other hand, Milligan (1980) found that Z_1 can lead to a limited reduction in recovery performance when the clusters were originally generated in the unstandardized variable space. Since the clusters were well-separated in the Milligan study, outliers did not exist in the error-free data.

Kaufman (1985) prepared simulated data for clustering in several ways. One condition used Z_1, while the other four conditions involved various strategies using principal components. Kaufman found that Z_1 produced cluster recovery results which were nearly as effective as the best principal components condition. In fact, Z_1 gave much better recovery than two of the principal components strategies. Kaufman did not include a condition in the study which used the nonstandardized data directly in the clustering process.

Virtually no other comparative information exists. In the Blashfield (1976) study, all data sets were standardized using Z_1. Bayne et al. (1980) seem to have standardized their simulation data using Z_1 or Z_2, but the exact form was not specified. Likewise, Scheibler and Schneider (1985) standardized all data sets while failing to specify the form. It appears that they used Z_5 and they argue that Späth (1980) prefers Z_5 over Z_1. However, Späth's book does not seem to support this statement. In fact, Späth seems to prefer Z_1 and gives two references in support of its use. Furthermore, Späth provides a Fortran subroutine for Z_1 in the text, but does not offer a program listing for Z_5.

2. Experimental Design

The present study examined the effect of standardization forms Z_1 to Z_7 on the recovery of cluster structure in a variety of artificial data configurations. Analyses based on the untransformed data were included to provide a basis for comparison. Let the results for the original data be denoted by Z_0. The different forms of standardization corresponded to one independent variable in a fully-crossed six factor design. Three other factors were used to alter the data configurations, and the two remaining factors were clustering methods and coverage levels. For the reader's convenience, the six factors and the names of the associated levels are given in Table 1.

2.1 Data Generation

The artificial data were generated with the use of the routine developed by Milligan (1985). The data generation routine produces data sets with differing characteristics. The varying features include data sets possessing 2 to 5 clusters embedded in a 4, 6, or 8 dimensional space, and three levels for controlling the relative cluster sizes. All points within a cluster were generated from multivariate normal distributions with uncorrelated variables. Each data point was constrained to fall within cluster boundaries that corresponded to ±1.5 standard deviations about the mean on each variable. Points farther from the centroid were deleted from the cluster. The standard deviation value on each dimension for a cluster was randomly selected from the uniform interval of 10 to 40 units. Thus, some elongation in cluster shapes was present. A total of 50 points was present in each data set. The initial error-free data sets can be characterized as possessing well-separated clusters which exhibit good external isolation and internal cohesion.

In the original version of the data generator (Milligan 1985), the centroids on the first dimension were located randomly subject to two constraints. First, cluster boundaries as determined by the ±1.5 standard deviation rule could not overlap. Second, the boundary limits themselves were required to be separated by a minimum of .5 to a maximum of 1.5 within-cluster standard deviations. The actual separation value was chosen randomly from within this uniform interval. On the remaining dimensions, the locations of the cluster centroids were chosen randomly and without regard to the other clusters present. Since the range of the remaining dimensions was constrained to be a maximum of two-third's of the first dimension, overlap of cluster boundaries was common.

Three design factors were introduced in the study by modifying the data generation algorithm. First, the minimum allowable spacing between clusters was controlled. The second factor adjusted the maximum within-cluster standard deviations. The third factor consisted of various methods for introducing error into the data. The three factors can be considered "between-subjects" in nature since 108 different data sets were generated for each combination of separation, maximum variance, and error type. The factors are discussed in turn.

The first factor with two levels varied the spacing between clusters. The first level corresponded to the original version of the data generator. Cluster overlap was not allowed on the first dimension of the variable space.

Table 1

Experimental Design Factors

Factor Name	Number of Levels	Name of Level
Standardization Procedure	6	Z_0
		Z_1 & Z_2
		Z_3
		Z_4 & Z_5
		Z_6
		Z_7
Cluster Separation	2	Near Spacing Distant Spacing
Maximum Within Cluster Variances	2	16 100
Error Conditions	4	Error Free Distance Perturbed 2-Dimensional Noise 20% Outliers
Clustering Methods	4	Single Link Complete link Group Average Ward's Minimum Variance
Coverage Level	4	k k + 3 k + 6 k + 9

NOTE: Cluster Separation, Maximum Variances, and Error Conditions can be considered between-subject factors. Standardization Procedures, Clustering Methods, and Coverage Level are within-subject factors since these are measured repeatedly on each data set. In the present study, "Subject" refers to a generated data set. Subject scores are the recovery values computed from the clustering solution of the data set.

The typical separation between cluster boundaries on the first dimension corresponded to about one within-cluster standard deviation. Overlap of cluster boundaries was allowed on the remaining dimensions of the variable space. The second level of the factor eliminated the separation between cluster boundaries on the first dimension. Rather, cluster boundaries were contiguous on that dimension. No change was made for the remaining dimensions. Further, the restriction that all elements must fall within ±1.5 standard deviations of the cluster centroid was dropped for all dimensions. That is, no truncation of the multivariate normal distributions occurred. As such, limited cluster overlap was present in the data. The condition may favor standardization procedures which reduce the impact of intermediate points between clusters.

The second factor with two levels adjusted the maximum variance of the clusters on any given dimension. The length of each cluster was used to define the within-cluster standard deviation on the dimension. The total length was specified to represent three standard deviations for the cluster on the dimension. The first level of the factor used the original version of the data generator which required that all cluster lengths be sampled from the uniform interval of 10 to 40 units. Thus, the maximum ratio of the largest to smallest within-cluster standard deviations was 4 (16 for the variance). The second level of the factor randomly sampled the cluster lengths from the uniform range of 10 to 100. The second level allowed for maximum ratios as large as 10 between the within-cluster standard deviations (100 for variance). The second level should provide a test of whether or not the various forms of standardization of variables are necessary to adjust for differing variances across clusters.

The third factor possessed four levels and represented differing methods for the introduction of error into the data. The degree of cluster distinctness can be greatly diminished or masked by such errors. The first level corresponded to the error-free condition as produced by the data generator algorithm. The second level involved the error perturbation of the distances between points by adding normally distributed noise to the original error-free coordinates. The level is identified as the high error condition in Milligan (1985). The distance perturbation condition would correspond to those applied situations where the measurement process is noisy, as is often the case with behavioral studies.

The third error condition involved adding two random noise dimensions to the basic set of variables that defined the error-free cluster structure in the data. This would correspond to those cases where the applied researcher is not sure which variables define the clustering in the data and has inadvertently included irrelevant variables in the analysis. This condition was found to be one of the most severe forms of error in the study by Milligan (1980).

The fourth error level involved the inclusion of 20% additional outlier points to each data set. Thus, the data sets consisted of 60 points rather than 50. The outliers were generated by using the existing cluster centroids, but with variances for each variable multiplied by a factor of 9. Before a generated point was accepted as an outlier, it was verified that the point did not fall within the cluster boundaries on all dimensions for any given cluster. It was expected that standardization procedures may provide better recovery by reducing the effect of outliers and intermediates between clusters.

The reader should note that as with any simulation study, generalization of the results is limited by the nature of the constructed data. Results based only on the error-free data in the distant-spacing condition would have limited applicability. Hopefully, the design factors and the various forms of error will provide a greater degree of generalization.

2.2 Clustering Methods and Recovery Measure

Four agglomerative hierarchical clustering methods were used to generate partition solutions and formed one factor in the overall design. These were the single link, complete link, group average (UPGMA), and Ward's minimum variance methods (see Cormack 1971). The methods were chosen to provide a wide range of recovery effectiveness and to offer some generalization for the results. The dissimilarity measure applied to the resulting values of Z_0 to Z_7 was the Euclidean distance between points (see the formula for **NT-SYS** in Blashfield 1977). Since the results within pairs Z_1 & Z_2 and Z_4 & Z_5 are identical, only one set of values for each pair is reported.

Recovery of cluster structure was measured at the level in the hierarchical solution that corresponded to the correct number of clusters known to exist in the data. Recovery also was measured at selected levels preceeding the correct solution partition in order to explore the effects of reduced coverage as suggested by Edelbrock (1979). Edelbrock argued that not all observations must be classified for an effective or useful classification. Thus, some elements can be left unclustered or incompletely clustered. If k represents the true number of clusters in the data, then recovery was measured in the present study when k, $k + 3$, $k + 6$, and $k + 9$ clusters were present in the solution. Hence, reduced coverage implies that more clusters are present in the obtained partition than actually exist in the underlying structure. The coverage factor is the last of the six factors in the design.

The Hubert and Arabie (1985) corrected Rand index was used as the recovery statistic and dependent variable. The index has been shown to possess a number of desirable properties by its originators and by Milligan and

Cooper (1986). The index produces a value of 1.00 only if the partitions recovered by the clustering method exactly match the clusters known to be present in the data. When the recovery of the true cluster structure falls to chance levels, the index returns values near 0.0. The closer the recovery value is to 1.00, the better the recovery of the underlying cluster structure.

The factors involving standardization procedures, clustering methods, and coverage level can be considered to be "repeated measures." That is, from each data set generated through the various combinations of separation, maximum variance, and error type, observations were obtained for all standardization procedures, clustering methods, and coverage levels.

3. Results

Recovery results are presented in Table 2 for the factors involving standardization procedures, three of the four error conditions, and separation levels. Results for the 20% outlier error condition are presented later. In determining the cell means, only the first coverage level was used (level *k*), which corresponds to the correct number of clusters known to exist in the data. Each entry in the table represents the average corrected Rand statistic value based on 864 data sets (108 data sets by 2 levels for maximum variance by 4 levels for clustering methods). The main effects tests for all three factors were found to be statistically significant. Further, the three-way interaction was highly significant among these three variables. As an aid in interpreting the effects in the table, the Ryan-Einot-Gabriel-Welsch multiple range test was conducted for each combination of error condition and separation level across the standardization procedures. This resulted in a pairwise comparison of the means within any given column. (For a discussion of the test, see Ramsey 1978, and the GLM procedure in the *SAS User's Guide: Statistics*, 1985.) The asterisk notation in the table indicates those standardization procedures which were equivalent to each other, but which gave significantly better recovery than the remaining procedures in the column. These sets of techniques are referred to as the statistically equivalent superior groups in the discussion. The (L)-notation is used when one or more standardization procedures were found to give recovery which was significantly lower than all other techniques.

As would be expected, the best recovery was obtained with error-free data and widely spaced clusters. Standardization procedures Z_0 and Z_3 to Z_6 produced excellent recovery. This condition was the only case where the untransformed data, Z_0, gave an average recovery value which was greater than that for any standardization procedure. The ranking procedure, Z_7, produced recovery which was significantly lower than any other technique. Effectively, the same comments hold for the error-free data with closely

Table 2

Effect of Separation and Error Type on Cluster Recovery for the Standardization Procedures at the Correct Cluster Level

	Cluster Separation					
	Near Spacing			Distant Spacing		
Standardization Procedure	Error Free	Distance Perturbed	2-Dim. Noise	Error Free	Distance Perturbed	2-Dim. Noise
Z_0	.808*	.611*	.567(L)	.981*	.799*	.684(L)
Z_1 & Z_2	.759	.582	.675*	.931	.741	.838*
Z_3	.819*	.630*	.619	.974*	.810*	.778
Z_4 & Z_5	.809*	.622*	.650*	.957*	.789*	.847*
Z_6	.812*	.593*	.616	.972*	.786*	.749
Z_7	.725(L)	.566	.626	.848(L)	.707(L)	.751
Overall	.788	.601	.635	.944	.770	.791

NOTE: Entries are mean adjusted Rand statistic values. Asterisk indicates membership in the statistically equivalent superior group as determined by the Ryan-Einot-Gabriel-Welsch multiple range test for the column. (L) indicates that the procedure gave significantly lower recovery than any other method in the column. The marginal cell means in the "Overall" entry were not included in the multiple range tests.

spaced clusters, except that recovery was about .156 less on the average for all procedures.

In the distance perturbed error condition, the same set of procedures were found in the superior group as for the error-free data (Z_0, Z_3 to Z_6). Again, procedure Z_7 gave the lowest average recovery, and this performance was significantly worse than the other techniques when clusters were widely spaced. The distance perturbation condition reduced recovery by an average of about .18 below the corresponding error-free data. Overall, the distance perturbed condition resulted in the lowest average recovery rates of all three error conditions.

It is interesting to note that for both the error-free and the distance perturbed conditions, the traditional z-score transformation (Z_1) did not place in the superior group for either distant or closely spaced clusters. Hence, the most popular standardization method in the literature is not the most effective for these error conditions.

It was anticipated that a reduction in cluster separation might have a relatively more serious impact on Z_0. However, the results in Table 2 for the near spacing condition indicate that Z_0 was in the superior group for the error-free data and the distance perturbed condition. This is the same result as found with the distant spacing condition. It seems that intermediates and the presence of some cluster overlap do not always seriously impact recovery performance with unstandardized data.

The results for the condition involving the addition of two random noise dimensions to the data show an average reduction of about .15 in recovery from the error-free data. It is important to note that membership in the superior group changes in the two-dimensional noise condition. The change provides insight into the source of the significant interaction between error, separation, and procedures. Specifically, we now have Z_1 & Z_2 and Z_4 & Z_5 as the best available techniques. Procedures Z_3 and Z_6 are no longer in the superior set. This implies that when irrelevant variables have been included in the data set, standardization by division by the maximum score (Z_3) or by ΣX (Z_6) is not helpful in the subsequent clustering. Similarly, the untransformed data, Z_0, was found to produce significantly lower recovery rates than for all other techniques.

A related aspect to the interaction effect should be noted. The most severe error environment for the untransformed data, Z_0, was the two-dimensional noise condition. This result lead Milligan (1980) to conclude that this form of error was a serious problem for an applied researcher. However, for the standardization procedures that comprise the superior group in the two-dimensional noise condition (Z_1 & Z_2, Z_4 & Z_5), a lower mean recovery level was found in the distance perturbed error condition. Hence, certain standardization procedures can change the relative impact of different forms of error in the data. Finally, the inclusion of Z_1 and Z_2 in the superior group indicates that there can be situations where these forms are useful. One can speculate that the popularity of Z_1 stems as much from its familiarity as its success in those real life cases where variables have been included in the data which are irrelevant to the cluster structure.

The overall marginal cell means for the separation effect and the maximum variance ratio factor are presented in Table 3. Again, the asterisk notation indicates the statistically equivalent superior procedures within each column. When considering the overall results for the separation factor, one can see the reduction in recovery when the spacing between clusters was reduced. Standardization Z_7 was found to give significantly lower recovery than any other procedure. Although procedures Z_3 at both separation levels

Table 3

Effect of Cluster Separation and Maximum Within Cluster Variance Ratios on Standardization Procedures at the Correct Cluster Level

Standardization Procedure	Separation Level		Max. Variance Ratio		Global Variances Experiment
	Near	Distant	16	100	
Z_0	.662	.821	.745	.739	.621(L)
Z_1 & Z_2	.672	.837	.755	.754	.936
Z_3	.689*	.854*	.771*	.772*	.984*
Z_4 & Z_5	.693*	.864*	.778*	.780*	.968*
Z_6	.674*	.836	.757	.753	.981*
Z_7	.639(L)	.768(L)	.693(L)	.713(L)	.839
Overall	.674	.835	.754	.756	.888

NOTE: See note to Table 2 for explanation of symbols. Means are presented for coverage level k. The means do not include the data from the 20% outlier condition. Means for the global variances experiment were based only on the error-free data with inflated variances as specified in the text.

and Z_6 at the level of near spacing are found in the superior group, their overall effectiveness is limited due to the interaction effect seen in Table 2. Only procedures Z_4 and Z_5 are found in the statistically equivalent superior group in all error conditions. Note also that the unstandardized data, Z_0, did not fall into the superior group in the tests in Table 3 despite its performance in the error-free and distance perturbed conditions. Apparently, the performance of Z_0 with the two-dimensional noise condition was so poor as to significantly lower the overall marginal mean.

The results for the maximum variance ratio factor presented in Table 3 are somewhat surprising. The main effect for this factor was found to be insignificant. The overall marginal means can be seen to be rather close in value for each standardization procedure.

Because of the unexpected results obtained for the maximum variance factor, additional experimental conditions were created to study this effect. First, a third level for the maximum variance ratios was considered. This involved maximum within-cluster variance ratios as large as 500. Again, no substantial differences were found between this level and the first two levels involving ratios of 16 and 100. Second, a different approach was taken which involved adjusting the overall variance on the variables, rather than the within-cluster variances. Several combinations were tested and results which were found to be typical of these conditions are presented as the last column in Table 3. The results for the global variance experiment were obtained by using data generated to be error-free with distant cluster spacing and variances adjusted for selected variables. In particular, the variances for the second and fourth variables were increased by a factor of 500 and 1,000, respectively. Variances for the remaining dimensions were left unchanged. Note that the first variable of the space was not changed because it could uniquely define the clusters. A vastly inflated variance for the first variable would serve to enhance rather than inhibit recovery. By contrast, the remaining variables, including the second and fourth, would involve some level of cluster overlap. Inflated variances for these variables should result in increased confusion as to the correct cluster assignments for a portion of each data set.

The results in Table 3 for the global variance experiment did indicate that the untransformed data, Z_0, resulted in significantly lower recovery than for any other form of standardization. Further, the obtained mean of .621 can be compared to the mean value of Z_0 for the uninflated variables which was found to be .895. Hence, substantial differences between the overall variances among variables can significantly reduce cluster recovery for Z_0.

As seen in the last column of Table 3, any form other than Z_0 resulted in improved recovery of cluster structure. This confirms the view that when substantial heterogeneity of variances exists, some form of standardization is needed. However, not all forms are equally effective. The use of ranks, Z_7, resulted in recovery which was significantly worse than standardization forms Z_1 to Z_6. Forms Z_1 & Z_2 did produce a significantly better recovery mean than for Z_0 or Z_7, but Z_1 & Z_2 were not found in the best performance group. Only Z_3, Z_4 & Z_5, and Z_6 were found in the statistically equivalent superior group.

As mentioned in the introduction, Späth (1980) suspected that Z_1 may not perform well if substantial differences existed between the within-cluster standard deviations. The results in Table 3 for Z_1 do not support this argument. The results for the maximum variance factor fail to provide direct support for the claim. Recovery for Z_1 was effectively the same in all ratio conditions. Apparently, differing within-cluster variances has little impact on recovery for the procedures and methods tested, at least within the range of the maximum variance ratios considered in the study. On the other hand, as found in the global variances experiment, substantial heterogeneity of variances strongly favors Z_1 over raw or ranked data (Z_0 and Z_7).

Table 4

Interaction of Standardization Procedure and Clustering Method

Standardization Procedure	Clustering Method			
	Single Link	Complete Link	Group Average	Ward's Method
Z_0	.608*	.750	.811	.798(L)
Z_1 & Z_2	.577	.778	.800	.864*
Z_3	.622*	.793*	.835*	.836
Z_4 & Z_5	.609*	.815*	.839*	.851*
Z_6	.616*	.761	.813	.828
Z_7	.494(L)	.730(L)	.810	.781(L)
Overall	.589	.777	.819	.834

NOTE: See note to Table 2 for explanation of symbols. Means are presented for coverage level k. The means do not include the data from the 20% outlier condition.

Returning to the original experimental design, the recovery results for clustering methods and standardization procedures are displayed in Table 4. As would be expected from previous simulation studies (see Milligan and Cooper 1987), the single link technique was least effective while the group average and Ward's methods gave the best overall recovery. The main effect of methods was found to be significant. More importantly, the interaction between methods and standardization procedures also was significant, One aspect of the interaction effect can be detected by noting that the superior equivalent group changes from one clustering method to the next. For one or more methods, each standardization procedure except Z_7 was a member of a superior group. However, only Z_4 & Z_5 were found in the best performing group for all methods. It would appear that the relative performance of Z_4 & Z_5 is independent of the clustering methods examined in the present study.

A different source of the interaction is the change in identity of the poorest performing standardization procedures. For the group average method, no standardization procedure was found to give significantly lower

Table 5

Effect of Coverage Level and Standardization Procedures on Cluster Recovery for Error Free Data

Standardization Procedure	Coverage Level			
	k	k+3	k+6	k+9
Z_0	.894*	.669*	.529*	.428*
Z_1 & Z_2	.845	.653	.521*	.423*
Z_3	.896*	.667*	.529*	.428*
Z_4 & Z_5	.883*	.661*	.524*	.424*
Z_6	.892*	.674*	.534*	.429*
Z_7	.787(L)	.568(L)	.440(L)	.350(L)

NOTE: See note to Table 2 for explanation of symbols.

recovery, whereas Z_7 was significantly worse for the single and complete link methods. Conversion to ranks (Z_7) had a particularly bad effect on the single link method. Although the single link method is monotone invariant with respect to the input distances, it is not invariant to standardization of the coordinates prior to the computation of the distances. Since the method attempts to minimize the largest chain link between two points (see Johnson 1967), the loss of information due to ranking the coordinate values apparently is quite severe.

In contrast to the group average method, both Z_0 and Z_7 formed a low performance set with Ward's method. It is worth noting that Z_1, the traditional standard score formula, was the most effective form for Ward's method. Procedure Z_1 was absent from the superior set for all other methods. Hence, it appears that the relative performance of Z_1 is dependent on the clustering method selected.

The effect of reduced coverage is presented in Table 5. The results in Table 5 do not include the data obtained from the outlier error condition. Clearly, as the coverage level is decreased from k to k + 9, recovery declined for all standardization procedures. As would be expected from the cell means in the table, the main effect of coverage was significant, and an interaction between coverage and procedures was detected. Again, post hoc tests help reveal the source of the interaction. At the highest coverage level (*k*), procedures Z_1 & Z_2 and Z_7 were not in the superior performance group. Recall that *k* is the level that corresponds to the exact number of clusters known to exist in the data. For coverage levels $k + 6$ and $k + 9$, the differential performance between procedures Z_0 to Z_6 has disappeared. Only Z_7 failed to match the performance of the other procedures at all coverage levels.

The results concerning Z_0 and Z_1 across coverage levels are consistent with the findings of Edelbrock (1979). That is, a significant difference existed between Z_0 and Z_1 at level *k*, and no significant effect existed at lower levels of coverage. However, a serious discrepancy exists concerning the overall direction of recovery as coverage decreases. Specifically, as coverage decreased in the present study, recovery declined as measured by the Hubert and Arabie corrected Rand index. This is the reverse of the effect found by Edelbrock (1979) and also by Scheibler and Schneider (1985). These studies found that recovery *improved* as coverage decreased. Typically, median recovery values were at or near 1.00 by level $k + 9$. Both studies used the kappa statistic as the recovery measure. There is something troubling in having the criterion statistic return values of 1.00 at level $k + 9$. In the best possible case, one has a set of subpartitions of the actual *k* clusters that exist in the data. An index that fails to indicate such incomplete clustering is misleading. The kappa statistic produces values of 1.00 when recovery is exactly correct, and in cases when an incomplete clustering is present in the solution. In fact, both studies indicate that kappa approaches 1.00 as the number of clusters in the solution approaches *n*, the number of elements in the data set. This property is shared with a number of other recovery indices, including the original version of the Rand index. The property is not desirable and is not shared with the Hubert and Arabie corrected Rand measure (for discussion and additional references, see Milligan and Cooper 1986).

The effect of the addition of outlier points to the data sets on the group average method appear in Table 6. The data sets consisted of 60 points each, of which 10 were outliers to all clusters. The recovery values presented in the table were determined from the 50 points that defined the major clusters in the data. The outliers were not included in the calculation of the recovery index. They were allowed to remain as single point entities or merge with existing clusters as determined by the algorithm.

As indicated in Table 6, recovery was best at low levels of coverage ($k + 9$) for all standardization procedures except Z_7 (ranking). At this level, the outliers tended to remain single point clusters. Recovery declined as coverage was increased to *k* for Z_0 to Z_6. At level *k*, recovery for these proced-

Table 6

Effect of Outliers on the Group Average Method Across Coverage Levels

Standardization Procedure	Coverage Level			
	k	k+3	k+6	k+9
Z_0	.338	.583	.771*	.831*
Z_1 & Z_2	.219(L)	.553	.719*	.814*
Z_3	.308	.616	.775*	.848*
Z_4 & Z_5	.303	.602	.781*	.829*
Z_6	.318	.602	.798*	.832*
Z_7	.623*	.751*	.675(L)	.537(L)

NOTE: See note to Table 2 for explanation of symbols.

ures was poor and corresponds to the lowest average values seen in the experiment. Note that recovery with Z_1 & Z_2 was significantly worse than all other procedures at level k. Combining this result with the findings from Table 4, Z_1 & Z_2 appear to be the poorest choices for use with the group average method. At level k, the outliers generally were forced to merge with existing clusters in the data. Clearly, the ability of the algorithm to recover the major embedded clusters was severely impaired by the outliers at high coverage levels. The only exception was for Z_7 which showed some improvement in recovery means from $k + 9$ to $k + 3$, but declined at level k. Procedure Z_7 performed significantly better than all other procedures at levels k and $k + 3$, but significantly worse than the other procedures at $k + 6$ and $k + 9$. Apparently, the ranking process attenuated the effect of the outliers at high coverage levels, but interfered with recovery at low levels.

The results for the outlier condition on Ward's method are presented in Table 7. The recovery pattern across coverage levels is the reverse of that for the group average method. That is, recovery improved as coverage increased. Standardization procedures Z_0 to Z_6 were statistically equivalent in terms of

Table 7

Effect of Outliers on Ward's Method Across Coverage Levels

Standardization Procedure	Coverage Level			
	k	k+3	k+6	k+9
Z_0	.869*	.731*	.587*	.477*
Z_1 & Z_2	.870*	.711*	.592*	.468*
Z_3	.878*	.707*	.565*	.481*
Z_4 & Z_5	.877*	.711*	.577*	.473*
Z_6	.859*	.715*	.591*	.474*
Z_7	.763(L)	.515(L)	.402(L)	.328(L)

NOTE: See note to Table 2 for explanation of symbols.

cluster recovery at each level. Procedure Z_7 resulted in significantly worse recovery at all coverage levels. The best recovery was found at level *k* for all procedures, despite the presence of the outliers. Mean values at this level are quite good and display better recovery performance than the results for the other error conditions in Table 2. The result does suggest an interesting paradox. Ward's method is a minimum sum of squares technique. It is generally recognized that least squares procedures are sensitive to outliers. Yet Ward's method shows a strong robustness against their presence in the data. Further work on this phenomenon is warranted.

In comparison, the effect of outliers on the single and complete link methods generally produced poorer recovery values. For the single link method, the pattern across coverage levels was similar to that found in Table 6 for the group average technique, except for mean values which were about .20 lower in all cases. Recovery for the complete link method was best at levels $k + 3$ and $k + 6$ using standardization procedures Z_0 to Z_3. However, the means were at least .10 lower than for the group average or Ward's techniques.

4. Conclusions

The results in Tables 2-5 indicate that standardization methods involving division by the range (Z_4 and Z_5) offer the best recovery of the underlying cluster structure. This result holds across error conditions, separation distances, clustering methods, and coverage levels. In no case did either Z_4 or Z_5 fail to be in the statistically equivalent superior group. Apparently, standardization by division by the range of the variable consistently aids in cluster recovery and is robust across a wide variety of conditions.

On at least one occasion, all other procedures were found to give recovery which was significantly worse than Z_4 or Z_5. This included the unstandardized values represented by Z_0. The unstandardized values were competitive only with error-free or distance perturbed data, and with the single link method over all conditions. Similarly, the traditional *z*-score formula, Z_1, was not especially effective. Procedure Z_1 was in the superior group only when used with Ward's method, or at low coverage levels, or when random noise dimensions had been added to the data. The performances of Z_0 and Z_1 are almost complementary. When one procedure was not working well, the other usually was effective.

Procedure Z_7, based on ranking the data, performed rather poorly. The procedure was never a member of the superior group in any condition. Often, it was found to produce significantly lower recovery than any other approach. Apparently, the loss of information about the magnitude of differences between coordinates was too severe. All other standardization procedures retained this information to some degree.

Finally, it is important to keep in perspective the relative improvement that can be obtained with the selection of an effective standardization procedure. The difference between the corrected Rand index values for the best and worst standardization procedures in each column of Tables 2-5 averages about .12. This contrasts to a difference of .15 to .19 in corrected Rand values between the error-free data and the other two error conditions. Similarly, the difference between the best and worst clustering methods was .24. Clearly, minimization of different forms of error in the data and selection of an effective clustering method may offer a greater return in terms of cluster recovery. Deciding on a suitable form of standardization of variables can improve recovery of the true cluster structure, but it is only one of several decisions faced by the applied researcher.

References

ANDERBERG, M.R. (1973), *Cluster Analysis for Applications*, New York: Academic Press.

BAYNE, C.K., BEAUCHAMP, J.J., BEGOVICH, C.L., and KANE, V.E. (1980), "Monte Carlo Comparisons of Selected Clustering Procedures," *Pattern Recognition, 12,* 51-62.

BLASHFIELD, R.K. (1976), "Mixture Model Tests of Cluster Analysis: Accuracy of Four Agglomerative Hierarchical Methods," *Psychological Bulletin, 83,* 377-388.

BLASHFIELD, R.K. (1977), "The Equivalence of Three Statistical Packages for Performing Hierarchical Cluster Analysis," *Psychometrika, 42,* 429-431.

BURR, E.J. (1968), "Clustering Sorting with Mixed Character Types: I. Standardization of Character Values," *Australian Computer Journal, 1,* 97-99.

CAIN, A.J., and HARRISON, G.A. (1958), "An Analysis of the Taxonomist's Judgement of Affinity," *Proceedings of the Zoological Society of London, 131,* 85-98.

CARMICHAEL, J.W., GEORGE, J.A., and JULIUS, R.S. (1968), "Finding Natural Clusters," *Systematic Zoology, 17,* 144-150.

CONOVER, W.J., and IMAN, R.L. (1981), "Rank Transformation as a Bridge Between Parametric and Nonparametric Statistics," *The American Statistician, 35,* 124-129.

CORMACK, R.M. (1971), "A Review of Classification," *Journal of the Royal Statistical Society, Series A, 134,* 321-367.

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

DUBES, R., and JAIN, A.K. (1980), "Clustering Methodologies in Exploratory Data Analysis," *Advances in Computers, 19,* 113-228.

EDELBROCK, C. (1979), "Comparing the Accuracy of Hierarchical Clustering Algorithms: The Problem of Classifying Everybody," *Multivariate Behavioral Research, 14,* 367-384.

EVERITT, B.S. (1980), *Cluster Analysis* (2nd ed.), London: Heinemann.

FLEISS, J.L., and ZUBIN, J. (1969), "On the Methods and Theory of Clustering," *Multivariate Behavioral Research, 4,* 235-250.

GORDON, A.D. (1981), *Classification: Methods for the Exploratory Analysis of Multivariate Data,* London: Chapman and Hall.

GOWER, J.C. (1971), "A General Coefficient of Similarity and Some of Its Properties," *Biometrics, 27,* 857-871.

HALL, A.V. (1965), "The Peculiarity Index, a New Function for Use in Numerical Taxonomy," *Nature, 206,* 952.

HALL, A.V. (1969), "Group Forming and Discrimination with Homogeneity Functions," in *Numerical Taxonomy,* ed. A.J. Cole, New York: Academic Press.

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

HOHENEGGER, J. (1986), "Weighted Standardization - A General Data Transformation Method Preceeding Classification Procedures," *Biometrical Journal, 28,* 295-303.

HUBERT, L., and ARABIE, P. (1985), "Comparing Partitions," *Journal of Classification, 2,* 193-218.

JARDINE, N., and SIBSON, R. (1971), *Mathematical Taxonomy,* New York: Wiley.

JOHNSON, S.C. (1967), "Hierarchical Clustering Schemes," *Psychometrika, 32,* 241-254.

KAUFMAN, R.L. (1985), "Issues in Multivariate Cluster Analysis: Some Simulation Results," *Sociological Methods and Research, 13,* 467-486.

LANCE, G.N., and WILLIAMS, W.T. (1967), "Mixed Data Classificatory Programs: I. Agglomerative Systems," *Australian Computer Journal, 1,* 15-20.

LORR, M. (1983), *Cluster Analysis for the Social Sciences,* San Francisco: Jossey-Bass.

MILLIGAN, G.W. (1980), "An Examination of the Effect of Six Types of Error Perturbation on Fifteen Clustering Algorithms," *Psychometrika, 45,* 325-342.

MILLIGAN, G.W. (1981), "A Review of Monte Carlo Tests of Cluster Analysis," *Multivariate Behavioral Research, 16,* 379-407.

MILLIGAN, G.W. (1985), "An Algorithm for Generating Artificial Test Clusters," *Psychometrika, 50*, 123-127.

MILLIGAN, G.W., and 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., and COOPER, M.C. (1987), "Methodological Review: Clustering Methods," *Applied Psychological Measurement, 11*, 329-354.

MORRISON, D.G. (1967), "Measurement Problems in Cluster Analysis," *Management Science, 13*, 775-780.

OVERALL, J.E., and KLETT, C.J. (1972), *Applied Multivariate Analysis*, New York: McGraw-Hill.

RAMSEY, P.H. (1978), "Power Differences Between Pairwise Multiple Comparisons," *Journal of the American Statistical Association, 73*, 479-487.

ROMESBURG, H.C. (1984), *Cluster Analysis for Researchers*, Belmont, CA: Lifetime Learning Publications.

SAS User's Guide: Statistics, (1985), Cary, NC: SAS Institute.

SAWERY, W.L., KELLER, L., and CONGER, J.J. (1960), "An Objective Method of Grouping Profiles by Distance Functions and Its Relation to Factor Analysis," *Educational and Psychological Measurement, 20*, 651-674.

SCHEIBLER, D., and SCHNEIDER, W. (1985), "Monte Carlo Tests of the Accuracy of Cluster Analysis Algorithms - A Comparison of Hierarchical and Nonhierarchical Methods," *Multivariate Behavioral Research, 20*, 283-304.

SNEATH, P.H.A., and SOKAL, R.R. (1973), *Numerical Taxonomy*, San Francisco: Freeman.

SOKAL, R.R. (1961), "Distance as a Measure of Taxonomic Similarity," *Systematic Zoology, 10*, 70-79.

SOKAL, R.R., and ROHLF, F.J. (1969), *Biometry, the Principles and Practice of Statistics in Biological Research*, San Francisco: Freeman.

SPATH, H. (1980), *Cluster Analysis Algorithms*, New York: Wiley.

STODDARD, A.M. (1979), "Standardization of Measures Prior to Cluster Analysis," *Biometrics, 35*, 765-773.

TUKEY, J.W. (1977), *Exploratory Data Analysis*, Reading, Ma.: Addison-Wesley.

WILLIAMS, W.T., DALE, M.B., and MAC NAUGHTON-SMITH, P. (1964), "An Objective Method of Weighting in Similarity Analysis," *Nature, 201*, 426.

WILLIAMS, W.T., LAMBERT, J.M., and LANCE, G.N. (1966), "Multivariate Methods in Plant Ecology. V. Similarity Analyses and Information Analysis," *Journal of Ecology, 54*, 427-445.