Chapter 23
The CLUSTER Procedure

Chapter Table of Contents

Chapter 23
The CLUSTER Procedure

Overview

The CLUSTER procedure hierarchically clusters the observations in a SAS data set using one of eleven methods. The CLUSTER procedure finds hierarchical clusters of the observations in a SAS data set. The data can be coordinates or distances. If the data are coordinates, PROC CLUSTER computes (possibly squared) Euclidean distances. If you want to perform a cluster analysis on non-Euclidean distance data, it is possible to do so by using a TYPE=DISTANCE data set as input. The %DISTANCE macro in the SAS/STAT sample library can compute many kinds of distance matrices.

One situation where analyzing non-Euclidean distance data can be useful is when you have categorical data, where the distance data are calculated using an association measure. For more information, see Example 23.5 on page 916.

The clustering methods available are average linkage, the centroid method, complete linkage, density linkage (including Wong's hybrid and kth-nearest-neighbor methods), maximum likelihood for mixtures of spherical multivariate normal distributions with equal variances but possibly unequal mixing proportions, the flexible-beta method, McQuitty's similarity analysis, the median method, single linkage, two-stage density linkage, and Ward's minimum-variance method.

All methods are based on the usual agglomerative hierarchical clustering procedure. Each observation begins in a cluster by itself. The two closest clusters are merged to form a new cluster that replaces the two old clusters. Merging of the two closest clusters is repeated until only one cluster is left. The various clustering methods differ in how the distance between two clusters is computed. Each method is described in the section "Clustering Methods" on page 854.

The CLUSTER procedure is not practical for very large data sets because, with most methods, the CPU time varies as the square or cube of the number of observations. The FASTCLUS procedure requires time proportional to the number of observations and can, therefore, be used with much larger data sets than PROC CLUSTER. If you want to cluster a very large data set hierarchically, you can use PROC FASTCLUS for a preliminary cluster analysis producing a large number of clusters and then use PROC CLUSTER to cluster the preliminary clusters hierarchically. This method is used to find clusters for the Fisher Iris data in Example 23.3, later in this chapter.

PROC CLUSTER displays a history of the clustering process, giving statistics useful for estimating the number of clusters in the population from which the data are sampled. PROC CLUSTER also creates an output data set that can be used by the TREE procedure to draw a tree diagram of the cluster hierarchy or to output the cluster membership at any desired level. For example, to obtain the six-cluster solution, you could first use PROC CLUSTER with the OUTTREE= option then use this output data set as the input data set to the TREE procedure. With PROC TREE, specify NCLUSTERS=6 and the OUT= options to obtain the six-cluster solution and draw a tree diagram. For an example, see Example 66.1 in Chapter 66, "The TREE Procedure."

Before you perform a cluster analysis on coordinate data, it is necessary to consider scaling or transforming the variables since variables with large variances tend to have more effect on the resulting clusters than those with small variances. The ACECLUS procedure is useful for performing linear transformations of the variables. You can also use the PRINCOMP procedure with the STD option, although in some cases it tends to obscure clusters or magnify the effect of error in the data when all components are retained. The STD option in the CLUSTER procedure standardizes the variables to mean 0 and standard deviation 1. Standardization is not always appropriate. See Milligan and Cooper (1987) for a Monte Carlo study on various methods of variable standardization. You should remove outliers before using PROC PRINCOMP or before using PROC CLUSTER with the STD option unless you specify the TRIM= option.

Nonlinear transformations of the variables may change the number of population clusters and should, therefore, be approached with caution. For most applications, the variables should be transformed so that equal differences are of equal practical importance. An interval scale of measurement is required if raw data are used as input. Ordinal or ranked data are generally not appropriate.

Agglomerative hierarchical clustering is discussed in all standard references on cluster analysis, for example, Anderberg (1973), Sneath and Sokal (1973), Hartigan (1975), Everitt (1980), and Spath (1980). An especially good introduction is given by Massart and Kaufman (1983). Anyone considering doing a hierarchical cluster analysis should study the Monte Carlo results of Milligan (1980), Milligan and Cooper (1985), and Cooper and Milligan (1988). Other essential, though more advanced, references on hierarchical clustering include Hartigan (1977, pp. 60–68; 1981), Wong (1982), Wong and Schaack (1982), and Wong and Lane (1983). Refer to Blashfield and Aldenderfer (1978) for a discussion of the confusing terminology in hierarchical cluster analysis.

Getting Started

The following example demonstrates how you can use the CLUSTER procedure to compute hierarchical clusters of observations in a SAS data set.

Suppose you want to determine whether national figures for birth rates, death rates, and infant death rates can be used to determine certain types or categories of countries. You want to perform a cluster analysis to determine whether the observations can be formed into groups suggested by the data. Previous studies indicate that the clusters computed from this type of data can be elongated and elliptical. Thus, you need to perform some linear transformation on the raw data before the cluster analysis.

The following data[*] from Rouncefield (1995) are birth rates, death rates, and infant death rates for 97 countries. The DATA step creates the SAS data set Poverty:

```
data Poverty;
   input Birth Death InfantDeath Country $20. @@;
   datalines;
24.7  5.7  30.8 Albania             12.5 11.9  14.4 Bulgaria
13.4 11.7  11.3 Czechoslovakia      12   12.4   7.6 Former_E._Germany
11.6 13.4  14.8 Hungary             14.3 10.2    16 Poland
13.6 10.7  26.9 Romania               14    9  20.2 Yugoslavia
17.7   10    23 USSR                15.2  9.5  13.1 Byelorussia_SSR
13.4 11.6    13 Ukrainian_SSR       20.7  8.4  25.7 Argentina
46.6   18   111 Bolivia             28.6  7.9    63 Brazil
23.4  5.8  17.1 Chile               27.4  6.1    40 Columbia
32.9  7.4    63 Ecuador             28.3  7.3    56 Guyana
34.8  6.6    42 Paraguay            32.9  8.3 109.9 Peru
  18  9.6  21.9 Uruguay             27.5  4.4  23.3 Venezuela
  29 23.2    43 Mexico                12 10.6   7.9 Belgium
13.2 10.1   5.8 Finland             12.4 11.9   7.5 Denmark
13.6  9.4   7.4 France              11.4 11.2   7.4 Germany
10.1  9.2    11 Greece              15.1  9.1   7.5 Ireland
 9.7  9.1   8.8 Italy               13.2  8.6   7.1 Netherlands
14.3 10.7   7.8 Norway              11.9  9.5  13.1 Portugal
10.7  8.2   8.1 Spain               14.5 11.1   5.6 Sweden
12.5  9.5   7.1 Switzerland         13.6 11.5   8.4 U.K.
14.9  7.4     8 Austria              9.9  6.7   4.5 Japan
14.5  7.3   7.2 Canada              16.7  8.1   9.1 U.S.A.
40.4 18.7 181.6 Afghanistan         28.4  3.8    16 Bahrain
42.5 11.5 108.1 Iran                42.6  7.8    69 Iraq
22.3  6.3   9.7 Israel              38.9  6.4    44 Jordan
26.8  2.2  15.6 Kuwait              31.7  8.7    48 Lebanon
45.6  7.8    40 Oman                42.1  7.6    71 Saudi_Arabia
29.2  8.4    76 Turkey              22.8  3.8    26 United_Arab_Emirates
42.2 15.5   119 Bangladesh          41.4 16.6   130 Cambodia
21.2  6.7    32 China               11.7  4.9   6.1 Hong_Kong
30.5 10.2    91 India               28.6  9.4    75 Indonesia
23.5 18.1    25 Korea               31.6  5.6    24 Malaysia
36.1  8.8    68 Mongolia            39.6 14.8   128 Nepal
```

[*] These data have been compiled from the United Nations Demographic Yearbook 1990 (United Nations publications, Sales No. E/F.91.XII.1, copyright 1991, United Nations, New York) and are reproduced with the permission of the United Nations.

```
30.3  8.1 107.7 Pakistan            33.2  7.7    45 Philippines
17.8  5.2   7.5 Singapore           21.3  6.2  19.4 Sri_Lanka
22.3  7.7    28 Thailand            31.8  9.5    64 Vietnam
35.5  8.3    74 Algeria             47.2 20.2   137 Angola
48.5 11.6    67 Botswana            46.1 14.6    73 Congo
38.8  9.5  49.4 Egypt               48.6 20.7   137 Ethiopia
39.4 16.8   103 Gabon               47.4 21.4   143 Gambia
44.4 13.1    90 Ghana                 47 11.3    72 Kenya
  44  9.4    82 Libya               48.3   25   130 Malawi
35.5  9.8    82 Morocco               45 18.5   141 Mozambique
  44 12.1   135 Namibia             48.5 15.6   105 Nigeria
48.2 23.4   154 Sierra_Leone        50.1 20.2   132 Somalia
32.1  9.9    72 South_Africa        44.6 15.8   108 Sudan
46.8 12.5   118 Swaziland           31.1  7.3    52 Tunisia
52.2 15.6   103 Uganda              50.5   14   106 Tanzania
45.6 14.2    83 Zaire               51.1 13.7    80 Zambia
41.7 10.3    66 Zimbabwe
;
```

The data set Poverty contains the character variable Country and the numeric variables Birth, Death, and InfantDeath, which represent the birth rate per thousand, death rate per thousand, and infant death rate per thousand. The $20. in the INPUT statement specifies that the variable Country is a character variable with a length of 20. The double trailing at sign (@@) in the INPUT statement holds the input line for further iterations of the DATA step, specifying that observations are input from each line until all values are read.

Because the variables in the data set do not have equal variance, you must perform some form of scaling or transformation. One method is to standardize the variables to mean zero and variance one. However, when you suspect that the data contain elliptical clusters, you can use the ACECLUS procedure to transform the data such that the resulting within-cluster covariance matrix is spherical. The procedure obtains approximate estimates of the pooled within-cluster covariance matrix and then computes canonical variables to be used in subsequent analyses.

The following statements perform the ACECLUS transformation using the SAS data set Poverty. The OUT= option creates an output SAS data set called Ace to contain the canonical variable scores.

```
proc aceclus data=Poverty out=Ace p=.03 noprint;
   var Birth Death InfantDeath;
run;
```

The P= option specifies that approximately three percent of the pairs are included in the estimation of the within-cluster covariance matrix. The NOPRINT option suppresses the display of the output. The VAR statement specifies that the variables Birth, Death, and InfantDeath are used in computing the canonical variables.

The following statements invoke the CLUSTER procedure, using the SAS data set ACE created in the previous PROC ACECLUS run.

```
proc cluster data=Ace outtree=Tree method=ward
             ccc pseudo print=15;
   var can1 can2 can3 ;
   id Country;
run;
```

The OUTTREE= option creates an output SAS data set called Tree that can be used by the TREE procedure to draw a tree diagram. Ward's minimum-variance clustering method is specified by the METHOD= option. The CCC option displays the cubic clustering criterion, and the PSEUDO option displays pseudo F and t^2 statistics. Only the last 15 generations of the cluster history are displayed, as defined by the PRINT= option.

The VAR statement specifies that the canonical variables computed in the ACECLUS procedure are used in the cluster analysis. The ID statement specifies that the variable Country should be added to the Tree output data set.

The results of this analysis are displayed in the following figures.

PROC CLUSTER first displays the table of eigenvalues of the covariance matrix for the three canonical variables (Figure 23.1). The first two columns list each eigenvalue and the difference between the eigenvalue and its successor. The last two columns display the individual and cumulative proportion of variation associated with each eigenvalue.

```
                              The CLUSTER Procedure
                     Ward's Minimum Variance Cluster Analysis

                       Eigenvalues of the Covariance Matrix

                   Eigenvalue    Difference    Proportion    Cumulative

              1    64.5500051    54.7313223        0.8091        0.8091
              2     9.8186828     4.4038309        0.1231        0.9321
              3     5.4148519                      0.0679        1.0000


          Root-Mean-Square Total-Sample Standard Deviation = 5.156987
          Root-Mean-Square Distance Between Observations   = 12.63199
```

Figure 23.1. Table of Eigenvalues of the Covariance Matrix

As displayed in the last column, the first two canonical variables account for about 93% of the total variation. Figure 23.1 also displays the root mean square of the total sample standard deviation and the root mean square distance between observations.

Figure 23.2 displays the last 15 generations of the cluster history. First listed are the number of clusters and the names of the clusters joined. The observations are identified either by the ID value or by CLn, where n is the number of the cluster. Next, PROC CLUSTER displays the number of observations in the new cluster and the semipartial R^2. The latter value represents the decrease in the proportion of variance accounted for by joining the two clusters.

```
                                  The CLUSTER Procedure
                            Ward's Minimum Variance Cluster Analysis

                       Root-Mean-Square Total-Sample Standard Deviation = 5.156987
                       Root-Mean-Square Distance Between Observations   = 12.63199


                                      Cluster History
                                                                                          T
                                                                                          i
NCL    --------------Clusters Joined---------------     FREQ    SPRSQ     RSQ    ERSQ     CCC     PSF    PST2    e

 15    Oman                    CL37                        5   0.0039    .957    .933    6.03     132    12.1
 14    CL31                    CL22                       13   0.0040    .953    .928    5.81     131     9.7
 13    CL41                    CL17                       32   0.0041    .949    .922    5.70     131    13.1
 12    CL19                    CL21                       10   0.0045    .945    .916    5.65     132     6.4
 11    CL39                    CL15                        9   0.0052    .940    .909    5.60     134     6.3
 10    CL76                    CL27                        6   0.0075    .932    .900    5.25     133    18.1
  9    CL23                    CL11                       15   0.0130    .919    .890    4.20     125    12.4
  8    CL10                    Afghanistan                 7   0.0134    .906    .879    3.55     122     7.3
  7    CL9                     CL25                       17   0.0217    .884    .864    2.26     114    11.6
  6    CL8                     CL20                       14   0.0239    .860    .846    1.42     112    10.5
  5    CL14                    CL13                       45   0.0307    .829    .822    0.65     112    59.2
  4    CL16                    CL7                        28   0.0323    .797    .788    0.57     122    14.8
  3    CL12                    CL6                        24   0.0323    .765    .732    1.84     153    11.6
  2    CL3                     CL4                        52   0.1782    .587    .613    -.82     135    48.9
  1    CL5                     CL2                        97   0.5866    .000    .000    0.00       .     135
```

Figure 23.2. Cluster Generation History and R-Square Values

Next listed is the squared multiple correlation, R^2, which is the proportion of variance accounted for by the clusters. Figure 23.2 shows that, when the data are grouped into three clusters, the proportion of variance accounted for by the clusters (R^2) is about 77%. The approximate expected value of R^2 is given in the column labeled "ERSQ."

The next three columns display the values of the cubic clustering criterion (CCC), pseudo F (PSF), and t^2 (PST2) statistics. These statistics are useful in determining the number of clusters in the data.

Values of the cubic clustering criterion greater than 2 or 3 indicate good clusters; values between 0 and 2 indicate potential clusters, but they should be considered with caution; large negative values can indicate outliers. In Figure 23.2, there is a local peak of the CCC when the number of clusters is 3. The CCC drops at 4 clusters and then steadily increases, levelling off at 11 clusters.

Another method of judging the number of clusters in a data set is to look at the pseudo F statistic (PSF). Relatively large values indicate a stopping point. Reading down the PSF column, you can see that this method indicates a possible stopping point at 11 clusters and another at 3 clusters.

A general rule for interpreting the values of the pseudo t^2 statistic is to move down the column until you find the first value markedly larger than the previous value and move back up the column by one cluster. Moving down the PST2 column, you can see possible clustering levels at 11 clusters, 6 clusters, 3 clusters, and 2 clusters.

The final column in Figure 23.2 lists ties for minimum distance; a blank value indicates the absence of a tie.

These statistics indicate that the data can be clustered into 11 clusters or 3 clusters. The following statements examine the results of clustering the data into 3 clusters.

A graphical view of the clustering process can often be helpful in interpreting the clusters. The following statements use the TREE procedure to produce a tree diagram of the clusters:

```
goptions vsize=8in htext=1pct htitle=2.5pct;
axis1 order=(0 to 1 by 0.2);
proc tree data=Tree out=New nclusters=3
          graphics haxis=axis1 horizontal;
   height _rsq_;
   copy can1 can2 ;
   id country;
run;
```

The AXIS1 statement defines axis parameters that are used in the TREE procedure. The ORDER= option specifies the data values in the order in which they should appear on the axis.

The preceding statements use the SAS data set Tree as input. The OUT= option creates an output SAS data set named New to contain information on cluster membership. The NCLUSTERS= option specifies the number of clusters desired in the data set New.

The GRAPHICS option directs the procedure to use high resolution graphics. The HAXIS= option specifies AXIS1 to customize the appearance of the horizontal axis. Use this option only when the GRAPHICS option is in effect. The HORIZONTAL option orients the tree diagram horizontally. The HEIGHT statement specifies the variable _RSQ_ (R^2) as the height variable.

The COPY statement copies the canonical variables can1 and can2 (computed in the ACECLUS procedure) into the output SAS data set New. Thus, the SAS output data set New contains information for three clusters and the first two of the original canonical variables.

Figure 23.3 displays the tree diagram. The figure provides a graphical view of the information in Figure 23.2. As the number of branches grows to the left from the root, the R^2 approaches 1; the first three clusters (branches of the tree) account for over half of the variation (about 77%, from Figure 23.2). In other words, only three clusters are necessary to explain over three-fourths of the variation.

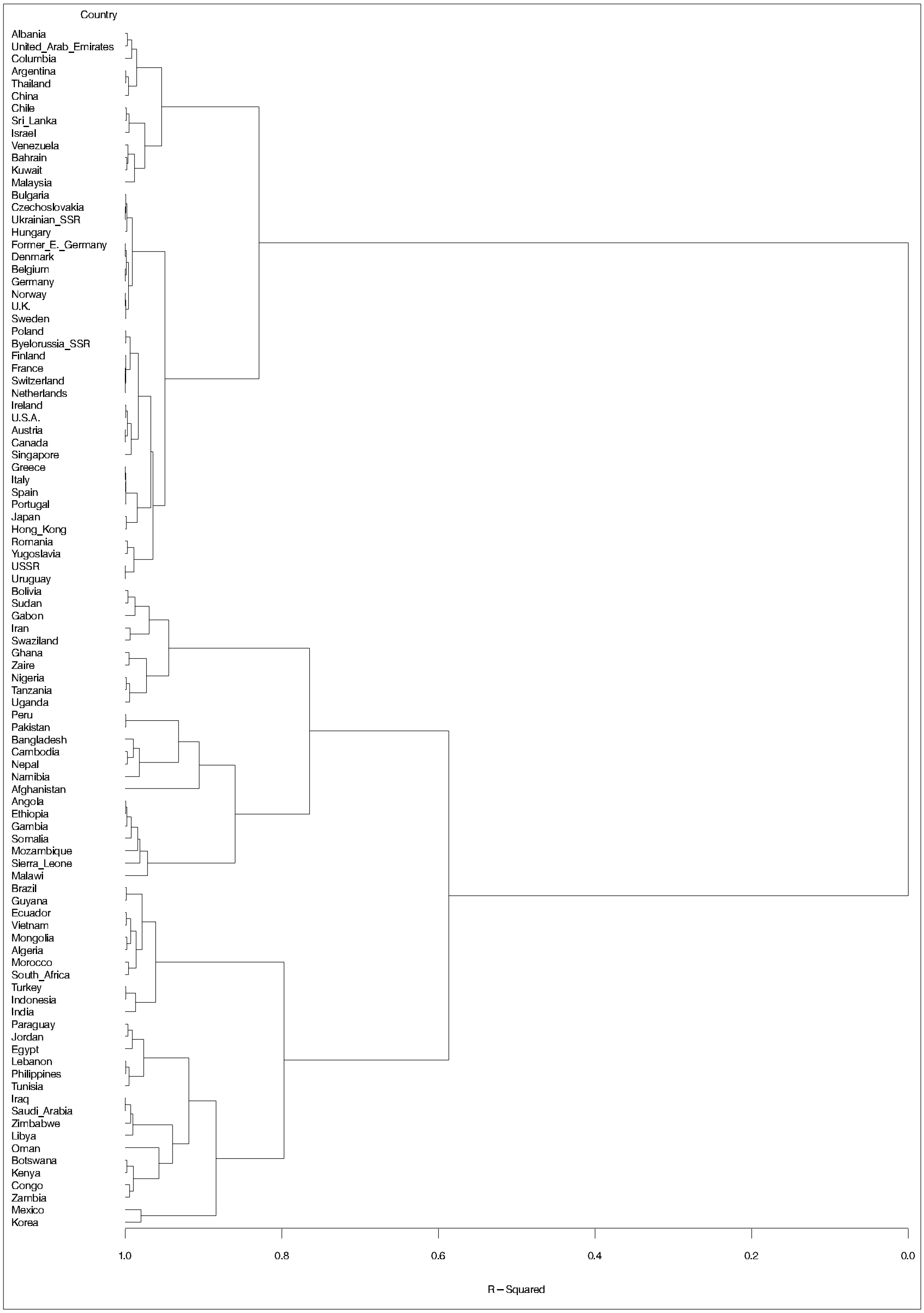


Figure 23.3. Tree Diagram of Clusters versus R-Square Values

The following statements invoke the GPLOT procedure on the SAS data set New.

```
legend1 frame cframe=ligr cborder=black
        position=center value=(justify=center);

axis1 label=(angle=90 rotate=0) minor=none order=(-10 to 20 by 5);
axis2 minor=none order=(-10 to 20 by 5);

proc gplot data=New ;
   plot can2*can1=cluster/frame cframe=ligr
                  legend=legend1 vaxis=axis1 haxis=axis2;
run;
```

The PLOT statement requests a plot of the two canonical variables, using the value of the variable cluster as the identification variable.

Figure 23.4 displays the separation of the clusters when three clusters are calculated. The plotting symbol is the cluster number.

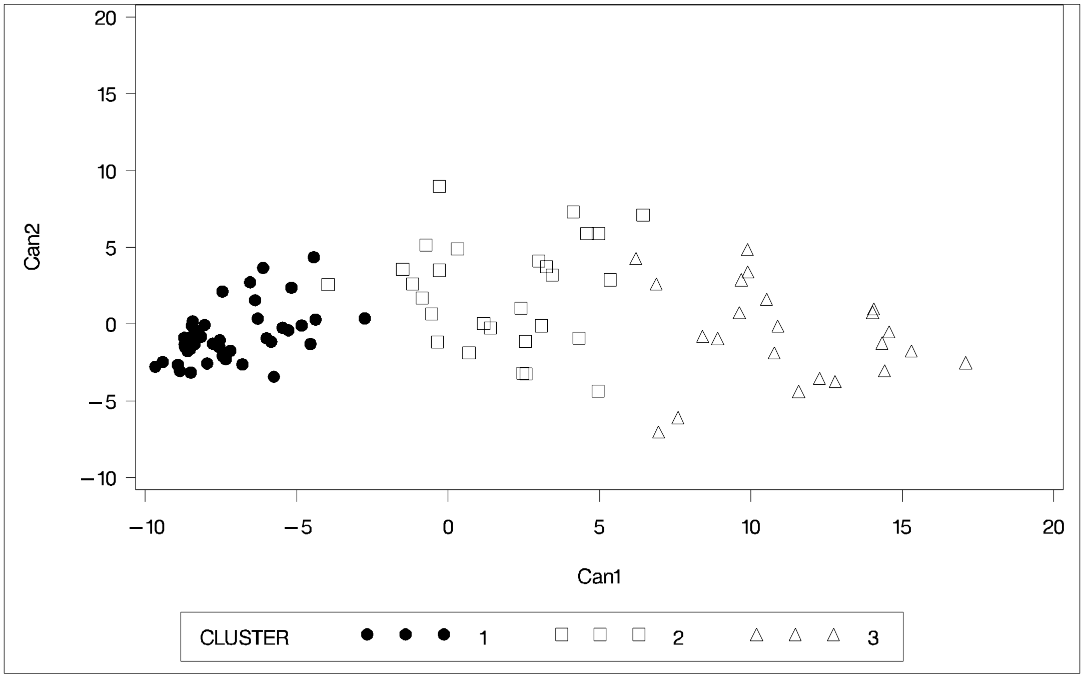


Figure 23.4. Plot of Canonical Variables and Cluster for Three Clusters

The statistics in Figure 23.2, the tree diagram in Figure 23.3, and the plot of the canonical variables assist in the determination of clusters in the data. There seems to be reasonable separation in the clusters. However, you must use this information, along with experience and knowledge of the field, to help in deciding the correct number of clusters.

Syntax

The following statements are available in the CLUSTER procedure.

PROC CLUSTER *METHOD = name < options >* **;**
BY *variables* **;**
COPY *variables* **;**
FREQ *variable* **;**
ID *variable* **;**
RMSSTD *variable* **;**
VAR *variables* **;**

Only the PROC CLUSTER statement is required, except that the FREQ statement is required when the RMSSTD statement is used; otherwise the FREQ statement is optional. Usually only the VAR statement and possibly the ID and COPY statements are needed in addition to the PROC CLUSTER statement. The rest of this section provides detailed syntax information for each of the preceding statements, beginning with the PROC CLUSTER statement. The remaining statements are covered in alphabetical order.

PROC CLUSTER Statement

PROC CLUSTER *METHOD=name < options >* **;**

The PROC CLUSTER statement starts the CLUSTER procedure, identifies a clustering method, and optionally identifies details for clustering methods, data sets, data processing, and displayed output. The METHOD= specification determines the clustering method used by the procedure. Any one of the following 11 methods can be specified for *name*:

AVERAGE | AVE requests average linkage (group average, unweighted pair-group method using arithmetic averages, UPGMA). Distance data are squared unless you specify the NOSQUARE option.

CENTROID | CEN requests the centroid method (unweighted pair-group method using centroids, UPGMC, centroid sorting, weighted-group method). Distance data are squared unless you specify the NOSQUARE option.

COMPLETE | COM requests complete linkage (furthest neighbor, maximum method, diameter method, rank order typal analysis). To reduce distortion of clusters by outliers, the TRIM= option is recommended.

DENSITY | DEN requests density linkage, which is a class of clustering methods using nonparametric probability density estimation. You must also specify one of the K=, R=, or HYBRID options to indicate the type of density estimation to be used. See also the MODE= and DIM= options in this section.

EML requests maximum-likelihood hierarchical clustering for mixtures of spherical multivariate normal distributions with equal variances but possibly unequal mixing proportions. Use METHOD=EML only with coordinate data. See the PENALTY= option on page 849. The NONORM option does not affect the reported likelihood values but does affect other unrelated criteria. The EML method is much slower than the other methods in the CLUSTER procedure.

FLEXIBLE | FLE requests the Lance-Williams flexible-beta method. See the BETA= option in this section.

MCQUITTY | MCQ requests McQuitty’s similarity analysis, which is weighted average linkage, weighted pair-group method using arithmetic averages (WPGMA).

MEDIAN | MED requests Gower’s median method, which is weighted pair-group method using centroids (WPGMC). Distance data are squared unless you specify the NOSQUARE option.

SINGLE | SIN requests single linkage (nearest neighbor, minimum method, connectedness method, elementary linkage analysis, or dendritic method). To reduce chaining, you can use the TRIM= option with METHOD=SINGLE.

TWOSTAGE | TWO requests two-stage density linkage. You must also specify the K=, R=, or HYBRID option to indicate the type of density estimation to be used. See also the MODE= and DIM= options in this section.

WARD | WAR requests Ward’s minimum-variance method (error sum of squares, trace W). Distance data are squared unless you specify the NOSQUARE option. To reduce distortion by outliers, the TRIM= option is recommended. See the NONORM option.

The following table summarizes the options in the PROC CLUSTER statement.

Tasks	Options
Specify input and output data sets	
specify input data set	DATA=
create output data set	OUTTREE=
Specify clustering methods	
specify clustering method	METHOD=
beta for flexible beta method	BETA=
minimum number of members for modal clusters	MODE=
penalty coefficient for maximum-likelihood	PENALTY=
Wong's hybrid clustering method	HYBRID
Control data processing prior to clustering	
suppress computation of eigenvalues	NOEIGEN
suppress normalizing of distances	NONORM
suppress squaring of distances	NOSQUARE
standardize variables	STANDARD
omit points with low probability densities	TRIM=
Control density estimation	
dimensionality for estimates	DIM=
number of neighbors for kth-nearest-neighbor	K=
radius of sphere of support for uniform-kernel	R=
Suppress checking for ties	NOTIE
Control display of the cluster history	
display cubic clustering criterion	CCC
suppress display of ID values	NOID
specify number of generations to display	PRINT=
display pseudo F and t^2 statistics	PSEUDO
display root-mean-square standard deviation	RMSSTD
display R^2 and semipartial R^2	RSQUARE
Control other aspects of output	
suppress display of all output	NOPRINT
display simple summary statistics	SIMPLE

The following list provides details on these options.

BETA=*n*

specifies the beta parameter for METHOD=FLEXIBLE. The value of n should be less than 1, usually between 0 and -1. By default, BETA=-0.25. Milligan (1987) suggests a somewhat smaller value, perhaps -0.5, for data with many outliers.

CCC

displays the cubic clustering criterion and approximate expected R^2 under the uniform null hypothesis (Sarle 1983). The statistics associated with the RSQUARE option, R^2 and semipartial R^2, are also displayed. The CCC option applies only to coordinate data. The CCC option is not appropriate with METHOD=SINGLE because of the method's tendency to chop off tails of distributions.

DATA=*SAS-data-set*

names the input data set containing observations to be clustered. By default, the procedure uses the most recently created SAS data set. If the data set is TYPE=DISTANCE, the data are interpreted as a distance matrix; the number of variables must equal the number of observations in the data set or in each BY group. The distances are assumed to be Euclidean, but the procedure accepts other types of distances or dissimilarities. If the data set is not TYPE=DISTANCE, the data are interpreted as coordinates in a Euclidean space, and Euclidean distances are computed. For more on TYPE=DISTANCE data sets, see Appendix A, “Special SAS Data Sets.”

You cannot use a TYPE=CORR data set as input to PROC CLUSTER, since the procedure uses dissimilarity measures. Instead, you can use a DATA step or the IML procedure to extract the correlation matrix from a TYPE=CORR data set and transform the values to dissimilarities such as $1-r$ or $1-r^2$, where r is the correlation.

All methods produce the same results when used with coordinate data as when used with Euclidean distances computed from the coordinates. However, the DIM= option must be used with distance data if you specify METHOD=TWOSTAGE or METHOD=DENSITY or if you specify the TRIM= option.

Certain methods that are most naturally defined in terms of coordinates require *squared* Euclidean distances to be used in the combinatorial distance formulas (Lance and Williams 1967). For this reason, distance data are automatically squared when used with METHOD=AVERAGE, METHOD=CENTROID, METHOD=MEDIAN, or METHOD=WARD. If you want the combinatorial formulas to be applied to the (unsquared) distances with these methods, use the NOSQUARE option.

DIM=*n*

specifies the dimensionality used when computing density estimates with the TRIM= option, METHOD=DENSITY, or METHOD=TWOSTAGE. The values of n must be greater than or equal to 1. The default is the number of variables if the data are coordinates; the default is 1 if the data are distances.

HYBRID

requests Wong’s (1982) hybrid clustering method in which density estimates are computed from a preliminary cluster analysis using the k-means method. The DATA= data set must contain means, frequencies, and root-mean-square standard deviations of the preliminary clusters (see the FREQ and RMSSTD statements). To use HYBRID, you must use either a FREQ statement or a DATA= data set that contains a _FREQ_ variable, and you must also use either an RMSSTD statement or a DATA= data set that contains a _RMSSTD_ variable.

The MEAN= data set produced by the FASTCLUS procedure is suitable for input to the CLUSTER procedure for hybrid clustering. Since this data set contains _FREQ_ and _RMSSTD_ variables, you can use it as input and then omit the FREQ and RMSSTD statements.

You must specify either METHOD=DENSITY or METHOD=TWOSTAGE with the HYBRID option. You cannot use this option in combination with the TRIM=, K=, or R= option.

K=*n*

specifies the number of neighbors to use for kth-nearest-neighbor density estimation (Silverman 1986, pp. 19–21 and 96–99). The number of neighbors (n) must be at least two but less than the number of observations. See the MODE= option, which follows.

If you request an analysis that requires density estimation (the TRIM= option, METHOD=DENSITY, or METHOD=TWOSTAGE), you must specify one of the K=, HYBRID, or R= options.

MODE=*n*

specifies that, when two clusters are joined, each must have at least n members for either cluster to be designated a modal cluster. If you specify MODE=1, each cluster must also have a maximum density greater than the fusion density for either cluster to be designated a modal cluster.

Use the MODE= option only with METHOD=DENSITY or METHOD=TWOSTAGE. With METHOD=TWOSTAGE, the MODE= option affects the number of modal clusters formed. With METHOD=DENSITY, the MODE= option does not affect the clustering process but does determine the number of modal clusters reported on the output and identified by the _MODE_ variable in the output data set.

If you specify the K= option, the default value of MODE= is the same as the value of K= because the use of kth-nearest-neighbor density estimation limits the resolution that can be obtained for clusters with fewer than k members. If you do not specify the K= option, the default is MODE=2.

If you specify MODE=0, the default value is used instead of 0.

If you specify a FREQ statement or if a _FREQ_ variable appears in the input data set, the MODE= value is compared with the number of actual observations in the clusters being joined, not with the sum of the frequencies in the clusters.

NOEIGEN

suppresses computation of eigenvalues for the cubic clustering criterion. Specifying the NOEIGEN option saves time if the number of variables is large, but it should be used only if the variables are nearly uncorrelated or if you are not interested in the cubic clustering criterion. If you specify the NOEIGEN option and the variables are highly correlated, the cubic clustering criterion may be very liberal. The NOEIGEN option applies only to coordinate data.

NOID

suppresses the display of ID values for the clusters joined at each generation of the cluster history.

NONORM

prevents the distances from being normalized to unit mean or unit root mean square with most methods. With METHOD=WARD, the NONORM option prevents the between-cluster sum of squares from being normalized by the total sum of squares to yield a squared semipartial correlation. The NONORM option does not affect the reported likelihood values with METHOD=EML, but it does affect other unrelated criteria, such as the _DIST_ variable.

NOPRINT

suppresses the display of all output. Note that this option temporarily disables the Output Delivery System (ODS). For more information, see Chapter 15, “Using the Output Delivery System.”

NOSQUARE

prevents input distances from being squared with METHOD=AVERAGE, METHOD=CENTROID, METHOD=MEDIAN, or METHOD=WARD.

If you specify the NOSQUARE option with distance data, the data are assumed to be squared Euclidean distances for computing R-squared and related statistics defined in a Euclidean coordinate system.

If you specify the NOSQUARE option with coordinate data with METHOD=CENTROID, METHOD=MEDIAN, or METHOD=WARD, then the combinatorial formula is applied to unsquared Euclidean distances. The resulting cluster distances do not have their usual Euclidean interpretation and are, therefore, labeled “False” in the output.

NOTIE

prevents PROC CLUSTER from checking for ties for minimum distance between clusters at each generation of the cluster history. If your data are measured with such sufficient precision that ties are unlikely, then you can specify the NOTIE option to reduce slightly the time and space required by the procedure. See the section “Ties” on page 865.

OUTTREE=*SAS-data-set*

creates an output data set that can be used by the TREE procedure to draw a tree diagram. You must give the data set a two-level name to save it. Refer to *SAS Language Reference: Concepts* for a discussion of permanent data sets. If you omit the OUTTREE= option, the data set is named using the DATAn convention and is not permanently saved. If you do not want to create an output data set, use OUTTREE=_NULL_.

PENALTY=*p*

specifies the penalty coefficient used with METHOD=EML. See the section “Clustering Methods” on page 854. Values for p must be greater than zero. By default, PENALTY=2.

PRINT=*n* | P=*n*

specifies the number of generations of the cluster history to display. The P= option displays the latest n generations; for example, P=5 displays the cluster history from 1 cluster through 5 clusters. The value of P= must be a nonnegative integer. The default is to display all generations. Specify PRINT=0 to suppress the cluster history.

PSEUDO

displays pseudo F and t^2 statistics. This option is effective only when the data are coordinates or when METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD. See the section “Miscellaneous Formulas” on page 861. The PSEUDO option is not appropriate with METHOD=SINGLE because of the method’s tendency to chop off tails of distributions.

R=*n*

specifies the radius of the sphere of support for uniform-kernel density estimation (Silverman 1986, pp. 11–13 and 75–94). The value of R= must be greater than zero.

If you request an analysis that requires density estimation (the TRIM= option, METHOD=DENSITY, or METHOD=TWOSTAGE), you must specify one of the K=, HYBRID, or R= options.

RMSSTD

displays the root-mean-square standard deviation of each cluster. This option is effective only when the data are coordinates or when METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD. See the section "Miscellaneous Formulas" on page 861.

RSQUARE | RSQ

displays the R^2 and semipartial R^2. This option is effective only when the data are coordinates or when METHOD=AVERAGE or METHOD=CENTROID. The R^2 and semipartial R^2 statistics are always displayed with METHOD=WARD. See the section "Miscellaneous Formulas" on page 861.

SIMPLE | S

displays means, standard deviations, skewness, kurtosis, and a coefficient of bimodality. The SIMPLE option applies only to coordinate data. See the section "Miscellaneous Formulas" on page 861.

STANDARD | STD

standardizes the variables to mean 0 and standard deviation 1. The STANDARD option applies only to coordinate data.

TRIM=*p*

omits points with low estimated probability densities from the analysis. Valid values for the TRIM= option are $0 \leq p < 100$. If $p < 1$, then p is the proportion of observations omitted. If $p \geq 1$, then p is interpreted as a percentage. A specification of TRIM=10, which trims 10 percent of the points, is a reasonable value for many data sets. Densities are estimated by the kth-nearest-neighbor or uniform-kernel methods. Trimmed points are indicated by a negative value of the _FREQ_ variable in the OUTTREE= data set.

You must use either the K= or R= option when you use TRIM=. You cannot use the HYBRID option in combination with TRIM=, so you may want to use the DIM= option instead. If you specify the STANDARD option in combination with TRIM=, the variables are standardized both before and after trimming.

The TRIM= option is useful for removing outliers and reducing chaining. Trimming is highly recommended with METHOD=WARD or METHOD=COMPLETE because clusters from these methods can be severely distorted by outliers. Trimming is also valuable with METHOD=SINGLE since single linkage is the method most susceptible to chaining. Most other methods also benefit from trimming. However, trimming is unnecessary with METHOD=TWOSTAGE or METHOD=DENSITY when kth-nearest-neighbor density estimation is used.

Use of the TRIM= option may spuriously inflate the cubic clustering criterion and the pseudo F and t^2 statistics. Trimming only outliers improves the accuracy of the statistics, but trimming saddle regions between clusters yields excessively large values.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC CLUSTER to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the CLUSTER procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables using the DATASETS procedure.

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

COPY Statement

COPY *variables* **;**

The variables in the COPY statement are copied from the input data set to the OUTTREE= data set. Observations in the OUTTREE= data set that represent clusters of more than one observation from the input data set have missing values for the COPY variables.

FREQ Statement

FREQ *variable* **;**

If one variable in the input data set represents the frequency of occurrence for other values in the observation, specify the variable's name in a FREQ statement. PROC CLUSTER then treats the data set as if each observation appeared n times, where n is the value of the FREQ variable for the observation. Noninteger values of the FREQ variable are truncated to the largest integer less than the FREQ value.

If you omit the FREQ statement but the DATA= data set contains a variable called _FREQ_, then frequencies are obtained from the _FREQ_ variable. If neither a FREQ statement nor a _FREQ_ variable is present, each observation is assumed to have a frequency of one.

If each observation in the DATA= data set represents a cluster (for example, clusters formed by PROC FASTCLUS), the variable specified in the FREQ statement should give the number of original observations in each cluster.

If you specify the RMSSTD statement, a FREQ statement is required. A FREQ statement or _FREQ_ variable is required when you specify the HYBRID option.

With most clustering methods, the same clusters are obtained from a data set with a FREQ variable as from a similar data set without a FREQ variable, if each observation is repeated as many times as the value of the FREQ variable in the first data set. The FLEXIBLE method can yield different results due to the nature of the combinatorial formula. The DENSITY and TWOSTAGE methods are also exceptions because two identical observations can be absorbed one at a time by a cluster with a higher density. If you are using a FREQ statement with either the DENSITY or TWOSTAGE method, see the MODE=option on page 848.

ID Statement

ID *variable* **;**

The values of the ID variable identify observations in the displayed cluster history and in the OUTTREE= data set. If the ID statement is omitted, each observation is denoted by OBn, where n is the observation number.

RMSSTD Statement

RMSSTD *variable* **;**

If the coordinates in the DATA= data set represent cluster means (for example, formed by the FASTCLUS procedure), you can obtain accurate statistics in the cluster histories for METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD if the data set contains

- a variable giving the number of original observations in each cluster (see the discussion of the FREQ statement earlier in this chapter)
- a variable giving the root-mean-square standard deviation of each cluster

Specify the name of the variable containing root-mean-square standard deviations in the RMSSTD statement. If you specify the RMSSTD statement, you must also specify a FREQ statement.

If you omit the RMSSTD statement but the DATA= data set contains a variable called _RMSSTD_, then root-mean-square standard deviations are obtained from the _RMSSTD_ variable.

An RMSSTD statement or _RMSSTD_ variable is required when you specify the HYBRID option.

A data set created by FASTCLUS using the MEAN= option contains _FREQ_ and _RMSSTD_ variables, so you do not have to use FREQ and RMSSTD statements when using such a data set as input to the CLUSTER procedure.

VAR Statement

VAR *variables* **;**

The VAR statement lists numeric variables to be used in the cluster analysis. If you omit the VAR statement, all numeric variables not listed in other statements are used.

Details

Clustering Methods

The following notation is used, with lowercase symbols generally pertaining to observations and uppercase symbols pertaining to clusters:

n	number of observations
v	number of variables if data are coordinates
G	number of clusters at any given level of the hierarchy
x_i or $\mathbf{x}_i$	ith observation (row vector if coordinate data)
C_K	Kth cluster, subset of $\{1, 2, \ldots, n\}$
N_K	number of observations in C_K
$\bar{\mathbf{x}}$	sample mean vector
$\bar{\mathbf{x}}_K$	mean vector for cluster C_K
$\|\mathbf{x}\|$	Euclidean length of the vector $\mathbf{x}$, that is, the square root of the sum of the squares of the elements of $\mathbf{x}$
T	$\sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|^2$
W_K	$\sum_{i \in C_k} \|\mathbf{x}_i - \bar{\mathbf{x}}_K\|^2$
P_G	$\sum W_J$, where summation is over the G clusters at the Gth level of the hierarchy
B_{KL}	$W_M - W_K - W_L$ if $C_M = C_K \cup C_L$
$d(\mathbf{x}, \mathbf{y})$	any distance or dissimilarity measure between observations or vectors $\mathbf{x}$ and $\mathbf{y}$
D_{KL}	any distance or dissimilarity measure between clusters C_K and C_L

The distance between two clusters can be defined either directly or combinatorially (Lance and Williams 1967), that is, by an equation for updating a distance matrix when two clusters are joined. In all of the following combinatorial formulas, it is assumed that clusters C_K and C_L are merged to form C_M, and the formula gives the distance between the new cluster C_M and any other cluster C_J.

For an introduction to most of the methods used in the CLUSTER procedure, refer to Massart and Kaufman (1983).

Average Linkage

The following method is obtained by specifying METHOD=AVERAGE. The distance between two clusters is defined by

$$D_{KL} = \frac{1}{N_K N_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j)$$

If $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, then

$$D_{KL} = \|\bar{\mathbf{x}}_K - \bar{\mathbf{x}}_L\|^2 + \frac{W_K}{N_K} + \frac{W_L}{N_L}$$

The combinatorial formula is

$$D_{JM} = \frac{N_K D_{JK} + N_L D_{JL}}{N_M}$$

In average linkage the distance between two clusters is the average distance between pairs of observations, one in each cluster. Average linkage tends to join clusters with small variances, and it is slightly biased toward producing clusters with the same variance.

Average linkage was originated by Sokal and Michener (1958).

Centroid Method

The following method is obtained by specifying METHOD=CENTROID. The distance between two clusters is defined by

$$D_{KL} = \|\bar{\mathbf{x}}_K - \bar{\mathbf{x}}_L\|^2$$

If $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, then the combinatorial formula is

$$D_{JM} = \frac{N_K D_{JK} + N_L D_{JL}}{N_M} - \frac{N_K N_L D_{KL}}{N_M^2}$$

In the centroid method, the distance between two clusters is defined as the (squared) Euclidean distance between their centroids or means. The centroid method is more robust to outliers than most other hierarchical methods but in other respects may not perform as well as Ward's method or average linkage (Milligan 1980).

The centroid method was originated by Sokal and Michener (1958).

Complete Linkage

The following method is obtained by specifying METHOD=COMPLETE. The distance between two clusters is defined by

$$D_{KL} = \max_{i \in C_K} \max_{j \in C_L} d(x_i, x_j)$$

The combinatorial formula is

$$D_{JM} = \max(D_{JK}, D_{JL})$$

In complete linkage, the distance between two clusters is the maximum distance between an observation in one cluster and an observation in the other cluster. Complete linkage is strongly biased toward producing clusters with roughly equal diameters, and it can be severely distorted by moderate outliers (Milligan 1980).

Complete linkage was originated by Sorensen (1948).

Density Linkage

The phrase *density linkage* is used here to refer to a class of clustering methods using nonparametric probability density estimates (for example, Hartigan 1975, pp. 205–212; Wong 1982; Wong and Lane 1983). Density linkage consists of two steps:

1. A new dissimilarity measure, d^*, based on density estimates and adjacencies is computed. If x_i and x_j are adjacent (the definition of *adjacency* depends on the method of density estimation), then $d^*(x_i, x_j)$ is the reciprocal of an estimate of the density midway between x_i and x_j; otherwise, $d^*(x_i, x_j)$ is infinite.

2. A single linkage cluster analysis is performed using d^*.

The CLUSTER procedure supports three types of density linkage: the kth-nearest-neighbor method, the uniform kernel method, and Wong's hybrid method. These are obtained by using METHOD=DENSITY and the K=, R=, and HYBRID options, respectively.

kth-Nearest Neighbor Method

The kth-nearest-neighbor method (Wong and Lane 1983) uses kth-nearest neighbor density estimates. Let $r_k(x)$ be the distance from point x to the kth-nearest observation, where k is the value specified for the K= option. Consider a closed sphere centered at x with radius $r_k(x)$. The estimated density at x, $f(x)$, is the proportion of observations within the sphere divided by the volume of the sphere. The new dissimilarity measure is computed as

$$d^*(x_i, x_j) = \begin{cases} \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) & \text{if } d(x_i, x_j) \le \max(r_k(x_i), r_k(x_j)) \\ \infty & \text{otherwise} \end{cases}$$

Wong and Lane (1983) show that kth-nearest-neighbor density linkage is strongly set consistent for high-density (density-contour) clusters if k is chosen such that $k/n \rightarrow 0$ and $k/\ln(n) \rightarrow \infty$ as $n \rightarrow \infty$. Wong and Schaack (1982) discuss methods for estimating the number of population clusters using kth-nearest-neighbor clustering.

Uniform-Kernel Method

The uniform-kernel method uses uniform-kernel density estimates. Let r be the value specified for the R= option. Consider a closed sphere centered at point x with radius r. The estimated density at x, $f(x)$, is the proportion of observations within the sphere divided by the volume of the sphere. The new dissimilarity measure is computed as

$$d^*(x_i, x_j) = \begin{cases} \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) & \text{if } d(x_i, x_j) \le r \\ \infty & \text{otherwise} \end{cases}$$

Wong's Hybrid Method

Wong's (1982) hybrid clustering method uses density estimates based on a preliminary cluster analysis by the k-means method. The preliminary clustering can be done by the FASTCLUS procedure, using the MEAN= option to create a data set containing cluster means, frequencies, and root-mean-square standard deviations. This data set is used as input to the CLUSTER procedure, and the HYBRID option is specified with METHOD=DENSITY to request the hybrid analysis. The hybrid method is appropriate for very large data sets but should not be used with small data sets, say fewer than 100 observations in the original data. The term *preliminary cluster* refers to an observation in the DATA= data set.

For preliminary cluster C_K, N_K and W_K are obtained from the input data set, as are the cluster means or the distances between the cluster means. Preliminary clusters C_K and C_L are considered adjacent if the midpoint between $\bar{x}_K$ and $\bar{\mathbf{x}}_L$ is closer to either $\bar{\mathbf{x}}_K$ or $\bar{\mathbf{x}}_L$ than to any other preliminary cluster mean or, equivalently, if $d^2(\bar{\mathbf{x}}_K, \bar{\mathbf{x}}_L) < d^2(\bar{\mathbf{x}}_K, \bar{\mathbf{x}}_M) + d^2(\bar{\mathbf{x}}_L, \bar{\mathbf{x}}_M)$ for all other preliminary clusters C_M, $M \neq K$ or L. The new dissimilarity measure is computed as

$$d^*(\bar{\mathbf{x}}_K, \bar{\mathbf{x}}_L) = \begin{cases} \frac{\left(W_K+W_L+\frac{1}{4}(N_K+N_L)d^2(\bar{\mathbf{x}}_K,\bar{\mathbf{x}}_L)\right)^{\frac{v}{2}}}{(N_K+N_L)^{1+\frac{v}{2}}} & \text{if } C_K \text{ and } C_L \text{ are adjacent} \\ \infty & \text{otherwise} \end{cases}$$

Using the K= and R= Options

The values of the K= and R= options are called *smoothing parameters*. Small values of K= or R= produce jagged density estimates and, as a consequence, many modes. Large values of K= or R= produce smoother density estimates and fewer modes. In the hybrid method, the smoothing parameter is the number of clusters in the preliminary cluster analysis. The number of modes in the final analysis tends to increase as the number of clusters in the preliminary analysis increases. Wong (1982) suggests using $n^{0.3}$ preliminary clusters, where n is the number of observations in the original data set. There is no general rule-of-thumb for selecting K= values. For all types of density linkage, you should repeat the analysis with several different values of the smoothing parameter (Wong and Schaack 1982).

There is no simple answer to the question of which smoothing parameter to use (Silverman 1986, pp. 43–61, 84–88, and 98–99). It is usually necessary to try several different smoothing parameters. A reasonable first guess for the R= option in many coordinate data sets is given by

$$\left[\frac{2^{v+2}(v+2)\Gamma(\frac{v}{2}+1)}{nv^2}\right]^{\frac{1}{v+4}} \sqrt{\sum_{l=1}^{v} s_l^2}$$

where s_l^2 is the standard deviation of the lth variable. The estimate for R= can be computed in a DATA step using the GAMMA function for Γ. This formula is derived under the assumption that the data are sampled from a multivariate normal distribution and tends, therefore, to be too large (oversmooth) if the true distribution is multimodal. Robust estimates of the standard deviations may be preferable if there are outliers. If the data are distances, the factor $\sum s_l^2$ can be replaced by an average (mean, trimmed mean, median, root-mean-square, and so on) distance divided by $\sqrt{2}$. To prevent outliers from appearing as separate clusters, you can also specify K=2, or more generally K=m, $m \geq 2$, which in most cases forces clusters to have at least m members.

If the variables all have unit variance (for example, if the STANDARD option is used), Table 23.1 can be used to obtain an initial guess for the R= option:

Table 23.1. Reasonable First Guess for the R= Option for Standardized Data

Number of	**Number of Variables**									
Observations	1	2	3	4	5	6	7	8	9	10
20	1.01	1.36	1.77	2.23	2.73	3.25	3.81	4.38	4.98	5.60
35	0.91	1.24	1.64	2.08	2.56	3.08	3.62	4.18	4.77	5.38
50	0.84	1.17	1.56	1.99	2.46	2.97	3.50	4.06	4.64	5.24
75	0.78	1.09	1.47	1.89	2.35	2.85	3.38	3.93	4.50	5.09
100	0.73	1.04	1.41	1.82	2.28	2.77	3.29	3.83	4.40	4.99
150	0.68	0.97	1.33	1.73	2.18	2.66	3.17	3.71	4.27	4.85
200	0.64	0.93	1.28	1.67	2.11	2.58	3.09	3.62	4.17	4.75
350	0.57	0.85	1.18	1.56	1.98	2.44	2.93	3.45	4.00	4.56
500	0.53	0.80	1.12	1.49	1.91	2.36	2.84	3.35	3.89	4.45
750	0.49	0.74	1.06	1.42	1.82	2.26	2.74	3.24	3.77	4.32
1000	0.46	0.71	1.01	1.37	1.77	2.20	2.67	3.16	3.69	4.23
1500	0.43	0.66	0.96	1.30	1.69	2.11	2.57	3.06	3.57	4.11
2000	0.40	0.63	0.92	1.25	1.63	2.05	2.50	2.99	3.49	4.03

Since infinite d^* values occur in density linkage, the final number of clusters can exceed one when there are wide gaps between the clusters or when the smoothing parameter results in little smoothing.

Density linkage applies no constraints to the shapes of the clusters and, unlike most other hierarchical clustering methods, is capable of recovering clusters with elongated or irregular shapes. Since density linkage employs less prior knowledge about the shape of the clusters than do methods restricted to compact clusters, density linkage is less effective at recovering compact clusters from small samples than are methods that always recover compact clusters, regardless of the data.

EML

The following method is obtained by specifying METHOD=EML. The distance between two clusters is given by

$$D_{KL} = nv \ln\left(1 + \frac{B_{KL}}{P_G}\right) - 2\left(N_M \ln(N_M) - N_K \ln(N_K) - N_L \ln(N_L)\right)$$

The EML method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions.

- multivariate normal mixture
- equal spherical covariance matrices
- unequal sampling probabilities

The EML method is similar to Ward's minimum-variance method but removes the bias toward equal-sized clusters. Practical experience has indicated that EML is somewhat biased toward unequal-sized clusters. You can specify the PENALTY= option to adjust the degree of bias. If you specify PENALTY=p, the formula is modified to

$$D_{KL} = nv \ln\left(1 + \frac{B_{KL}}{P_G}\right) - p\left(N_M \ln(N_M) - N_K \ln(N_K) - N_L \ln(N_L)\right)$$

The EML method was derived by W.S. Sarle of SAS Institute Inc. from the maximum-likelihood formula obtained by Symons (1981, p. 37, equation 8) for disjoint clustering. There are currently no other published references on the EML method.

Flexible-Beta Method

The following method is obtained by specifying METHOD=FLEXIBLE. The combinatorial formula is

$$D_{JM} = (D_{JK} + D_{JL})\frac{1-b}{2} + D_{KL}b$$

where b is the value of the BETA= option, or -0.25 by default.

The flexible-beta method was developed by Lance and Williams (1967). See also Milligan (1987).

McQuitty's Similarity Analysis

The following method is obtained by specifying METHOD=MCQUITTY. The combinatorial formula is

$$D_{JM} = \frac{D_{JK} + D_{JL}}{2}$$

The method was independently developed by Sokal and Michener (1958) and McQuitty (1966).

Median Method

The following method is obtained by specifying METHOD=MEDIAN. If $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2$, then the combinatorial formula is

$$D_{JM} = \frac{D_{JK} + D_{JL}}{2} - \frac{D_{KL}}{4}$$

The median method was developed by Gower (1967).

Single Linkage

The following method is obtained by specifying METHOD=SINGLE. The distance between two clusters is defined by

$$D_{KL} = \min_{i \in C_K} \min_{j \in C_L} d(x_i, x_j)$$

The combinatorial formula is

$$D_{JM} = \min(D_{JK}, D_{JL})$$

In single linkage, the distance between two clusters is the minimum distance between an observation in one cluster and an observation in the other cluster. Single linkage has many desirable theoretical properties (Jardine and Sibson 1971; Fisher and Van Ness 1971; Hartigan 1981) but has fared poorly in Monte Carlo studies (for example, Milligan 1980). By imposing no constraints on the shape of clusters, single linkage sacrifices performance in the recovery of compact clusters in return for the ability to detect elongated and irregular clusters. You must also recognize that single linkage tends to chop off the tails of distributions before separating the main clusters (Hartigan 1981). The notorious chaining tendency of single linkage can be alleviated by specifying the TRIM= option (Wishart 1969, pp. 296–298).

Density linkage and two-stage density linkage retain most of the virtues of single linkage while performing better with compact clusters and possessing better asymptotic properties (Wong and Lane 1983).

Single linkage was originated by Florek et al. (1951a, 1951b) and later reinvented by McQuitty (1957) and Sneath (1957).

Two-Stage Density Linkage

If you specify METHOD=DENSITY, the modal clusters often merge before all the points in the tails have clustered. The option METHOD=TWOSTAGE is a modification of density linkage that ensures that all points are assigned to modal clusters before the modal clusters are allowed to join. The CLUSTER procedure supports the same three varieties of two-stage density linkage as of ordinary density linkage: kth-nearest neighbor, uniform kernel, and hybrid.

In the first stage, disjoint modal clusters are formed. The algorithm is the same as the single linkage algorithm ordinarily used with density linkage, with one exception: two clusters are joined only if at least one of the two clusters has fewer members than the number specified by the MODE= option. At the end of the first stage, each point belongs to one modal cluster.

In the second stage, the modal clusters are hierarchically joined by single linkage. The final number of clusters can exceed one when there are wide gaps between the clusters or when the smoothing parameter is small.

Each stage forms a tree that can be plotted by the TREE procedure. By default, the TREE procedure plots the tree from the first stage. To obtain the tree for the second stage, use the option HEIGHT=MODE in the PROC TREE statement. You can also produce a single tree diagram containing both stages, with the number of clusters as the height axis, by using the option HEIGHT=N in the PROC TREE statement. To produce an output data set from PROC TREE containing the modal clusters, use _HEIGHT_ for the HEIGHT variable (the default) and specify LEVEL=0.

Two-stage density linkage was developed by W.S. Sarle of SAS Institute Inc. There are currently no other published references on two-stage density linkage.

Ward's Minimum-Variance Method

The following method is obtained by specifying METHOD=WARD. The distance between two clusters is defined by

$$D_{KL} = B_{KL} = \frac{\|\bar{\mathbf{x}}_K - \bar{\mathbf{x}}_L\|^2}{\frac{1}{N_K} + \frac{1}{N_L}}$$

If $d(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|^2$, then the combinatorial formula is

$$D_{JM} = \frac{(N_J + N_K)D_{JK} + (N_J + N_L)D_{JL} - N_J D_{KL}}{N_J + N_M}$$

In Ward's minimum-variance method, the distance between two clusters is the *ANOVA* sum of squares between the two clusters added up over all the variables. At each generation, the within-cluster sum of squares is minimized over all partitions obtainable by merging two clusters from the previous generation. The sums of squares are easier to interpret when they are divided by the total sum of squares to give proportions of variance (squared semipartial correlations).

Ward's method joins clusters to maximize the likelihood at each level of the hierarchy under the following assumptions:

- multivariate normal mixture
- equal spherical covariance matrices
- equal sampling probabilities

Ward's method tends to join clusters with a small number of observations, and it is strongly biased toward producing clusters with roughly the same number of observations. It is also very sensitive to outliers (Milligan 1980).

Ward (1963) describes a class of hierarchical clustering methods including the minimum variance method.

Miscellaneous Formulas

The root-mean-square standard deviation of a cluster C_K is

$$\text{RMSSTD} = \sqrt{\frac{W_K}{v(N_K - 1)}}$$

The R^2 statistic for a given level of the hierarchy is

$$R^2 = 1 - \frac{P_G}{T}$$

The squared semipartial correlation for joining clusters C_K and C_L is

$$\text{semipartial } R^2 = \frac{B_{KL}}{T}$$

The bimodality coefficient is

$$b = \frac{m_3^2 + 1}{m_4 + \frac{3(n-1)^2}{(n-2)(n-3)}}$$

where m_3 is skewness and m_4 is kurtosis. Values of b greater than 0.555 (the value for a uniform population) may indicate bimodal or multimodal marginal distributions. The maximum of 1.0 (obtained for the Bernoulli distribution) is obtained for a population with only two distinct values. Very heavy-tailed distributions have small values of b regardless of the number of modes.

Formulas for the cubic-clustering criterion and approximate expected R^2 are given in Sarle (1983).

The pseudo F statistic for a given level is

$$\text{pseudo } F = \frac{\frac{T-P_G}{G-1}}{\frac{P_G}{n-G}}$$

The pseudo t^2 statistic for joining C_K and C_L is

$$\text{pseudo } t^2 = \frac{B_{KL}}{\frac{W_K+W_L}{N_K+N_L-2}}$$

The pseudo F and t^2 statistics may be useful indicators of the number of clusters, but they are *not* distributed as F and t^2 random variables. If the data are independently sampled from a multivariate normal distribution with a scalar covariance matrix and if the clustering method allocates observations to clusters randomly (which no clustering method actually does), then the pseudo F statistic is distributed as an F random variable with $v(G-1)$ and $v(n-G)$ degrees of freedom. Under the same assumptions, the pseudo t^2 statistic is distributed as an F random variable with v and $v(N_K+N_L-2)$ degrees of freedom. The pseudo t^2 statistic differs computationally from Hotelling's T^2 in that the latter uses a general symmetric covariance matrix instead of a scalar covariance matrix. The pseudo F statistic was suggested by Calinski and Harabasz (1974). The pseudo t^2 statistic is related to the $J_e(2)/J_e(1)$ statistic of Duda and Hart (1973) by

$$\frac{J_e(2)}{J_e(1)} = \frac{W_K+W_L}{W_M} = \frac{1}{1+\frac{t^2}{N_K+N_L-2}}$$

See Milligan and Cooper (1985) and Cooper and Milligan (1988) regarding the performance of these statistics in estimating the number of population clusters. Conservative tests for the number of clusters using the pseudo F and t^2 statistics can be obtained by the Bonferroni approach (Hawkins, Muller, and ten Krooden 1982, pp. 337–340).

Ultrametrics

A dissimilarity measure $d(x, y)$ is called an *ultrametric* if it satisfies the following conditions:

- $d(x, x) = 0$ for all x
- $d(x, y) \geq 0$ for all x, y
- $d(x, y) = d(y, x)$ for all x, y
- $d(x, y) \leq \max\left(d(x, z), d(y, z)\right)$ for all x, y, and z

Any hierarchical clustering method induces a dissimilarity measure on the observations, say $h(x_i, x_j)$. Let C_M be the cluster with the fewest members that contains both x_i and x_j. Assume C_M was formed by joining C_K and C_L. Then define $h(x_i, x_j) = D_{KL}$.

If the fusion of C_K and C_L reduces the number of clusters from g to $g-1$, then define $D_{(g)} = D_{KL}$. Johnson (1967) shows that if

$$0 \leq D_{(n)} \leq D_{(n-1)} \leq \cdots \leq D_{(2)}$$

then $h(\cdot, \cdot)$ is an ultrametric. A method that always satisfies this condition is said to be a *monotonic* or *ultrametric clustering method*. All methods implemented in PROC CLUSTER except CENTROID, EML, and MEDIAN are ultrametric (Milligan 1979; Batagelj 1981).

Algorithms

Anderberg (1973) describes three algorithms for implementing agglomerative hierarchical clustering: stored data, stored distance, and sorted distance. The algorithms used by PROC CLUSTER for each method are indicated in Table 23.2. For METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD, either the stored data or the stored distance algorithm can be used. For these methods, if the data are distances or if you specify the NOSQUARE option, the stored distance algorithm is used; otherwise, the stored data algorithm is used.

Table 23.2. Three Algorithms for Implementing Agglomerative Hierarchical Clustering

	Algorithm		
Stored Method	**Stored Data**	**Stored Distance**	**Sorted Distance**
AVERAGE	x	x	
CENTROID	x	x	
COMPLETE		x	
DENSITY			x
EML	x		
FLEXIBLE		x	
MCQUITTY		x	
MEDIAN		x	
SINGLE		x	
TWOSTAGE			x
WARD	x	x	

Computational Resources

The CLUSTER procedure stores the data (including the COPY and ID variables) in memory or, if necessary, on disk. If eigenvalues are computed, the covariance matrix is stored in memory. If the stored distance or sorted distance algorithm is used, the distances are stored in memory or, if necessary, on disk.

With coordinate data, the increase in CPU time is roughly proportional to the number of variables. The VAR statement should list the variables in order of decreasing variance for greatest efficiency.

For both coordinate and distance data, the dominant factor determining CPU time is the number of observations. For density methods with coordinate data, the asymptotic time requirements are somewhere between $n \ln(n)$ and n^2, depending on how the smoothing parameter increases. For other methods except EML, time is roughly proportional to n^2. For the EML method, time is roughly proportional to n^3.

PROC CLUSTER runs much faster if the data can be stored in memory and, if the stored distance algorithm is used, the distance matrix can be stored in memory as well. To estimate the bytes of memory needed for the data, use the following equation and round up to the nearest multiple of d.

$$\begin{array}{lcll}
n(vd & + & 8d\ +\ i & \\
 & + & i & \text{if density estimation or the} \\
 & & & \text{sorted distance algorithm used} \\
 & + & 3d & \text{if stored data algorithm used} \\
 & + & 3d & \text{if density estimation used} \\
 & + & \text{max(8, length of ID variable)} & \text{if ID variable used} \\
 & + & \text{length of ID variable} & \text{if ID variable used} \\
 & + & \text{sum of lengths of COPY variables)} & \text{if COPY variables used}
\end{array}$$

where

n is the number of observations

v is the number of variables

d is the size of a C variable of type *double*. For most computers, $d = 8$.

i is the size of a C variable of type *int*. For most computers, $i = 4$.

The number of bytes needed for the distance matrix is $dn(n + 1)/2$.

Missing Values

If the data are coordinates, observations with missing values are excluded from the analysis. If the data are distances, missing values are not allowed in the lower triangle of the distance matrix. The upper triangle is ignored. For more on TYPE=DISTANCE data sets, see Appendix A, “Special SAS Data Sets.”

Ties

At each level of the clustering algorithm, PROC CLUSTER must identify the pair of clusters with the minimum distance. Sometimes, usually when the data are discrete, there may be two or more pairs with the same minimum distance. In such cases the tie must be broken in some arbitrary way. If there are ties, then the results of the cluster analysis depend on the order of the observations in the data set. The presence of ties is reported in the SAS log and in the column of the cluster history labeled “Tie” unless the NOTIE option is specified.

PROC CLUSTER breaks ties as follows. Each cluster is identified by the smallest observation number among its members. For each pair of clusters, there is a smaller identification number and a larger identification number. If two or more pairs of clusters are tied for minimum distance between clusters, the pair that has the minimum larger identification number is merged. If there is a tie for minimum larger identification number, the pair that has the minimum smaller identification number is merged. This method for breaking ties is different from that used in Version 5. The change in the algorithm may produce changes in the resulting clusters.

A tie means that the level in the cluster history at which the tie occurred and possibly some of the subsequent levels are not uniquely determined. Ties that occur early in the cluster history usually have little effect on the later stages. Ties that occur in the middle part of the cluster history are cause for further investigation. Ties late in the cluster history indicate important indeterminacies.

The importance of ties can be assessed by repeating the cluster analysis for several different random permutations of the observations. The discrepancies at a given level can be examined by crosstabulating the clusters obtained at that level for all of the permutations. See Example 23.4 for details.

Size, Shape, and Correlation

In some biological applications, the organisms that are being clustered may be at different stages of growth. Unless it is the growth process itself that is being studied, differences in size among such organisms are not of interest. Therefore, distances among organisms should be computed in such a way as to control for differences in size while retaining information about differences in shape.

If coordinate data are measured on an interval scale, you can control for size by subtracting a measure of the overall size of each observation from each datum. For example, if no other direct measure of size is available, you could subtract the mean of each row of the data matrix, producing a row-centered coordinate matrix. An easy way to subtract the mean of each row is to use PROC STANDARD on the transposed coordinate matrix:

```
proc transpose data= coordinate-datatype ;
proc standard m=0;
proc transpose out=row-centered-coordinate-data;
```

Another way to remove size effects from interval-scale coordinate data is to do a principal component analysis and discard the first component (Blackith and Reyment 1971).

If the data are measured on a ratio scale, you can control for size by dividing each datum by a measure of overall size; in this case, the geometric mean is a more natural measure of size than the arithmetic mean. However, it is often more meaningful to analyze the logarithms of ratio-scaled data, in which case you can subtract the arithmetic mean after taking logarithms. You must also consider the dimensions of measurement. For example, if you have measures of both length and weight, you may need to cube the measures of length or take the cube root of the weights. Various other complications may also arise in real applications, such as different growth rates for different parts of the body (Sneath and Sokal 1973).

Issues of size and shape are pertinent to many areas besides biology (for example, Hamer and Cunningham 1981). Suppose you have data consisting of subjective ratings made by several different raters. Some raters may tend to give higher overall ratings than other raters. Some raters may also tend to spread out their ratings over more of the scale than do other raters. If it is impossible for you to adjust directly for rater differences, then distances should be computed in such a way as to control for both differences in size and variability. For example, if the data are considered to be measured on an interval scale, you can subtract the mean of each observation and divide by the standard deviation, producing a row-standardized coordinate matrix. With some clustering methods, analyzing squared Euclidean distances from a row-standardized coordinate matrix is equivalent to analyzing the matrix of correlations among rows, since squared Euclidean distance is an affine transformation of the correlation (Hartigan 1975, p. 64).

If you do an analysis of row-centered or row-standardized data, you need to consider whether the columns (variables) should be standardized before centering or standardizing the rows, after centering or standardizing the rows, or both before and after. If you standardize the columns after standardizing the rows, then strictly speaking you are not analyzing shape because the profiles are distorted by standardizing the columns; however, this type of double standardization may be necessary in practice to get reasonable results. It is not clear whether iterating the standardization of rows and columns may be of any benefit.

The choice of distance or correlation measure should depend on the meaning of the data and the purpose of the analysis. Simulation studies that compare distance and correlation measures are useless unless the data are generated to mimic data from your field of application; conclusions drawn from artificial data cannot be generalized because it is possible to generate data such that distances that include size effects work better or such that correlations work better.

You can standardize the rows of a data set by using a DATA step or by using the TRANSPOSE and STANDARD procedures. You can also use PROC TRANSPOSE and then have PROC CORR create a TYPE=CORR data set containing a correlation matrix. If you want to analyze a TYPE=CORR data set with PROC CLUSTER, you must use a DATA step to perform the following steps:

1. Set the data set TYPE= to DISTANCE.
2. Convert the correlations to dissimilarities by computing $1 - r$, $\sqrt{1 - r}$, $1 - r^2$, or some other decreasing function.
3. Delete observations for which the variable _TYPE_ does not have the value 'CORR'.

See Example 23.6 for an analysis of a data set in which size information is detrimental to the classification.

Output Data Set

The OUTTREE= data set contains one observation for each observation in the input data set, plus one observation for each cluster of two or more observations (that is, one observation for each node of the cluster tree). The total number of output observations is usually $2n - 1$, where n is the number of input observations. The density methods may produce fewer output observations when the number of clusters cannot be reduced to one.

The label of the OUTTREE= data set identifies the type of cluster analysis performed and is automatically displayed when the TREE procedure is invoked.

The variables in the OUTTREE= data set are as follows:

- the BY variables, if you use a BY statement
- the ID variable, if you use an ID statement
- the COPY variables, if you use a COPY statement
- _NAME_, a character variable giving the name of the node. If the node is a cluster, the name is CLn, where n is the number of the cluster. If the node is an observation, the name is OBn, where n is the observation number. If the node is an observation and the ID statement is used, the name is the formatted value of the ID variable.
- _PARENT_, a character variable giving the value of _NAME_ of the parent of the node
- _NCL_, the number of clusters
- _FREQ_, the number of observations in the current cluster
- _HEIGHT_, the distance or similarity between the last clusters joined, as defined in the section "Clustering Methods" on page 854. The variable _HEIGHT_ is used by the TREE procedure as the default height axis. The label of the _HEIGHT_ variable identifies the between-cluster distance measure. For METHOD=TWOSTAGE, the _HEIGHT_ variable contains the densities at which clusters joined in the first stage; for clusters formed in the second stage, _HEIGHT_ is a very small negative number.

If the input data set contains coordinates, the following variables appear in the output data set:

- the variables containing the coordinates used in the cluster analysis. For output observations that correspond to input observations, the values of the coordinates are the same in both data sets except for some slight numeric error possibly introduced by standardizing and unstandardizing if the STANDARD option is used. For output observations that correspond to clusters of more than one input observation, the values of the coordinates are the cluster means.
- _ERSQ_, the approximate expected value of R^2 under the uniform null hypothesis
- _RATIO_, equal to $\frac{1-\text{\_ERSQ\_}}{1-\text{\_RSQ\_}}$
- _LOGR_, natural logarithm of _RATIO_
- _CCC_, the cubic clustering criterion

The variables _ERSQ_, _RATIO_, _LOGR_, and _CCC_ have missing values when the number of clusters is greater than one-fifth the number of observations.

If the input data set contains coordinates and METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD, then the following variables appear in the output data set.

- _DIST_, the Euclidean distance between the means of the last clusters joined
- _AVLINK_, the average distance between the last clusters joined

If the input data set contains coordinates or METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD, then the following variables appear in the output data set:

- _RMSSTD_, the root-mean-square standard deviation of the current cluster
- _SPRSQ_, the semipartial squared multiple correlation or the decrease in the proportion of variance accounted for due to joining two clusters to form the current cluster
- _RSQ_, the squared multiple correlation
- _PSF_, the pseudo F statistic
- _PST2_, the pseudo t^2 statistic

If METHOD=EML, then the following variable appears in the output data set:

- _LNLR_, the log-likelihood ratio

If METHOD=TWOSTAGE or METHOD=DENSITY, the following variable appears in the output data set:

- _MODE_, pertaining to the modal clusters. With METHOD=DENSITY, the _MODE_ variable indicates the number of modal clusters contained by the current cluster. With METHOD=TWOSTAGE, the _MODE_ variable gives the maximum density in each modal cluster and the fusion density, d^*, for clusters containing two or more modal clusters; for clusters containing no modal clusters, _MODE_ is missing.

If nonparametric density estimates are requested (when METHOD=DENSITY or METHOD=TWOSTAGE and the HYBRID option is not used; or when the TRIM= option is used), the output data set contains

- _DENS_, the maximum density in the current cluster

Displayed Output

If you specify the SIMPLE option and the data are coordinates, PROC CLUSTER produces simple descriptive statistics for each variable:

- the Mean
- the standard deviation, Std Dev
- the Skewness
- the Kurtosis
- a coefficient of Bimodality

If the data are coordinates and you do not specify the NOEIGEN option, PROC CLUSTER displays

- the Eigenvalues of the Correlation or Covariance Matrix
- the Difference between successive eigenvalues
- the Proportion of variance explained by each eigenvalue
- the Cumulative proportion of variance explained

If the data are coordinates, PROC CLUSTER displays the Root-Mean-Square Total-Sample Standard Deviation of the variables

If the distances are normalized, PROC CLUSTER displays one of the following, depending on whether squared or unsquared distances are used:

- the Root-Mean-Square Distance Between Observations
- the Mean Distance Between Observations

For the generations in the clustering process specified by the PRINT= option, PROC CLUSTER displays

- the Number of Clusters or NCL
- the names of the Clusters Joined. The observations are identified by the formatted value of the ID variable, if any; otherwise, the observations are identified by OBn, where n is the observation number. The CLUSTER procedure displays the entire value of the ID variable in the cluster history instead of truncating at 16 characters. Long ID values may be flowed onto several lines. Clusters of two or more observations are identified as CLn, where n is the number of clusters existing after the cluster in question is formed.
- the number of observations in the new cluster, Frequency of New Cluster or FREQ

If you specify the RMSSTD option and if the data are coordinates or if you specify METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD, then PROC CLUSTER displays the root-mean-square standard deviation of the new cluster, RMS Std of New Cluster or RMS Std.

PROC CLUSTER displays the following items if you specify METHOD=WARD. It also displays them if you specify the RSQUARE option and either the data are coordinates or you specify METHOD=AVERAGE or METHOD=CENTROID:

- the decrease in the proportion of variance accounted for resulting from joining the two clusters, Semipartial R-Squared or SPRSQ. This equals the between-cluster sum of squares divided by the corrected total sum of squares.
- the squared multiple correlation, R-Squared or RSQ. R^2 is the proportion of variance accounted for by the clusters.

If you specify the CCC option and the data are coordinates, PROC CLUSTER displays

- Approximate Expected R-Squared or ERSQ, the approximate expected value of R^2 under the uniform null hypothesis
- the Cubic Clustering Criterion or CCC. The cubic clustering criterion and approximate expected R^2 are given missing values when the number of clusters is greater than one-fifth the number of observations.

If you specify the PSEUDO option and if the data are coordinates or METHOD=AVERAGE, METHOD=CENTROID, or METHOD=WARD, then PROC CLUSTER displays

- Pseudo F or PSF, the pseudo F statistic measuring the separation among all the clusters at the current level
- Pseudo t^2 or PST2, the pseudo t^2 statistic measuring the separation between the two clusters most recently joined

If you specify the NOSQUARE option and METHOD=AVERAGE, PROC CLUSTER displays the (Normalized) Average Distance or (Norm) Aver Dist, the average distance between pairs of objects in the two clusters joined with one object from each cluster.

If you do not specify the NOSQUARE option and METHOD=AVERAGE, PROC CLUSTER displays the (Normalized) RMS Distance or (Norm) RMS Dist, the root-mean-square distance between pairs of objects in the two clusters joined with one object from each cluster.

If METHOD=CENTROID, PROC CLUSTER displays the (Normalized) Centroid Distance or (Norm) Cent Dist, the distance between the two cluster centroids.

If METHOD=COMPLETE, PROC CLUSTER displays the (Normalized) Maximum Distance or (Norm) Max Dist, the maximum distance between the two clusters.

If METHOD=DENSITY or METHOD=TWOSTAGE, PROC CLUSTER displays

- Normalized Fusion Density or Normalized Fusion Dens, the value of d^* as defined in the section “Clustering Methods” on page 854
- the Normalized Maximum Density in Each Cluster joined, including the Lesser or Min, and the Greater or Max, of the two maximum density values

If METHOD=EML, PROC CLUSTER displays

- Log Likelihood Ratio or LNLR
- Log Likelihood or LNLIKE

If METHOD=FLEXIBLE, PROC CLUSTER displays the (Normalized) Flexible Distance or (Norm) Flex Dist, the distance between the two clusters based on the Lance-Williams flexible formula.

If METHOD=MEDIAN, PROC CLUSTER displays the (Normalized) Median Distance or (Norm) Med Dist, the distance between the two clusters based on the median method.

If METHOD=MCQUITTY, PROC CLUSTER displays the (Normalized) McQuitty’s Similarity or (Norm) MCQ, the distance between the two clusters based on McQuitty’s similarity method.

If METHOD=SINGLE, PROC CLUSTER displays the (Normalized) Minimum Distance or (Norm) Min Dist, the minimum distance between the two clusters.

If you specify the NONORM option and METHOD=WARD, PROC CLUSTER displays the Between-Cluster Sum of Squares or BSS, the *ANOVA* sum of squares between the two clusters joined.

If you specify neither the NOTIE option nor METHOD=TWOSTAGE or METHOD=DENSITY, PROC CLUSTER displays Tie, where a T in the column indicates a tie for minimum distance and a blank indicates the absence of a tie.

After the cluster history, if METHOD=TWOSTAGE or METHOD=DENSITY, PROC CLUSTER displays the number of modal clusters.

ODS Table Names

PROC CLUSTER assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, “Using the Output Delivery System.”

Table 23.3. ODS Tables Produced in PROC CLUSTER

ODS Table Name	Description	Statement	Option
ClusterHistory	Obs or clusters joined, frequencies and other cluster statistics	PROC	default
SimpleStatistics	Simple statistics, before or after trimming	PROC	SIMPLE
EigenvalueTable	Eigenvalues of the CORR or COV matrix	PROC	default

Examples

Example 23.1. Cluster Analysis of Flying Mileages between Ten American Cities

This first example clusters ten American cities based on the flying mileages between them. Six clustering methods are shown with corresponding tree diagrams produced by the TREE procedure. The EML method cannot be used because it requires coordinate data. The other omitted methods produce the same clusters, although not the same distances between clusters, as one of the illustrated methods: complete linkage and the flexible-beta method yield the same clusters as Ward's method, McQuitty's similarity analysis produces the same clusters as average linkage, and the median method corresponds to the centroid method.

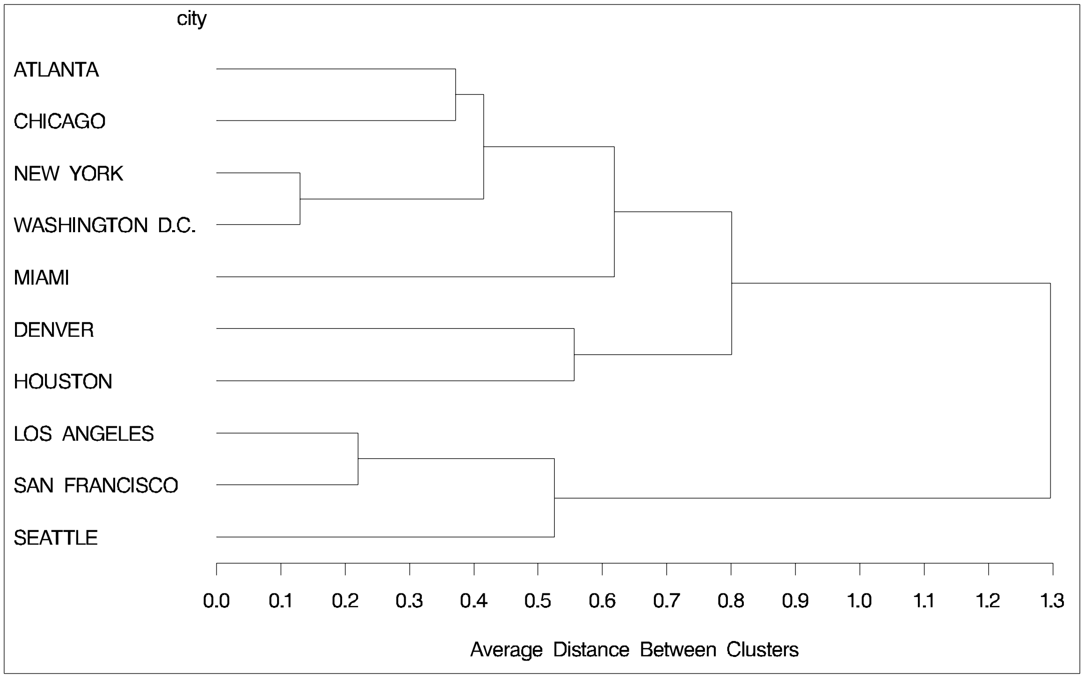
All of the methods suggest a division of the cities into two clusters along the east-west dimension. There is disagreement, however, about which cluster Denver should belong to. Some of the methods indicate a possible third cluster containing Denver and Houston. The following statements produce Output 23.1.1:

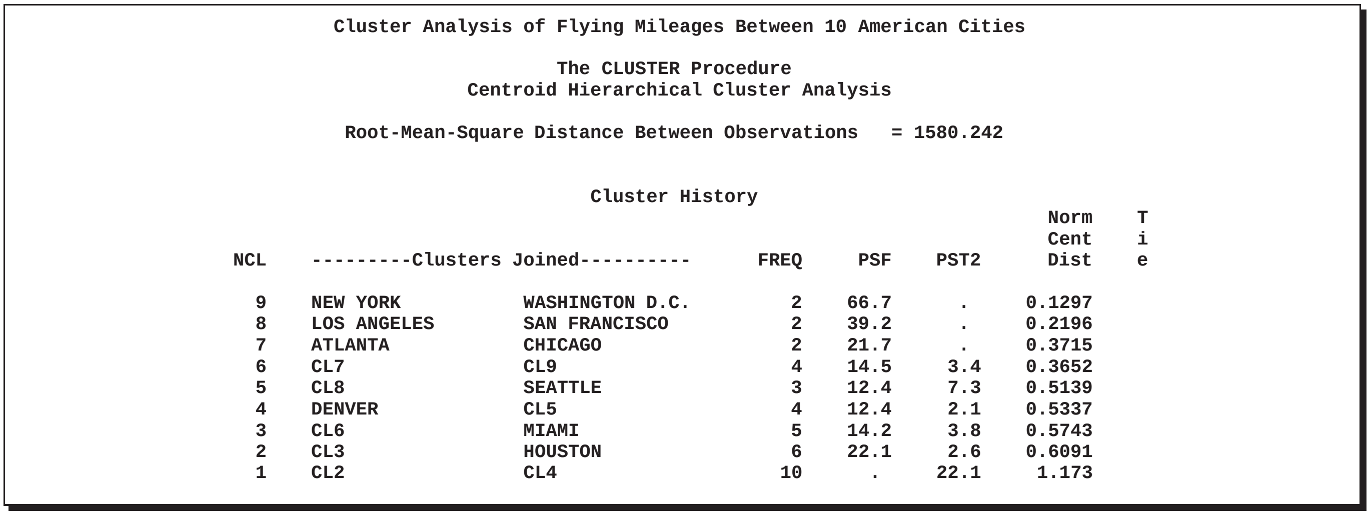
```
title 'Cluster Analysis of Flying Mileages Between 10 American Cities';
data mileages(type=distance);
   input (atlanta chicago denver houston losangeles
         miami newyork sanfran seattle washdc) (5.)
         @55 city $15.;
   datalines;
    0                                                 ATLANTA
  587    0                                            CHICAGO
 1212  920    0                                       DENVER
  701  940  879    0                                  HOUSTON
 1936 1745  831 1374    0                             LOS ANGELES
  604 1188 1726  968 2339    0                        MIAMI
  748  713 1631 1420 2451 1092    0                   NEW YORK
 2139 1858  949 1645  347 2594 2571    0              SAN FRANCISCO
 2182 1737 1021 1891  959 2734 2408  678    0         SEATTLE
  543  597 1494 1220 2300  923  205 2442 2329    0    WASHINGTON D.C.
;
```

```
/*---------------------- Average linkage --------------------*/
 proc cluster data=mileages method=average pseudo;
   id city;
run;

proc tree horizontal spaces=2;
   id city;
run;

/*---------------------- Centroid method --------------------*/
proc cluster data=mileages method=centroid pseudo;
   id city;
run;

proc tree horizontal spaces=2;
   id city;
run;

/*-------- Density linkage with 3rd-nearest-neighbor --------*/
proc cluster data=mileages method=density k=3;
   id city;
run;

proc tree horizontal spaces=2;
   id city;
run;

/*--------------------- Single linkage ----------------------*/
proc cluster data=mileages method=single;
   id city;
run;

proc tree horizontal spaces=2;
   id city;
run;

/*--- Two-stage density linkage with 3rd-nearest-neighbor ---*/
proc cluster data=mileages method=twostage k=3;
   id city;
run;

proc tree horizontal spaces=2;
   id city;
run;

/* Ward’s minimum variance with pseudo $F$ and $t^2$ statistics */
proc cluster data=mileages method=ward pseudo;
   id city;
run;

proc tree horizontal spaces=2;
   id city;
run;
```

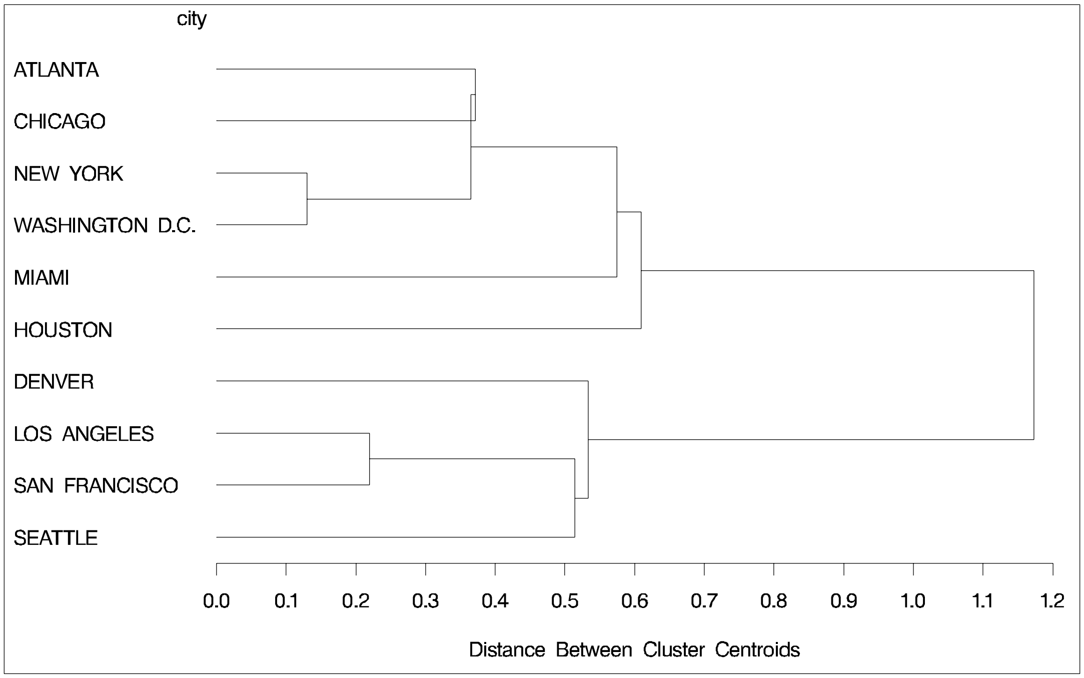
Output 23.1.1. Statistics and Tree Diagrams for Six Different Clustering Methods

```
                      Cluster Analysis of Flying Mileages Between 10 American Cities

                                          The CLUSTER Procedure
                                     Average Linkage Cluster Analysis

                        Root-Mean-Square Distance Between Observations   = 1580.242


                                             Cluster History
                                                                                  Norm    T
                                                                                   RMS    i
            NCL    ---------Clusters Joined----------      FREQ     PSF    PST2      Dist    e

              9    NEW YORK           WASHINGTON D.C.         2    66.7       .    0.1297
              8    LOS ANGELES        SAN FRANCISCO           2    39.2       .    0.2196
              7    ATLANTA            CHICAGO                 2    21.7       .    0.3715
              6    CL7                CL9                     4    14.5     3.4    0.4149
              5    CL8                SEATTLE                 3    12.4     7.3    0.5255
              4    DENVER             HOUSTON                 2    13.9       .    0.5562
              3    CL6                MIAMI                   5    15.5     3.8    0.6185
              2    CL3                CL4                     7    16.0     5.3    0.8005
              1    CL2                CL5                    10       .    16.0    1.2967
```

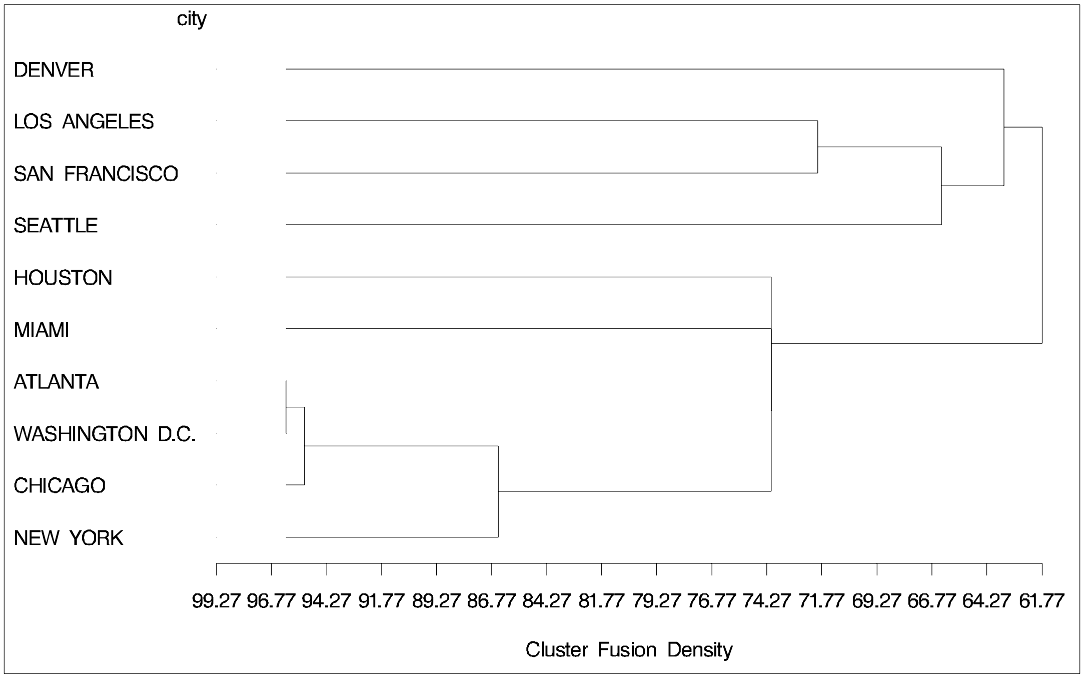




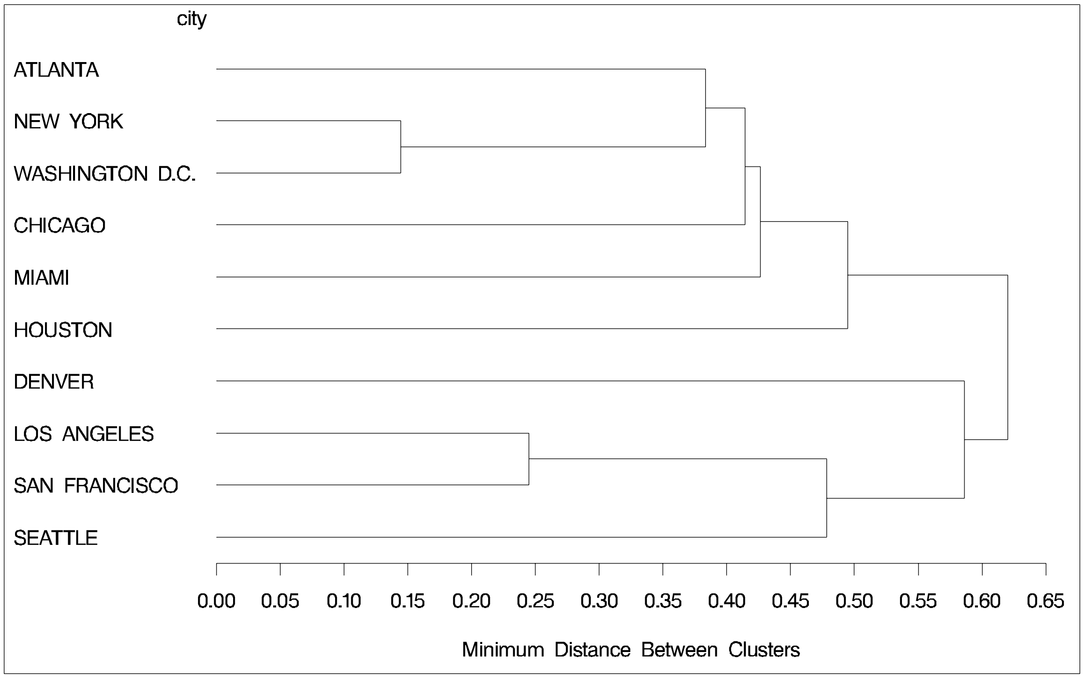
```
                     Cluster Analysis of Flying Mileages Between 10 American Cities

                                         The CLUSTER Procedure
                                  Centroid Hierarchical Cluster Analysis

                      Root-Mean-Square Distance Between Observations   = 1580.242


                                            Cluster History
                                                                                   Norm    T
                                                                                   Cent    i
          NCL    ---------Clusters Joined----------      FREQ     PSF    PST2      Dist    e

            9    NEW YORK          WASHINGTON D.C.          2    66.7       .    0.1297
            8    LOS ANGELES       SAN FRANCISCO            2    39.2       .    0.2196
            7    ATLANTA           CHICAGO                  2    21.7       .    0.3715
            6    CL7               CL9                      4    14.5     3.4    0.3652
            5    CL8               SEATTLE                  3    12.4     7.3    0.5139
            4    DENVER            CL5                      4    12.4     2.1    0.5337
            3    CL6               MIAMI                    5    14.2     3.8    0.5743
            2    CL3               HOUSTON                  6    22.1     2.6    0.6091
            1    CL2               CL4                     10       .    22.1     1.173
```



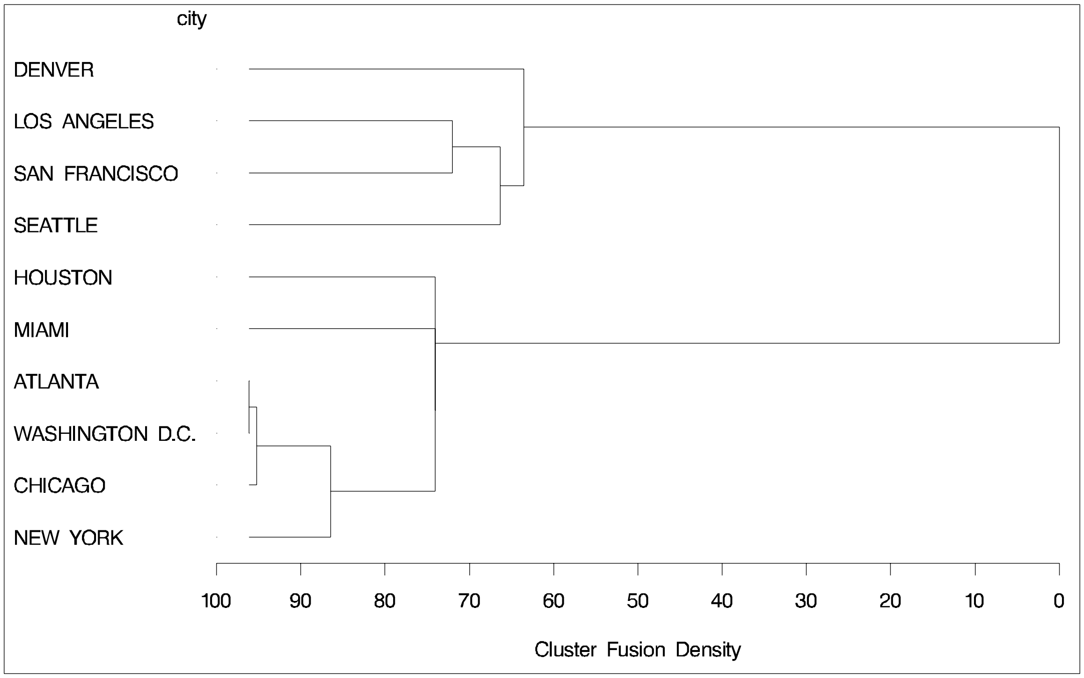
```
                  Cluster Analysis of Flying Mileages Between 10 American Cities

                                    The CLUSTER Procedure
                                Density Linkage Cluster Analysis

                                             K = 3


                                        Cluster History
                                                      Normalized           Maximum Density      T
                                                          Fusion           in Each Cluster      i
  NCL    ---------Clusters Joined----------      FREQ    Density           Lesser     Greater    e

    9    ATLANTA            WASHINGTON D.C.         2     96.106          92.5043       100.0
    8    CL9                CHICAGO                 3     95.263          90.9548       100.0
    7    CL8                NEW YORK                4     86.465          76.1571       100.0
    6    CL7                MIAMI                   5     74.079          58.8299       100.0    T
    5    CL6                HOUSTON                 6     74.079          61.7747       100.0
    4    LOS ANGELES        SAN FRANCISCO           2     71.968          65.3430     80.0885
    3    CL4                SEATTLE                 3     66.341          56.6215     80.0885
    2    CL3                DENVER                  4     63.509          61.7747     80.0885
    1    CL5                CL2                    10     61.775    *     80.0885       100.0

                      * indicates fusion of two modal or multimodal clusters
                                2 modal clusters have been formed.
```



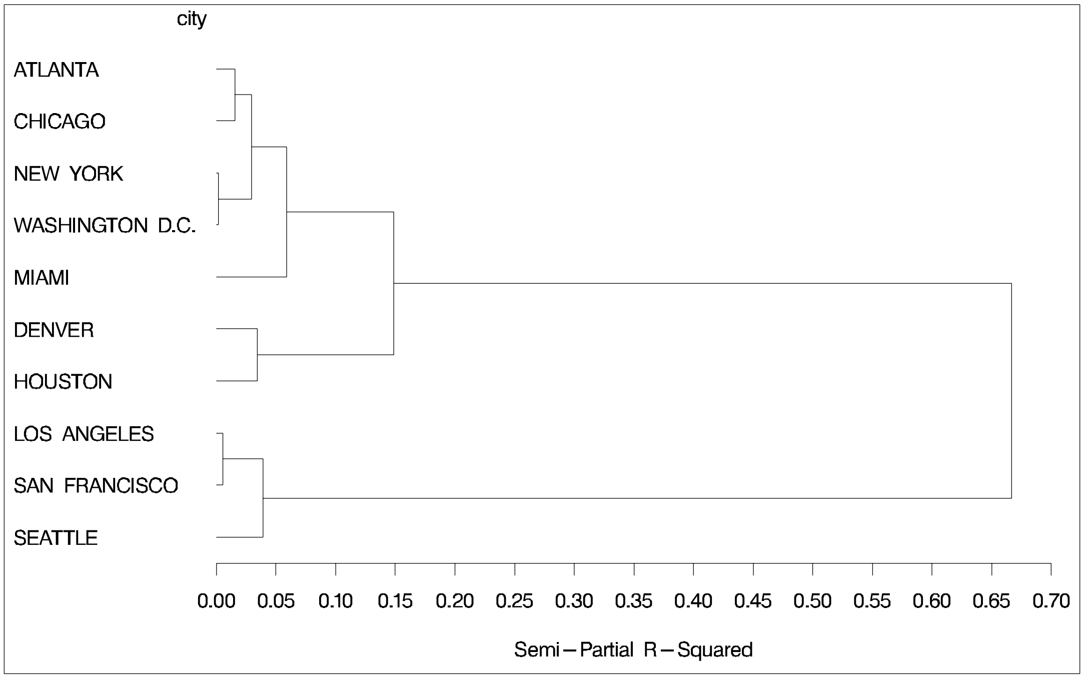
```
                Cluster Analysis of Flying Mileages Between 10 American Cities

                                    The CLUSTER Procedure
                               Single Linkage Cluster Analysis

                 Mean Distance Between Observations               = 1417.133


                                       Cluster History
                                                                        Norm    T
                                                                         Min    i
               NCL    ---------Clusters Joined----------      FREQ      Dist    e

                 9    NEW YORK           WASHINGTON D.C.         2    0.1447
                 8    LOS ANGELES        SAN FRANCISCO           2    0.2449
                 7    ATLANTA            CL9                     3    0.3832
                 6    CL7                CHICAGO                 4    0.4142
                 5    CL6                MIAMI                   5    0.4262
                 4    CL8                SEATTLE                 3    0.4784
                 3    CL5                HOUSTON                 6    0.4947
                 2    DENVER             CL4                     4    0.5864
                 1    CL3                CL2                    10    0.6203
```



```
                  Cluster Analysis of Flying Mileages Between 10 American Cities

                                      The CLUSTER Procedure
                                Two-Stage Density Linkage Clustering

                                               K = 3


                                          Cluster History
                                                       Normalized       Maximum Density      T
                                                           Fusion       in Each Cluster      i
  NCL    ---------Clusters Joined----------      FREQ       Density      Lesser     Greater    e

    9    ATLANTA           WASHINGTON D.C.         2        96.106     92.5043       100.0
    8    CL9               CHICAGO                 3        95.263     90.9548       100.0
    7    CL8               NEW YORK                4        86.465     76.1571       100.0
    6    CL7               MIAMI                   5        74.079     58.8299       100.0    T
    5    CL6               HOUSTON                 6        74.079     61.7747       100.0
    4    LOS ANGELES       SAN FRANCISCO           2        71.968     65.3430     80.0885
    3    CL4               SEATTLE                 3        66.341     56.6215     80.0885
    2    CL3               DENVER                  4        63.509     61.7747     80.0885
    1    CL5               CL2                    10        61.775     80.0885       100.0
                             2 modal clusters have been formed.
```



```
                     Cluster Analysis of Flying Mileages Between 10 American Cities

                                         The CLUSTER Procedure
                                  Ward's Minimum Variance Cluster Analysis

                      Root-Mean-Square Distance Between Observations   = 1580.242


                                            Cluster History
                                                                                          T
                                                                                          i
     NCL    ---------Clusters Joined----------      FREQ     SPRSQ     RSQ     PSF    PST2    e

       9    NEW YORK          WASHINGTON D.C.          2    0.0019    .998    66.7      .
       8    LOS ANGELES       SAN FRANCISCO            2    0.0054    .993    39.2      .
       7    ATLANTA           CHICAGO                  2    0.0153    .977    21.7      .
       6    CL7               CL9                      4    0.0296    .948    14.5     3.4
       5    DENVER            HOUSTON                  2    0.0344    .913    13.2      .
       4    CL8               SEATTLE                  3    0.0391    .874    13.9     7.3
       3    CL6               MIAMI                    5    0.0586    .816    15.5     3.8
       2    CL3               CL5                      7    0.1488    .667    16.0     5.3
       1    CL2               CL4                     10    0.6669    .000      .     16.0
```



Example 23.2. Crude Birth and Death Rates

The following example uses the SAS data set Poverty created in the “Getting Started” section beginning on page 837. The data, from Rouncefield (1995), are birth rates, death rates, and infant death rates for 97 countries. Six cluster analyses are performed with eight methods. Scatter plots showing cluster membership at selected levels are produced instead of tree diagrams.

Each cluster analysis is performed by a macro called ANALYZE. The macro takes two arguments. The first, &METHOD, specifies the value of the METHOD= option to be used in the PROC CLUSTER statement. The second, &NCL, must be specified as a list of integers, separated by blanks, indicating the number of clusters desired in each scatter plot. For example, the first invocation of ANALYZE specifies the AVERAGE method and requests plots of 3 and 8 clusters. When two-stage density linkage is used, the K= and R= options are specified as part of the first argument.

The ANALYZE macro first invokes the CLUSTER procedure with METHOD=&METHOD, where &METHOD represents the value of the first argument to ANALYZE. This part of the macro produces the PROC CLUSTER output shown.

The %DO loop processes &NCL, the list of numbers of clusters to plot. The macro variable &K is a counter that indexes the numbers within &NCL. The %SCAN function picks out the &Kth number in &NCL, which is then assigned to the macro variable &N. When &K exceeds the number of numbers in &NCL, %SCAN returns a null string. Thus, the %DO loop executes while &N is not equal to a null string. In the %WHILE condition, a null string is indicated by the absence of any nonblank characters between the comparison operator (NE) and the right parenthesis that terminates the condition.

Within the %DO loop, the TREE procedure creates an output data set containing &N clusters. The GPLOT procedure then produces a scatter plot in which each observation is identified by the number of the cluster to which it belongs. The TITLE2 statement uses double quotes so that &N and &METHOD can be used within the title. At the end of the loop, &K is incremented by 1, and the next number is extracted from &NCL by %SCAN.

For this example, plots are obtained only for average linkage. To generate plots for other methods, follow the example shown in the first macro call. The following statements produce Output 23.2.1 through Output 23.2.7.

```
title ’Cluster Analysis of Birth and Death Rates’;

%macro analyze(method,ncl);
proc cluster data=poverty outtree=tree method=&method p=15 ccc pseudo;
   var birth death;
   title2;
run;
%let k=1;
%let n=%scan(&ncl,&k);
%do %while(&n NE);
   proc tree data=tree noprint out=out ncl=&n;
      copy birth death;
   run;
   legend1 frame cframe=ligr cborder=black
           position=center value=(justify=center);
   axis1 label=(angle=90 rotate=0) minor=none;
   axis2 minor=none;
   proc gplot;
      plot death*birth=cluster /
      frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
      title2 "Plot of &n Clusters from METHOD=&METHOD";
   run;
   %let k=%eval(&k+1);
   %let n=%scan(&ncl,&k);
%end;
%mend;
```

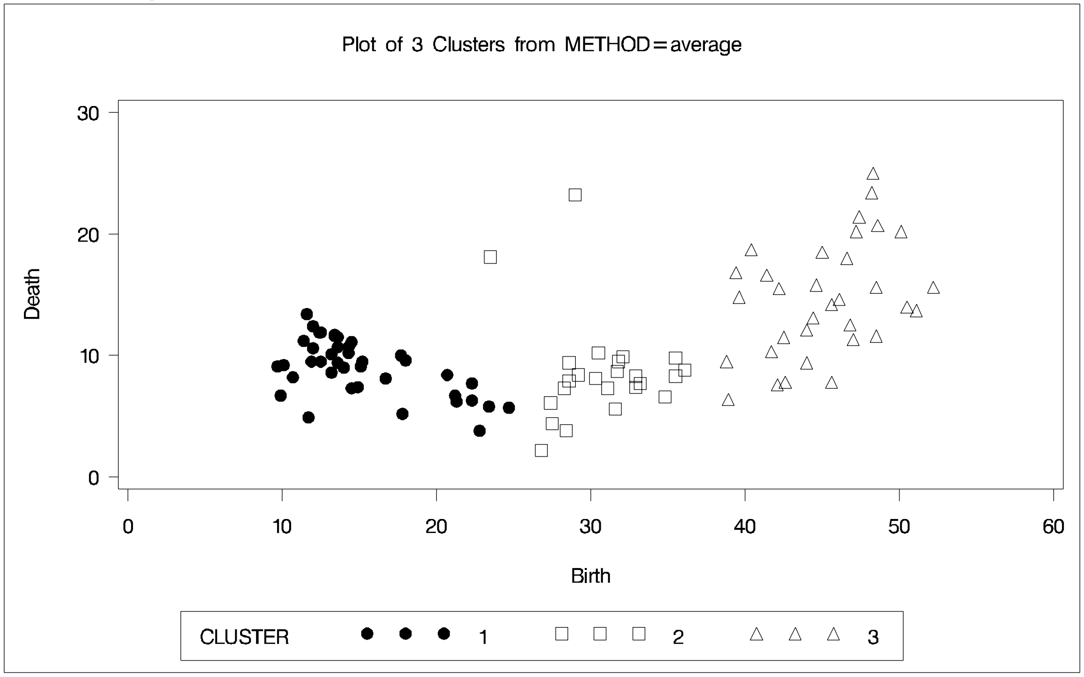
```
%analyze(average,3 8)
%analyze(complete,3)
%analyze(single,7 10)
%analyze(two k=10,3)
%analyze(two k=18,2)
```

For average linkage, the CCC has peaks at 3, 8, 10, and 12 clusters, but the 3-cluster peak is lower than the 8-cluster peak. The pseudo F statistic has peaks at 3, 8, and 12 clusters. The pseudo t^2 statistic drops sharply at 3 clusters, continues to fall at 4 clusters, and has a particularly low value at 12 clusters. However, there are not enough data to seriously consider as many as 12 clusters. Scatter plots are given for 3 and 8 clusters. The results are shown in Output 23.2.1 through Output 23.2.3. In Output 23.2.3, the eighth cluster consists of the two outlying observations, Mexico and Korea.

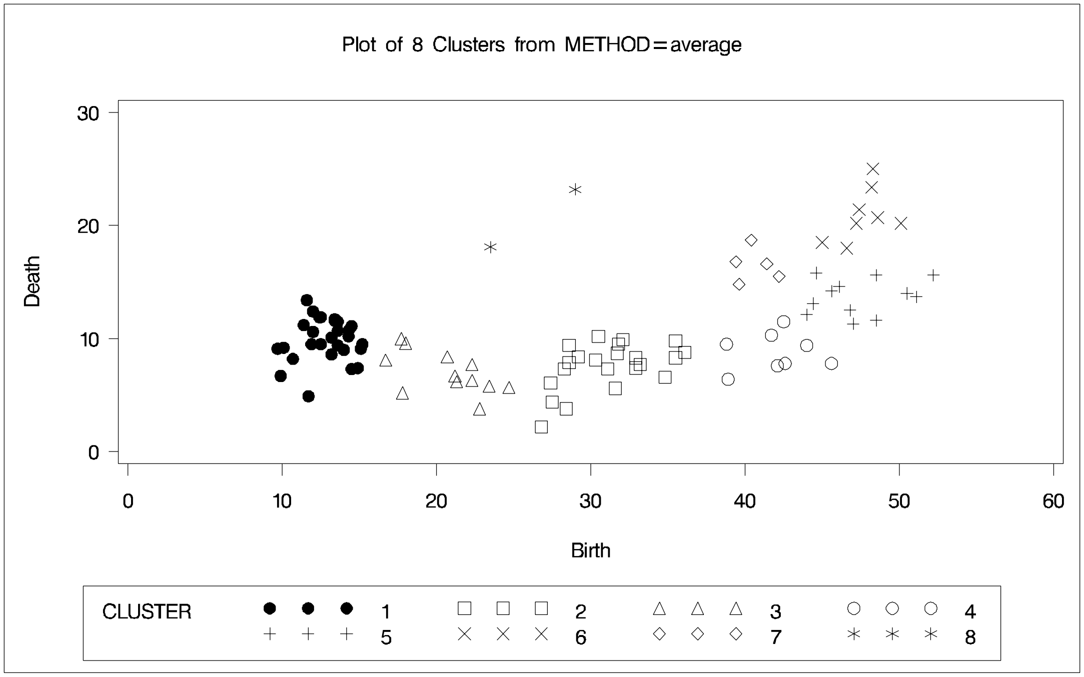
Output 23.2.1. Clusters for Birth and Death Rates: METHOD=AVERAGE

```
                              Cluster Analysis of Birth and Death Rates

                                       The CLUSTER Procedure
                                  Average Linkage Cluster Analysis

                                Eigenvalues of the Covariance Matrix

                             Eigenvalue    Difference    Proportion    Cumulative

                        1    189.106588    173.101020        0.9220        0.9220
                        2     16.005568                      0.0780        1.0000


                    Root-Mean-Square Total-Sample Standard Deviation =   10.127
                    Root-Mean-Square Distance Between Observations   = 20.25399


                                          Cluster History
                                                                                       Norm    T
                                                                                        RMS    i
NCL    --Clusters Joined---      FREQ     SPRSQ     RSQ    ERSQ     CCC     PSF    PST2      Dist    e

 15    CL27        CL20            18    0.0035    .980    .975    2.61     292    18.6    0.2325
 14    CL23        CL17            28    0.0034    .977    .972    1.97     271    17.7    0.2358
 13    CL18        CL54             8    0.0015    .975    .969    2.35     279     7.1    0.2432
 12    CL21        CL26             8    0.0015    .974    .966    2.85     290     6.1    0.2493
 11    CL19        CL24            12    0.0033    .971    .962    2.78     285    14.8    0.2767
 10    CL22        CL16            12    0.0036    .967    .957    2.84     284    17.4    0.2858
  9    CL15        CL28            22    0.0061    .961    .951    2.45     271    17.5    0.3353
  8    OB23        OB61             2    0.0014    .960    .943    3.59     302       .    0.3703
  7    CL25        CL11            17    0.0098    .950    .933    3.01     284    23.3    0.4033
  6    CL7         CL12            25    0.0122    .938    .920    2.63     273    14.8    0.4132
  5    CL10        CL14            40    0.0303    .907    .902    0.59     225    82.7    0.4584
  4    CL13        CL6             33    0.0244    .883    .875    0.77     234    22.2    0.5194
  3    CL9         CL8             24    0.0182    .865    .827    2.13     300    27.7     0.735
  2    CL5         CL3             64    0.1836    .681    .697    -.55     203     148    0.8402
  1    CL2         CL4             97    0.6810    .000    .000    0.00       .     203    1.3348
```

Output 23.2.2. Plot of Three Clusters, METHOD=AVERAGE



Output 23.2.3. Plot of Eight Clusters, METHOD=AVERAGE

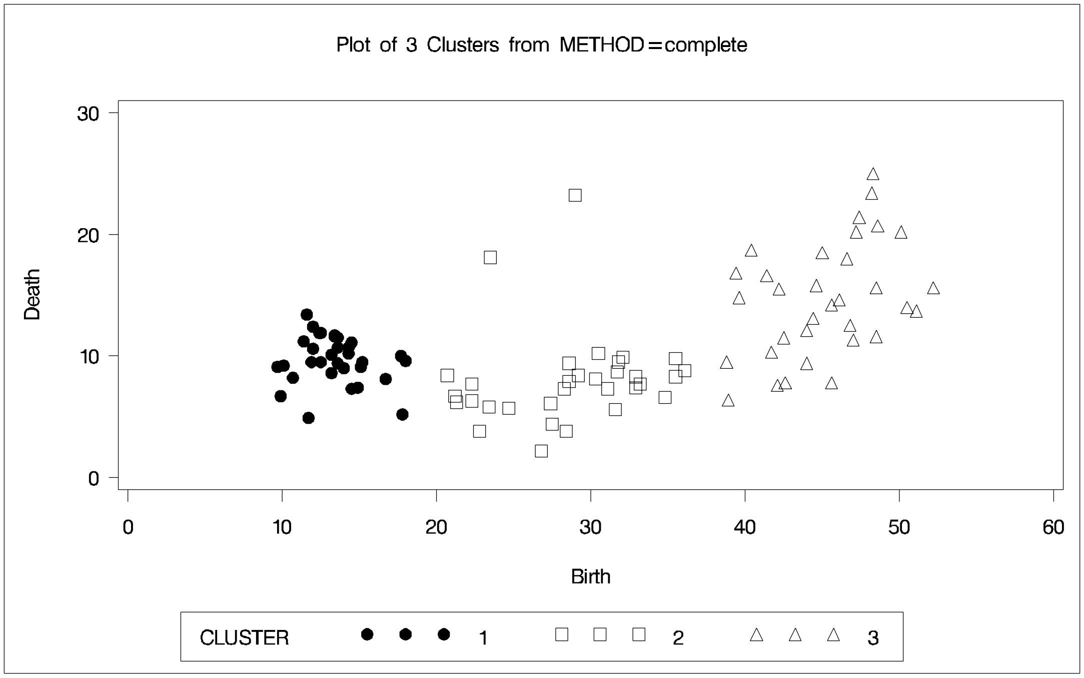


Complete linkage shows CCC peaks at 3, 8 and 12 clusters. The pseudo F statistic peaks at 3 and 12 clusters. The pseudo t^2 statistic indicates 3 clusters.

The scatter plot for 3 clusters is shown. The results are shown in Output 23.2.4.

Output 23.2.4. Clusters for Birth and Death Rates: METHOD=COMPLETE

```
                                  Cluster Analysis of Birth and Death Rates

                                           The CLUSTER Procedure
                                      Complete Linkage Cluster Analysis

                                     Eigenvalues of the Covariance Matrix

                                 Eigenvalue    Difference    Proportion    Cumulative

                             1    189.106588    173.101020        0.9220        0.9220
                             2     16.005568                      0.0780        1.0000

                        Root-Mean-Square Total-Sample Standard Deviation =   10.127
                        Mean Distance Between Observations               = 17.13099


                                              Cluster History
                                                                                                  Norm    T
                                                                                                   Max    i
          NCL    --Clusters Joined---      FREQ     SPRSQ     RSQ    ERSQ     CCC      PSF    PST2    Dist    e

           15    CL22        CL33             8    0.0015    .983    .975    3.80      329     6.1    0.4092
           14    CL56        CL18             8    0.0014    .981    .972    3.97      331     6.6    0.4255
           13    CL30        CL44             8    0.0019    .979    .969    4.04      330    19.0    0.4332
           12    OB23        OB61             2    0.0014    .978    .966    4.45      340      .     0.4378
           11    CL19        CL24            24    0.0034    .974    .962    4.17      327    24.1    0.4962
           10    CL17        CL28            12    0.0033    .971    .957    4.18      325    14.8    0.5204
            9    CL20        CL13            16    0.0067    .964    .951    3.38      297    25.2    0.5236
            8    CL11        CL21            32    0.0054    .959    .943    3.44      297    19.7    0.6001
            7    CL26        CL15            13    0.0096    .949    .933    2.93      282    28.9    0.7233
            6    CL14        CL10            20    0.0128    .937    .920    2.46      269    27.7    0.8033
            5    CL9         CL16            30    0.0237    .913    .902    1.29      241    47.1    0.8993
            4    CL6         CL7             33    0.0240    .889    .875    1.38      248    21.7    1.2165
            3    CL5         CL12            32    0.0178    .871    .827    2.56      317    13.6    1.2326
            2    CL3         CL8             64    0.1900    .681    .697    -.55      203     167    1.5412
            1    CL2         CL4             97    0.6810    .000    .000    0.00       .      203    2.5233
```

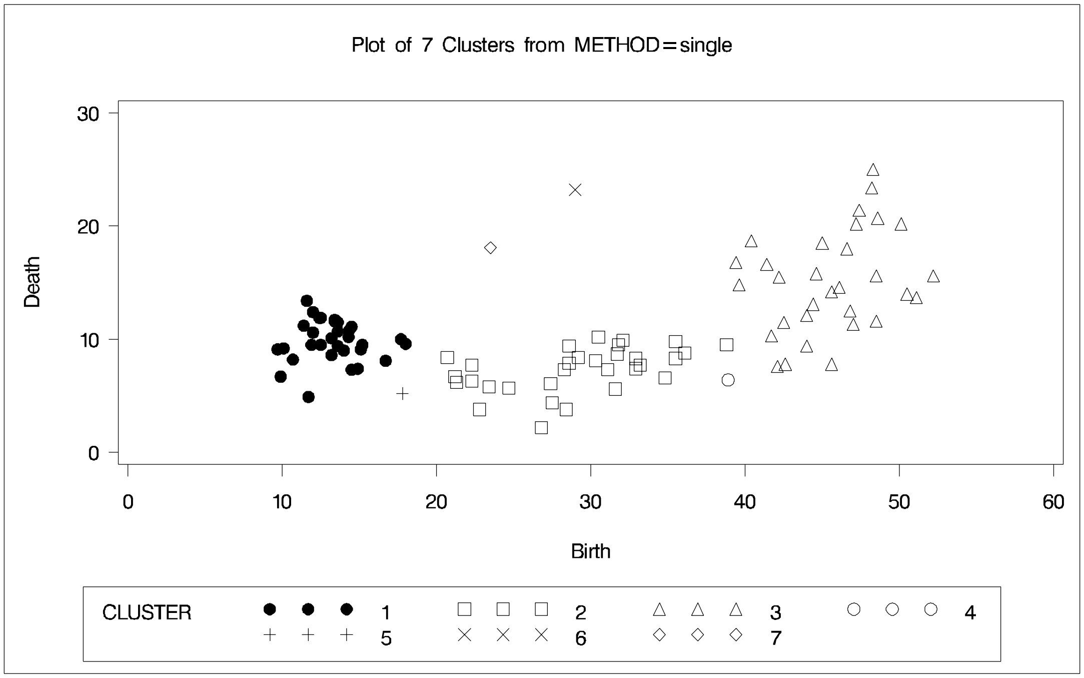


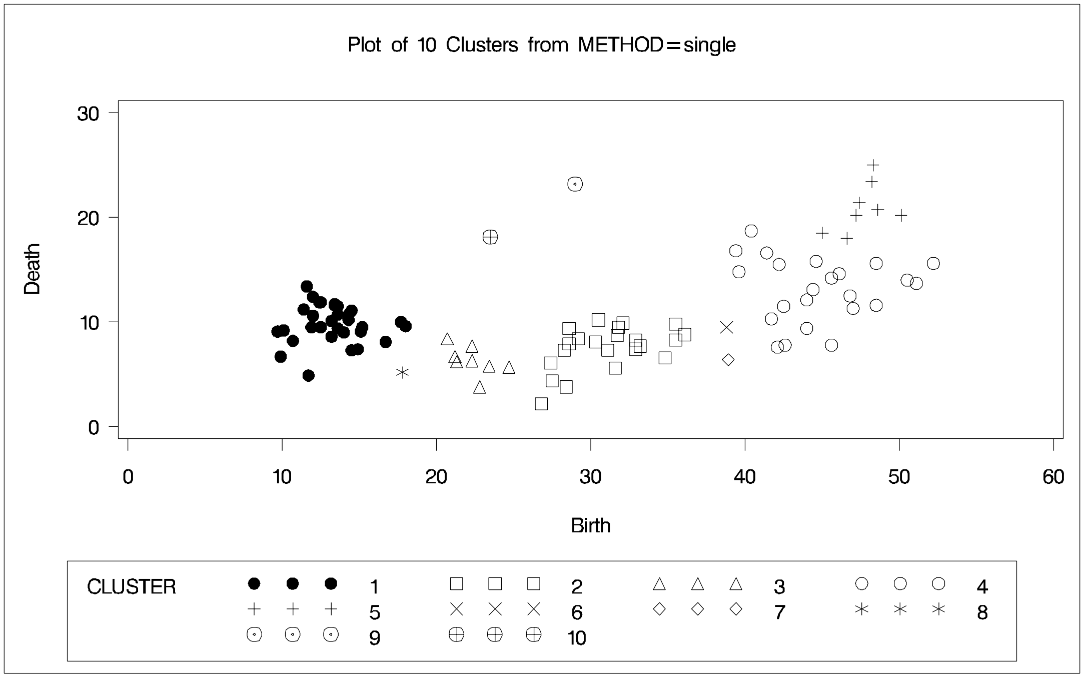
The CCC and pseudo F statistics are not appropriate for use with single linkage because of the method's tendency to chop off tails of distributions. The pseudo t^2 statistic can be used by looking for *large* values and taking the number of clusters to be one greater than the level at which the large pseudo t^2 value is displayed. For these data, there are large values at levels 6 and 9, suggesting 7 or 10 clusters.

The scatter plots for 7 and 10 clusters are shown. The results are shown in Output 23.2.5.

Output 23.2.5. Clusters for Birth and Death Rates: METHOD=SINGLE

```
                                   Cluster Analysis of Birth and Death Rates

                                            The CLUSTER Procedure
                                        Single Linkage Cluster Analysis

                                      Eigenvalues of the Covariance Matrix

                                  Eigenvalue    Difference    Proportion    Cumulative

                             1    189.106588    173.101020        0.9220        0.9220
                             2     16.005568                      0.0780        1.0000

                           Root-Mean-Square Total-Sample Standard Deviation =   10.127
                           Mean Distance Between Observations               = 17.13099


                                                Cluster History
                                                                                              Norm    T
                                                                                               Min    i
          NCL    --Clusters Joined---      FREQ     SPRSQ     RSQ    ERSQ     CCC     PSF    PST2     Dist    e

           15    CL37        CL19             8    0.0014    .968    .975    -2.3     178     6.6   0.1331
           14    CL20        CL23            15    0.0059    .962    .972    -3.1     162    18.7   0.1412
           13    CL14        CL16            19    0.0054    .957    .969    -3.4     155     8.8   0.1442
           12    CL26        OB58            31    0.0014    .955    .966    -2.7     165     4.0   0.1486
           11    OB86        CL18             4    0.0003    .955    .962    -1.6     183     3.8   0.1495
           10    CL13        CL11            23    0.0088    .946    .957    -2.3     170    11.3   0.1518
            9    CL22        CL17            30    0.0235    .923    .951    -4.7     131    45.7   0.1593    T
            8    CL15        CL10            31    0.0210    .902    .943    -5.8     117    21.8   0.1593
            7    CL9         OB75            31    0.0052    .897    .933    -4.7     130     4.0   0.1628
            6    CL7         CL12            62    0.2023    .694    .920     -15    41.3     223   0.1725
            5    CL6         CL8             93    0.6681    .026    .902     -26     0.6     199   0.1756
            4    CL5         OB48            94    0.0056    .021    .875     -24     0.7     0.5   0.1811    T
            3    CL4         OB67            95    0.0083    .012    .827     -15     0.6     0.8   0.1811
            2    OB23        OB61             2    0.0014    .011    .697     -13     1.0       .   0.4378
            1    CL3         CL2             97    0.0109    .000    .000    0.00       .     1.0   0.5815
```


Plot of 7 Clusters from METHOD=single
Death
30
20
10
0
Birth
0
10
20
30
40
50
60
CLUSTER
● ● ● 1
□ □ □ 2
△ △ △ 3
○ ○ ○ 4
+ + + 5
× × × 6
◇ ◇ ◇ 7


Plot of 10 Clusters from METHOD=single
Death
30
20
10
0
Birth
0
10
20
30
40
50
60
CLUSTER
● ● ● 1
□ □ □ 2
△ △ △ 3
○ ○ ○ 4
+ + + 5
× × × 6
◇ ◇ ◇ 7
* * * 8
⊙ ⊙ ⊙ 9
⊕ ⊕ ⊕ 10

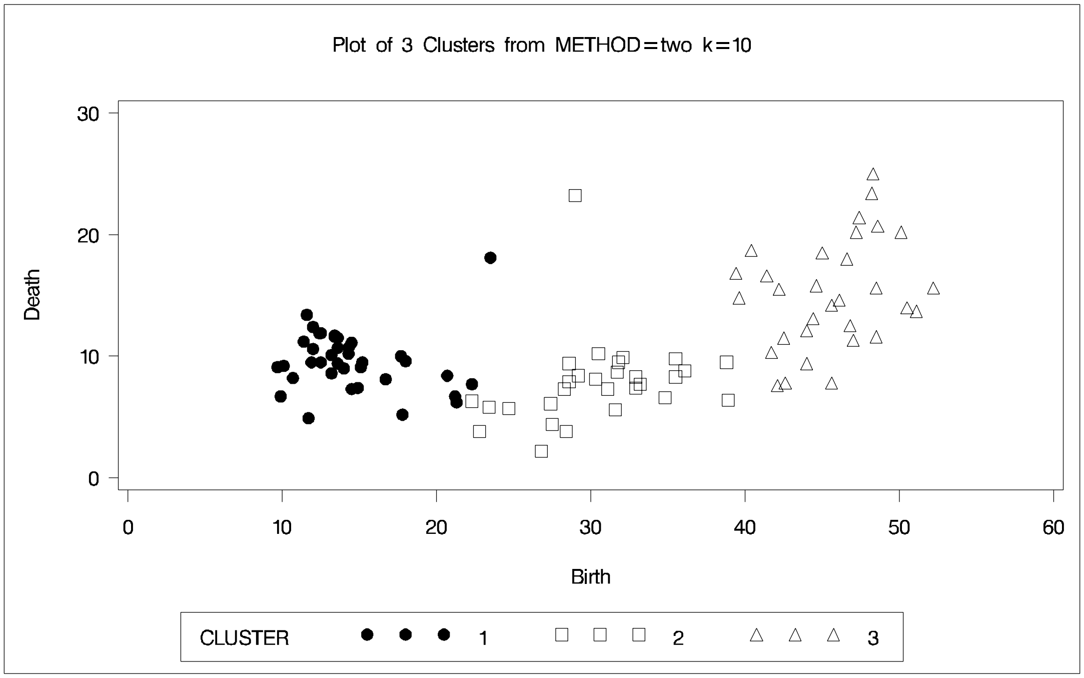
For kth-nearest-neighbor density linkage, the number of modes as a function of k is as follows (not all of these analyses are shown):

k	**modes**
3	13
4	6
5-7	4
8-15	3
16-21	2
22+	1

Thus, there is strong evidence of 3 modes and an indication of the possibility of 2 modes. Uniform-kernel density linkage gives similar results. For K=10 (10th-nearest-neighbor density linkage), the scatter plot for 3 clusters is shown; and for K=18, the scatter plot for 2 clusters is shown. The results are shown in Output 23.2.6.

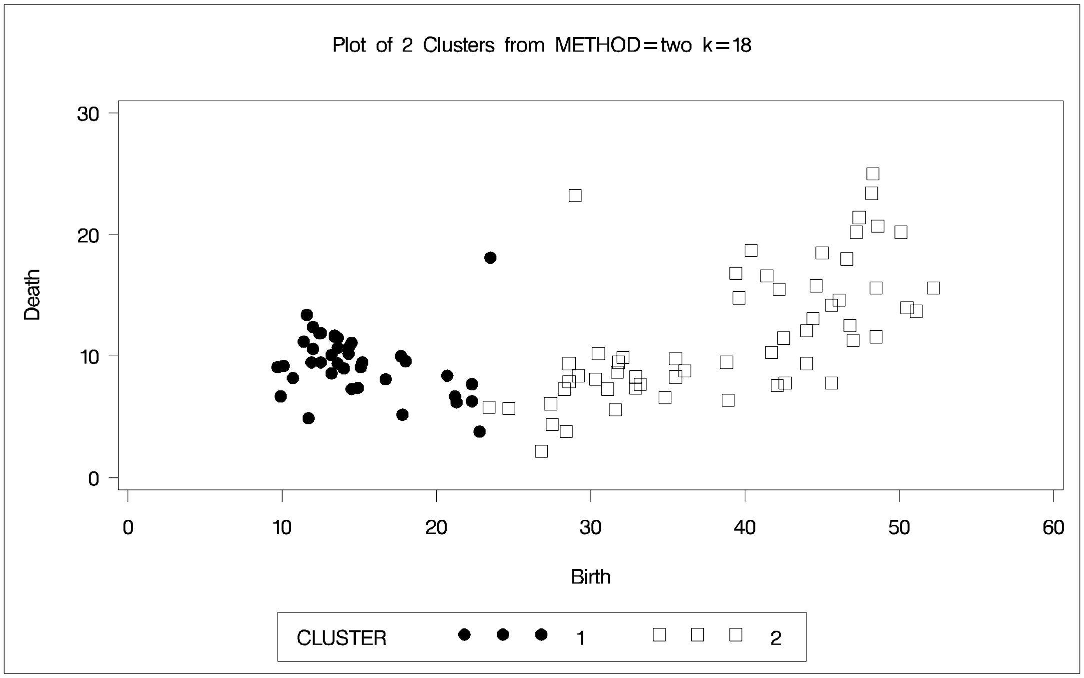
Output 23.2.6. Clusters for Birth and Death Rates: METHOD=TWOSTAGE, K=10

```
                                        Cluster Analysis of Birth and Death Rates

                                                  The CLUSTER Procedure
                                           Two-Stage Density Linkage Clustering

                                           Eigenvalues of the Covariance Matrix

                                       Eigenvalue    Difference    Proportion    Cumulative

                                  1    189.106588    173.101020        0.9220        0.9220
                                  2     16.005568                      0.0780        1.0000
                                                        K = 10
                               Root-Mean-Square Total-Sample Standard Deviation =   10.127


                                                     Cluster History
                                                                                      Normalized     Maximum Density     T
                                                                                          Fusion     in Each Cluster     i
   NCL   --Clusters Joined--     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2        Density     Lesser    Greater   e

    15   CL16       OB94           22   0.0015   .921   .975    -11   68.4    1.4         9.2234     6.7927    15.3069
    14   CL19       OB49           28   0.0021   .919   .972    -11   72.4    1.8         8.7369     5.9334    33.4385
    13   CL15       OB52           23   0.0024   .917   .969    -10   76.9    2.3         8.5847     5.9651    15.3069
    12   CL13       OB96           24   0.0018   .915   .966   -9.3   83.0    1.6         7.9252     5.4724    15.3069
    11   CL12       OB93           25   0.0025   .912   .962   -8.5   89.5    2.2         7.8913     5.4401    15.3069
    10   CL11       OB78           26   0.0031   .909   .957   -7.7   96.9    2.5          7.787     5.4082    15.3069
     9   CL10       OB76           27   0.0026   .907   .951   -6.7    107    2.1         7.7133     5.4401    15.3069
     8   CL9        OB77           28   0.0023   .904   .943   -5.5    120    1.7         7.4256     4.9017    15.3069
     7   CL8        OB43           29   0.0022   .902   .933   -4.1    138    1.6          6.927     4.4764    15.3069
     6   CL7        OB87           30   0.0043   .898   .920   -2.7    160    3.1          4.932     2.9977    15.3069
     5   CL6        OB82           31   0.0055   .892   .902   -1.1    191    3.7         3.7331     2.1560    15.3069
     4   CL22       OB61           37   0.0079   .884   .875   0.93    237   10.6         3.1713     1.6308      100.0
     3   CL14       OB23           29   0.0126   .872   .827   2.60    320   10.4         2.0654     1.0744    33.4385
     2   CL4        CL3            66   0.2129   .659   .697   -1.3    183    172         12.409    33.4385      100.0
     1   CL2        CL5            97   0.6588   .000   .000   0.00      .    183         10.071    15.3069      100.0
                                            3 modal clusters have been formed.
```



Output 23.2.7. Clusters for Birth and Death Rates: METHOD=TWOSTAGE, K=18

```
                                      Cluster Analysis of Birth and Death Rates

                                                The CLUSTER Procedure
                                         Two-Stage Density Linkage Clustering

                                         Eigenvalues of the Covariance Matrix

                                     Eigenvalue    Difference    Proportion    Cumulative

                                1    189.106588    173.101020        0.9220        0.9220
                                2     16.005568                      0.0780        1.0000
                                                        K = 18
                             Root-Mean-Square Total-Sample Standard Deviation =   10.127


                                                    Cluster History
                                                                                   Normalized     Maximum Density     T
                                                                                       Fusion     in Each Cluster     i
  NCL   --Clusters Joined--     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2      Density     Lesser    Greater   e

   15   CL16       OB72           46   0.0107   .799   .975    -21   23.3    3.0       10.118     7.7445    23.4457
   14   CL15       OB94           47   0.0098   .789   .972    -21   23.9    2.7        9.676     7.1257    23.4457
   13   CL14       OB51           48   0.0037   .786   .969    -20   25.6    1.0        9.409     6.8398    23.4457   T
   12   CL13       OB96           49   0.0099   .776   .966    -19   26.7    2.6        9.409     6.8398    23.4457
   11   CL12       OB76           50   0.0114   .764   .962    -19   27.9    2.9       8.8136     6.3138    23.4457
   10   CL11       OB77           51   0.0021   .762   .957    -18   31.0    0.5       8.6593     6.0751    23.4457
    9   CL10       OB78           52   0.0103   .752   .951    -17   33.3    2.5       8.6007     6.0976    23.4457
    8   CL9        OB43           53   0.0034   .748   .943    -16   37.8    0.8       8.4964     5.9160    23.4457
    7   CL8        OB93           54   0.0109   .737   .933    -15   42.1    2.6        8.367     5.7913    23.4457
    6   CL7        OB88           55   0.0110   .726   .920    -13   48.3    2.6        7.916     5.3679    23.4457
    5   CL6        OB87           56   0.0120   .714   .902    -12   57.5    2.7       6.6917     4.3415    23.4457
    4   CL20       OB61           39   0.0077   .707   .875   -9.8   74.7    8.3       6.2578     3.2882      100.0
    3   CL5        OB82           57   0.0138   .693   .827   -5.0    106    3.0       5.3605     3.2834    23.4457
    2   CL3        OB23           58   0.0117   .681   .697   -.54    203    2.5       3.2687     1.7568    23.4457
    1   CL2        CL4            97   0.6812   .000   .000   0.00      .    203       13.764    23.4457      100.0
                                          2 modal clusters have been formed.
```



In summary, most of the clustering methods indicate 3 or 8 clusters. Most methods agree at the 3-cluster level, but at the other levels, there is considerable disagreement about the composition of the clusters. The presence of numerous ties also complicates the analysis; see Example 23.4.

Example 23.3. Cluster Analysis of Fisher Iris Data

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on fifty iris specimens from each of three species, *Iris setosa, I. versicolor,* and *I. virginica*. Mezzich and Solomon (1980) discuss a variety of cluster analyses of the iris data.

This example analyzes the iris data by Ward's method and two-stage density linkage and then illustrates how the FASTCLUS procedure can be used in combination with PROC CLUSTER to analyze large data sets.

```
title 'Cluster Analysis of Fisher (1936) Iris Data';
proc format;
   value specname
      1='Setosa    '
      2='Versicolor'
      3='Virginica ';
run;

data iris;
   input SepalLength SepalWidth PetalLength PetalWidth Species @@;
   format Species specname.;
   label SepalLength='Sepal Length in mm.'
```

```
         SepalWidth =’Sepal Width in mm.’
         PetalLength=’Petal Length in mm.’
         PetalWidth =’Petal Width in mm.’;
   symbol = put(species, specname10.);
   datalines;
50 33 14 02 1 64 28 56 22 3 65 28 46 15 2 67 31 56 24 3
63 28 51 15 3 46 34 14 03 1 69 31 51 23 3 62 22 45 15 2
59 32 48 18 2 46 36 10 02 1 61 30 46 14 2 60 27 51 16 2
65 30 52 20 3 56 25 39 11 2 65 30 55 18 3 58 27 51 19 3
68 32 59 23 3 51 33 17 05 1 57 28 45 13 2 62 34 54 23 3
77 38 67 22 3 63 33 47 16 2 67 33 57 25 3 76 30 66 21 3
49 25 45 17 3 55 35 13 02 1 67 30 52 23 3 70 32 47 14 2
64 32 45 15 2 61 28 40 13 2 48 31 16 02 1 59 30 51 18 3
55 24 38 11 2 63 25 50 19 3 64 32 53 23 3 52 34 14 02 1
49 36 14 01 1 54 30 45 15 2 79 38 64 20 3 44 32 13 02 1
67 33 57 21 3 50 35 16 06 1 58 26 40 12 2 44 30 13 02 1
77 28 67 20 3 63 27 49 18 3 47 32 16 02 1 55 26 44 12 2
50 23 33 10 2 72 32 60 18 3 48 30 14 03 1 51 38 16 02 1
61 30 49 18 3 48 34 19 02 1 50 30 16 02 1 50 32 12 02 1
61 26 56 14 3 64 28 56 21 3 43 30 11 01 1 58 40 12 02 1
51 38 19 04 1 67 31 44 14 2 62 28 48 18 3 49 30 14 02 1
51 35 14 02 1 56 30 45 15 2 58 27 41 10 2 50 34 16 04 1
46 32 14 02 1 60 29 45 15 2 57 26 35 10 2 57 44 15 04 1
50 36 14 02 1 77 30 61 23 3 63 34 56 24 3 58 27 51 19 3
57 29 42 13 2 72 30 58 16 3 54 34 15 04 1 52 41 15 01 1
71 30 59 21 3 64 31 55 18 3 60 30 48 18 3 63 29 56 18 3
49 24 33 10 2 56 27 42 13 2 57 30 42 12 2 55 42 14 02 1
49 31 15 02 1 77 26 69 23 3 60 22 50 15 3 54 39 17 04 1
66 29 46 13 2 52 27 39 14 2 60 34 45 16 2 50 34 15 02 1
44 29 14 02 1 50 20 35 10 2 55 24 37 10 2 58 27 39 12 2
47 32 13 02 1 46 31 15 02 1 69 32 57 23 3 62 29 43 13 2
74 28 61 19 3 59 30 42 15 2 51 34 15 02 1 50 35 13 03 1
56 28 49 20 3 60 22 40 10 2 73 29 63 18 3 67 25 58 18 3
49 31 15 01 1 67 31 47 15 2 63 23 44 13 2 54 37 15 02 1
56 30 41 13 2 63 25 49 15 2 61 28 47 12 2 64 29 43 13 2
51 25 30 11 2 57 28 41 13 2 65 30 58 22 3 69 31 54 21 3
54 39 13 04 1 51 35 14 03 1 72 36 61 25 3 65 32 51 20 3
61 29 47 14 2 56 29 36 13 2 69 31 49 15 2 64 27 53 19 3
68 30 55 21 3 55 25 40 13 2 48 34 16 02 1 48 30 14 01 1
45 23 13 03 1 57 25 50 20 3 57 38 17 03 1 51 38 15 03 1
55 23 40 13 2 66 30 44 14 2 68 28 48 14 2 54 34 17 02 1
51 37 15 04 1 52 35 15 02 1 58 28 51 24 3 67 30 50 17 2
63 33 60 25 3 53 37 15 02 1
;
```

The following macro, SHOW, is used in the subsequent analyses to display cluster results. It invokes the FREQ procedure to crosstabulate clusters and species. The CANDISC procedure computes canonical variables for discriminating among the clusters, and the first two canonical variables are plotted to show cluster membership. See Chapter 21, “The CANDISC Procedure,” for a canonical discriminant analysis of the iris species.

```
%macro show;
proc freq;
   tables cluster*species;
run;
proc candisc noprint out=can;
   class cluster;
   var petal: sepal:;
run;
legend1 frame cframe=ligr cborder=black
        position=center value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none;
axis2 minor=none;
proc gplot;
   plot can2*can1=cluster /
      frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;
%mend;
```

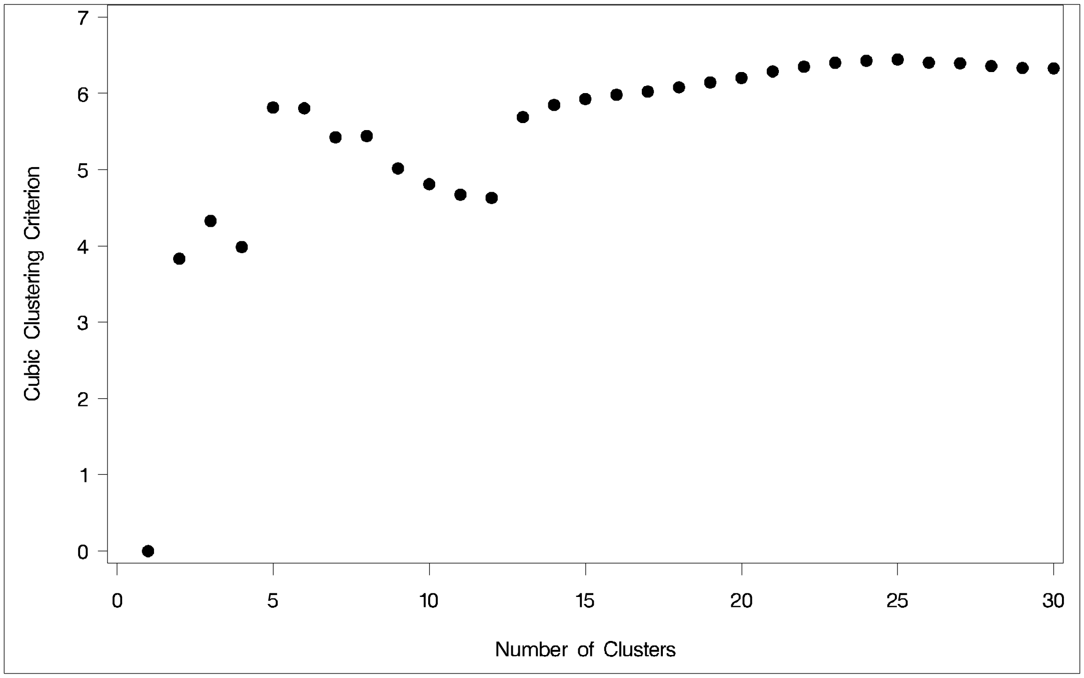
The first analysis clusters the iris data by Ward’s method and plots the CCC and pseudo F and t^2 statistics. The CCC has a local peak at 3 clusters but a higher peak at 5 clusters. The pseudo F statistic indicates 3 clusters, while the pseudo t^2 statistic suggests 3 or 6 clusters. For large numbers of clusters, Version 6 of the SAS System produces somewhat different results than previous versions of PROC CLUSTER. This is due to changes in the treatment of ties. Results are identical for 5 or fewer clusters.

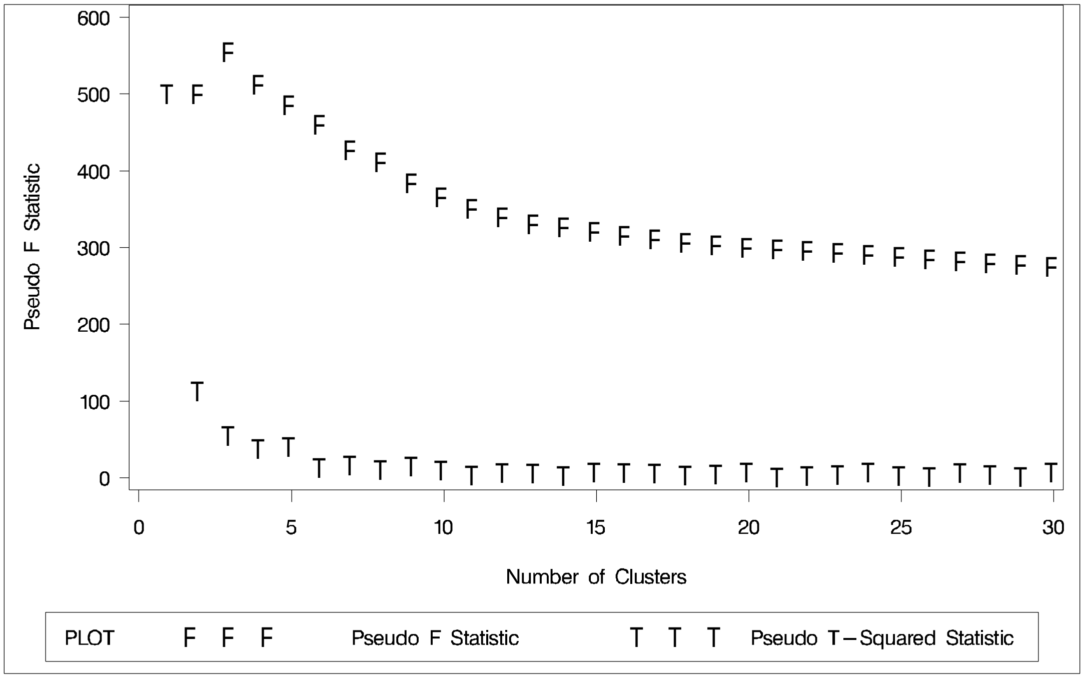
The TREE procedure creates an output data set containing the 3-cluster partition for use by the SHOW macro. The FREQ procedure reveals 16 misclassifications. The results are shown in Output 23.3.1.

```
title2 ’By Ward’’s Method’;
proc cluster data=iris method=ward print=15 ccc pseudo;
   var petal: sepal:;
   copy species;
run;
legend1 frame cframe=ligr cborder=black
        position=center value=(justify=center);
axis1 label=(angle=90 rotate=0) minor=none order=(0 to 600 by 100);
axis2 minor=none order=(1 to 30 by 1);
axis3 label=(angle=90 rotate=0) minor=none order=(0 to 7 by 1);
proc gplot;
   plot _ccc_*_ncl_  /
      frame cframe=ligr legend=legend1 vaxis=axis3 haxis=axis2;
   plot _psf_*_ncl_  _pst2_*_ncl_  /overlay
      frame cframe=ligr legend=legend1 vaxis=axis1 haxis=axis2;
run;

proc tree noprint ncl=3 out=out;
   copy petal: sepal: species;
run;

%show;
```

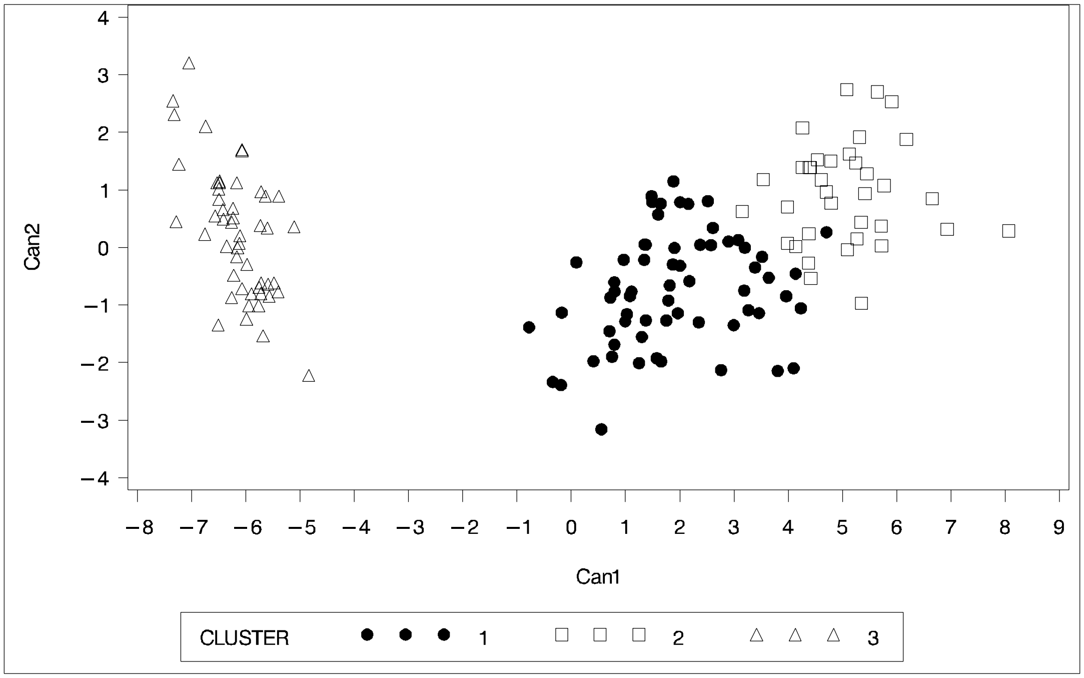
Output 23.3.1. Cluster Analysis of Fisher Iris Data: CLUSTER with METHOD=WARD

```
                               Cluster Analysis of Fisher (1936) Iris Data
                                            By Ward's Method

                                         The CLUSTER Procedure
                                 Ward's Minimum Variance Cluster Analysis

                                   Eigenvalues of the Covariance Matrix

                               Eigenvalue    Difference    Proportion    Cumulative

                          1    422.824171    398.557096        0.9246        0.9246
                          2     24.267075     16.446125        0.0531        0.9777
                          3      7.820950      5.437441        0.0171        0.9948
                          4      2.383509                      0.0052        1.0000


                      Root-Mean-Square Total-Sample Standard Deviation = 10.69224
                      Root-Mean-Square Distance Between Observations   = 30.24221


                                            Cluster History
                                                                                             T
                                                                                             i
         NCL    --Clusters Joined---      FREQ     SPRSQ     RSQ    ERSQ     CCC     PSF    PST2    e

          15    CL24        CL28            15    0.0016    .971    .958    5.93     324     9.8
          14    CL21        CL53             7    0.0019    .969    .955    5.85     329     5.1
          13    CL18        CL48            15    0.0023    .967    .953    5.69     334     8.9
          12    CL16        CL23            24    0.0023    .965    .950    4.63     342     9.6
          11    CL14        CL43            12    0.0025    .962    .946    4.67     353     5.8
          10    CL26        CL20            22    0.0027    .959    .942    4.81     368    12.9
           9    CL27        CL17            31    0.0031    .956    .936    5.02     387    17.8
           8    CL35        CL15            23    0.0031    .953    .930    5.44     414    13.8
           7    CL10        CL47            26    0.0058    .947    .921    5.43     430    19.1
           6    CL8         CL13            38    0.0060    .941    .911    5.81     463    16.3
           5    CL9         CL19            50    0.0105    .931    .895    5.82     488    43.2
           4    CL12        CL11            36    0.0172    .914    .872    3.99     515    41.0
           3    CL6         CL7             64    0.0301    .884    .827    4.33     558    57.2
           2    CL4         CL3            100    0.1110    .773    .697    3.83     503     116
           1    CL5         CL2            150    0.7726    .000    .000    0.00       .     503
```





```
                   Cluster Analysis of Fisher (1936) Iris Data

                              The FREQ Procedure

                         Table of CLUSTER by Species

                   CLUSTER     Species

                   Frequency|
                   Percent  |
                   Row Pct  |
                   Col Pct  |Setosa  |Versicol|Virginic|  Total
                            |        |or      |a       |
                   ---------+--------+--------+--------+
                          1 |      0 |     49 |     15 |     64
                            |   0.00 |  32.67 |  10.00 |  42.67
                            |   0.00 |  76.56 |  23.44 |
                            |   0.00 |  98.00 |  30.00 |
                   ---------+--------+--------+--------+
                          2 |      0 |      1 |     35 |     36
                            |   0.00 |   0.67 |  23.33 |  24.00
                            |   0.00 |   2.78 |  97.22 |
                            |   0.00 |   2.00 |  70.00 |
                   ---------+--------+--------+--------+
                          3 |     50 |      0 |      0 |     50
                            |  33.33 |   0.00 |   0.00 |  33.33
                            | 100.00 |   0.00 |   0.00 |
                            | 100.00 |   0.00 |   0.00 |
                   ---------+--------+--------+--------+
                   Total          50       50       50      150
                               33.33    33.33    33.33   100.00
```



The second analysis uses two-stage density linkage. The raw data suggest 2 or 6 modes instead of 3:

k	**modes**
3	12
4-6	6
7	4
8	3
9-50	2
51+	1

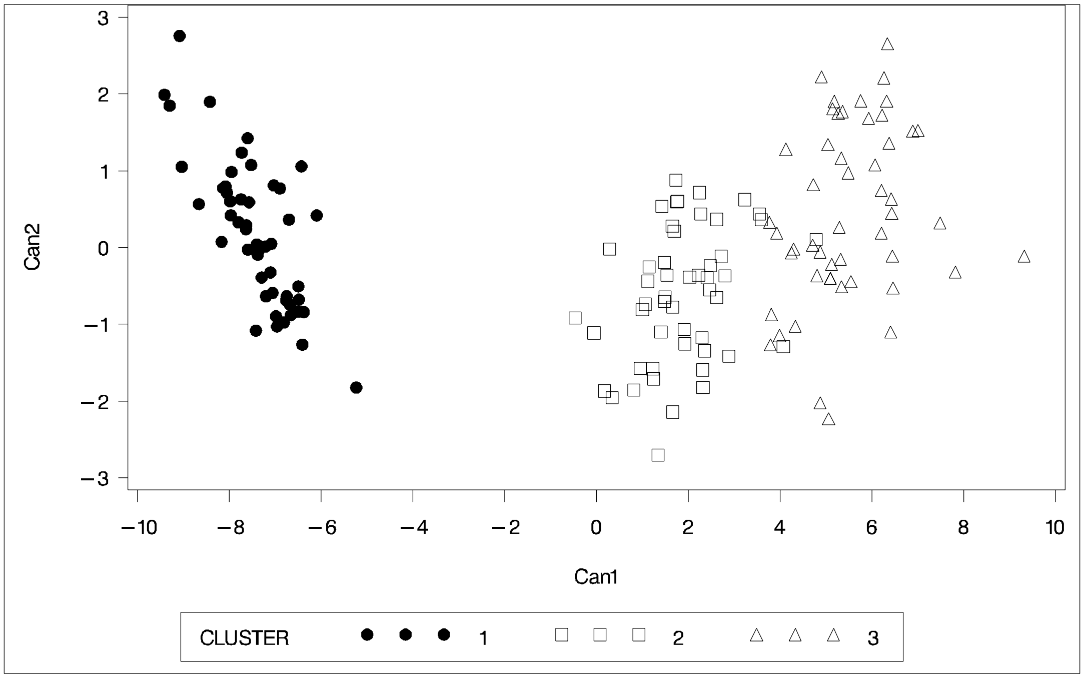
However, the ACECLUS procedure can be used to reveal 3 modes. This analysis uses K=8 to produce 3 clusters for comparison with other analyses. There are only 6 misclassifications. The results are shown in Output 23.3.2.

```
title2 ’By Two-Stage Density Linkage’;
proc cluster data=iris method=twostage k=8 print=15 ccc pseudo;
   var petal: sepal:;
   copy species;
run;

proc tree noprint ncl=3 out=out;
   copy petal: sepal: species;
run;

%show;
```

Output 23.3.2. Cluster Analysis of Fisher Iris Data: CLUSTER with METHOD=TWOSTAGE

```
                                    Cluster Analysis of Fisher (1936) Iris Data
                                            By Two-Stage Density Linkage

                                               The CLUSTER Procedure
                                        Two-Stage Density Linkage Clustering

                                        Eigenvalues of the Covariance Matrix

                                    Eigenvalue    Difference    Proportion    Cumulative

                               1    422.824171    398.557096        0.9246        0.9246
                               2     24.267075     16.446125        0.0531        0.9777
                               3      7.820950      5.437441        0.0171        0.9948
                               4      2.383509                      0.0052        1.0000
                                                       K = 8
                            Root-Mean-Square Total-Sample Standard Deviation = 10.69224


                                                   Cluster History
                                                                                  Normalized     Maximum Density     T
                                                                                      Fusion     in Each Cluster     i
 NCL   --Clusters Joined--     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2      Density     Lesser    Greater   e

  15   CL17       OB127          44   0.0025   .916   .958    -11    105    3.4       0.3903     0.2066     3.5156
  14   CL16       OB137          50   0.0023   .913   .955    -11    110    5.6       0.3637     0.1837      100.0
  13   CL15       OB74           45   0.0029   .910   .953    -10    116    3.7       0.3553     0.2130     3.5156
  12   CL28       OB49           46   0.0036   .907   .950   -8.0    122    5.2       0.3223     0.1736     8.3678   T
  11   CL12       OB85           47   0.0036   .903   .946   -7.6    130    4.8       0.3223     0.1736     8.3678
  10   CL11       OB98           48   0.0033   .900   .942   -7.1    140    4.1       0.2879     0.1479     8.3678
   9   CL13       OB24           46   0.0037   .896   .936   -6.5    152    4.4       0.2802     0.2005     3.5156
   8   CL10       OB25           49   0.0019   .894   .930   -5.5    171    2.2       0.2699     0.1372     8.3678
   7   CL8        OB121          50   0.0035   .891   .921   -4.5    194    4.0       0.2586     0.1372     8.3678
   6   CL9        OB45           47   0.0042   .886   .911   -3.3    225    4.6       0.1412     0.0832     3.5156
   5   CL6        OB39           48   0.0049   .882   .895   -1.7    270    5.0        0.107     0.0605     3.5156
   4   CL5        OB21           49   0.0049   .877   .872   0.35    346    4.7       0.0969     0.0541     3.5156
   3   CL4        OB90           50   0.0047   .872   .827   3.28    500    4.1       0.0715     0.0370     3.5156
   2   CL3        CL7           100   0.0993   .773   .697   3.83    503   91.9       2.6277     3.5156     8.3678
                                         3 modal clusters have been formed.
```

```
                                    Cluster Analysis of Fisher (1936) Iris Data

                                                 The FREQ Procedure

                                            Table of CLUSTER by Species

                                    CLUSTER     Species

                                    Frequency|
                                    Percent  |
                                    Row Pct  |
                                    Col Pct  |Setosa  |Versicol|Virginic|  Total
                                             |        |or      |a       |
                                    ---------+--------+--------+--------+
                                           1 |     50 |      0 |      0 |     50
                                             |  33.33 |   0.00 |   0.00 |  33.33
                                             | 100.00 |   0.00 |   0.00 |
                                             | 100.00 |   0.00 |   0.00 |
                                    ---------+--------+--------+--------+
                                           2 |      0 |     47 |      3 |     50
                                             |   0.00 |  31.33 |   2.00 |  33.33
                                             |   0.00 |  94.00 |   6.00 |
                                             |   0.00 |  94.00 |   6.00 |
                                    ---------+--------+--------+--------+
                                           3 |      0 |      3 |     47 |     50
                                             |   0.00 |   2.00 |  31.33 |  33.33
                                             |   0.00 |   6.00 |  94.00 |
                                             |   0.00 |   6.00 |  94.00 |
                                    ---------+--------+--------+--------+
                                    Total          50       50       50      150
                                                33.33    33.33    33.33   100.00
```



The CLUSTER procedure is not practical for very large data sets because, with most methods, the CPU time varies as the square or cube of the number of observations. The FASTCLUS procedure requires time proportional to the number of observations and can, therefore, be used with much larger data sets than PROC CLUSTER. If you want to hierarchically cluster a very large data set, you can use PROC FASTCLUS for a preliminary cluster analysis producing a large number of clusters and then use PROC CLUSTER to hierarchically cluster the preliminary clusters.

FASTCLUS automatically creates variables _FREQ_ and _RMSSTD_ in the MEAN= output data set. These variables are then automatically used by PROC CLUSTER in the computation of various statistics.

The iris data are used to illustrate the process of clustering clusters. In the preliminary analysis, PROC FASTCLUS produces ten clusters, which are then crosstabulated with species. The data set containing the preliminary clusters is sorted in preparation for later merges. The results are shown in Output 23.3.3.

```
title2 'Preliminary Analysis by FASTCLUS';
proc fastclus data=iris summary maxc=10 maxiter=99 converge=0
              mean=mean out=prelim cluster=preclus;
   var petal: sepal:;
run;

proc freq;
   tables preclus*species;
run;

proc sort data=prelim;
   by preclus;
run;
```

Output 23.3.3. Preliminary Analysis of Fisher Iris Data

```
                           Cluster Analysis of Fisher (1936) Iris Data
                                 Preliminary Analysis by FASTCLUS

                                      The FASTCLUS Procedure
                  Replace=FULL  Radius=0  Maxclusters=10 Maxiter=99  Converge=0

                                         Cluster Summary

                                      Maximum Distance
                           RMS Std           from Seed     Radius     Nearest     Distance Between
Cluster     Frequency    Deviation      to Observation    Exceeded    Cluster    Cluster Centroids
--------------------------------------------------------------------------------------------------
    1               9       2.7067              8.2027                      5               8.7362
    2              19       2.2001              7.7340                      4               6.2243
    3              18       2.1496              6.2173                      8               7.5049
    4               4       2.5249              5.3268                      2               6.2243
    5               3       2.7234              5.8214                      1               8.7362
    6               7       2.2939              5.1508                      2               9.3318
    7              17       2.0274              6.9576                     10               7.9503
    8              18       2.2628              7.1135                      3               7.5049
    9              22       2.2666              7.5029                      8               9.0090
   10              33       2.0594             10.0033                      7               7.9503


                                  Pseudo F Statistic =   370.58


                              Observed Over-All R-Squared =  0.95971


                       Approximate Expected Over-All R-Squared =   0.82928


                              Cubic Clustering Criterion =   27.077

               WARNING: The two values above are invalid for correlated variables.
```

```
Cluster Analysis of Fisher (1936) Iris Data
      Preliminary Analysis by FASTCLUS

             The FREQ Procedure

        Table of PRECLUS by Species

PRECLUS(Cluster)     Species

Frequency|
Percent  |
Row Pct  |
Col Pct  |Setosa  |Versicol|Virginic|  Total
         |        |or      |a       |
---------+--------+--------+--------+
       1 |      0 |      0 |      9 |      9
         |   0.00 |   0.00 |   6.00 |   6.00
         |   0.00 |   0.00 | 100.00 |
         |   0.00 |   0.00 |  18.00 |
---------+--------+--------+--------+
       2 |      0 |     19 |      0 |     19
         |   0.00 |  12.67 |   0.00 |  12.67
         |   0.00 | 100.00 |   0.00 |
         |   0.00 |  38.00 |   0.00 |
---------+--------+--------+--------+
       3 |      0 |     18 |      0 |     18
         |   0.00 |  12.00 |   0.00 |  12.00
         |   0.00 | 100.00 |   0.00 |
         |   0.00 |  36.00 |   0.00 |
---------+--------+--------+--------+
       4 |      0 |      3 |      1 |      4
         |   0.00 |   2.00 |   0.67 |   2.67
         |   0.00 |  75.00 |  25.00 |
         |   0.00 |   6.00 |   2.00 |
---------+--------+--------+--------+
       5 |      0 |      0 |      3 |      3
         |   0.00 |   0.00 |   2.00 |   2.00
         |   0.00 |   0.00 | 100.00 |
         |   0.00 |   0.00 |   6.00 |
---------+--------+--------+--------+
       6 |      0 |      7 |      0 |      7
         |   0.00 |   4.67 |   0.00 |   4.67
         |   0.00 | 100.00 |   0.00 |
         |   0.00 |  14.00 |   0.00 |
---------+--------+--------+--------+
       7 |     17 |      0 |      0 |     17
         |  11.33 |   0.00 |   0.00 |  11.33
         | 100.00 |   0.00 |   0.00 |
         |  34.00 |   0.00 |   0.00 |
---------+--------+--------+--------+
       8 |      0 |      3 |     15 |     18
         |   0.00 |   2.00 |  10.00 |  12.00
         |   0.00 |  16.67 |  83.33 |
         |   0.00 |   6.00 |  30.00 |
---------+--------+--------+--------+
       9 |      0 |      0 |     22 |     22
         |   0.00 |   0.00 |  14.67 |  14.67
         |   0.00 |   0.00 | 100.00 |
         |   0.00 |   0.00 |  44.00 |
---------+--------+--------+--------+
      10 |     33 |      0 |      0 |     33
         |  22.00 |   0.00 |   0.00 |  22.00
         | 100.00 |   0.00 |   0.00 |
         |  66.00 |   0.00 |   0.00 |
---------+--------+--------+--------+
Total          50       50       50      150
            33.33    33.33    33.33   100.00
```

The following macro, CLUS, clusters the preliminary clusters. There is one argument to choose the METHOD= specification to be used by PROC CLUSTER. The TREE procedure creates an output data set containing the 3-cluster partition, which is sorted and merged with the OUT= data set from PROC FASTCLUS to determine to which cluster each of the original 150 observations belongs. The SHOW macro is then used to display the results. In this example, the CLUS macro is invoked using

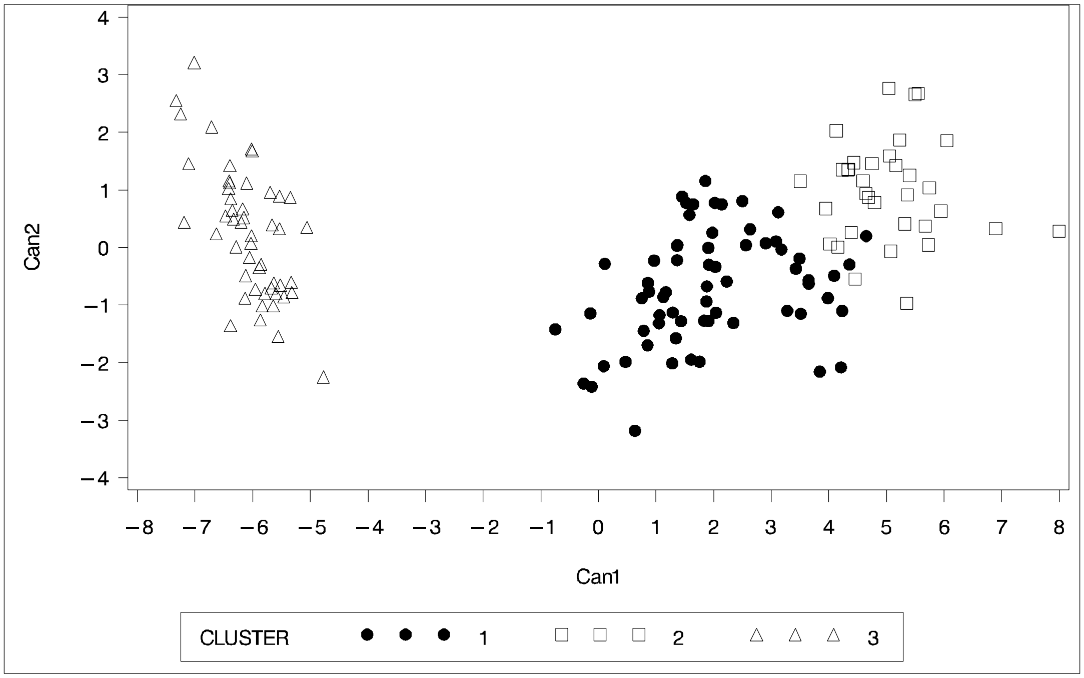
Ward’s method, which produces 16 misclassifications, and Wong’s hybrid method, which produces 22 misclassifications. The results are shown in Output 23.3.4 and Output 23.3.5.

```
%macro clus(method);
proc cluster data=mean method=&method ccc pseudo;
   var petal: sepal:;
   copy preclus;
run;
proc tree noprint ncl=3 out=out;
   copy petal: sepal: preclus;
run;
proc sort data=out;
   by preclus;
run;
data clus;
   merge prelim out;
   by preclus;
run;
%show;
%mend;

title2 ’Clustering Clusters by Ward’’s Method’;
%clus(ward);

title2 ’Clustering Clusters by Wong’’s Hybrid Method’;
%clus(twostage hybrid);
```

Output 23.3.4. Clustering Clusters: with Ward’s Method

```
                          Cluster Analysis of Fisher (1936) Iris Data
                              Clustering Clusters by Ward’s Method

                                     The CLUSTER Procedure
                            Ward’s Minimum Variance Cluster Analysis

                               Eigenvalues of the Covariance Matrix

                          Eigenvalue    Difference    Proportion    Cumulative

                     1    416.976349    398.666421        0.9501        0.9501
                     2     18.309928     14.952922        0.0417        0.9918
                     3      3.357006      3.126943        0.0076        0.9995
                     4      0.230063                      0.0005        1.0000


                  Root-Mean-Square Total-Sample Standard Deviation = 10.69224
                  Root-Mean-Square Distance Between Observations   = 30.24221


                                        Cluster History
                                                                                    T
                                                                                    i
     NCL    --Clusters Joined---      FREQ     SPRSQ     RSQ    ERSQ     CCC     PSF    PST2    e

       9    OB2         OB4             23    0.0019    .958    .932    6.26     400     6.3
       8    OB1         OB5             12    0.0025    .955    .926    6.75     434     5.8
       7    CL9         OB6             30    0.0069    .948    .918    6.28     438    19.5
       6    OB3         OB8             36    0.0074    .941    .907    6.21     459    26.0
       5    OB7         OB10            50    0.0104    .931    .892    6.15     485    42.2
       4    CL8         OB9             34    0.0162    .914    .870    4.28     519    39.3
       3    CL7         CL6             66    0.0318    .883    .824    4.39     552    59.7
       2    CL4         CL3            100    0.1099    .773    .695    3.94     503     113
       1    CL2         CL5            150    0.7726    .000    .000    0.00       .     503
```

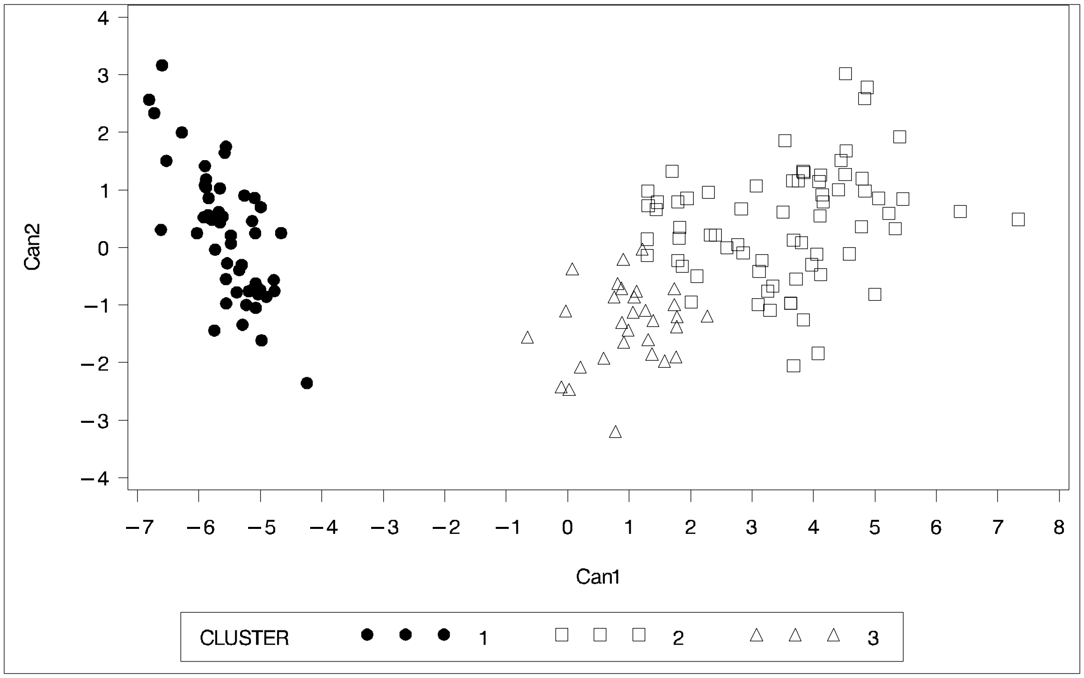
```
                    Cluster Analysis of Fisher (1936) Iris Data

                               The FREQ Procedure

                            Table of CLUSTER by Species

                    CLUSTER     Species

                    Frequency|
                    Percent  |
                    Row Pct  |
                    Col Pct  |Setosa  |Versicol|Virginic|  Total
                             |        |or      |a       |
                    ---------+--------+--------+--------+
                           1 |      0 |     50 |     16 |     66
                             |   0.00 |  33.33 |  10.67 |  44.00
                             |   0.00 |  75.76 |  24.24 |
                             |   0.00 | 100.00 |  32.00 |
                    ---------+--------+--------+--------+
                           2 |      0 |      0 |     34 |     34
                             |   0.00 |   0.00 |  22.67 |  22.67
                             |   0.00 |   0.00 | 100.00 |
                             |   0.00 |   0.00 |  68.00 |
                    ---------+--------+--------+--------+
                           3 |     50 |      0 |      0 |     50
                             |  33.33 |   0.00 |   0.00 |  33.33
                             | 100.00 |   0.00 |   0.00 |
                             | 100.00 |   0.00 |   0.00 |
                    ---------+--------+--------+--------+
                    Total          50       50       50      150
                                33.33    33.33    33.33   100.00
```



Output 23.3.5. Clustering Clusters: PROC CLUSTER with Wong’s Hybrid Method

```
                                   Cluster Analysis of Fisher (1936) Iris Data
                                   Clustering Clusters by Wong’s Hybrid Method

                                              The CLUSTER Procedure
                                       Two-Stage Density Linkage Clustering

                                       Eigenvalues of the Covariance Matrix

                                   Eigenvalue    Difference    Proportion    Cumulative

                              1    416.976349    398.666421        0.9501        0.9501
                              2     18.309928     14.952922        0.0417        0.9918
                              3      3.357006      3.126943        0.0076        0.9995
                              4      0.230063                      0.0005        1.0000
                           Root-Mean-Square Total-Sample Standard Deviation = 10.69224


                                                  Cluster History
                                                                                 Normalized     Maximum Density     T
                                                                                     Fusion     in Each Cluster     i
NCL   --Clusters Joined--     FREQ    SPRSQ    RSQ    ERSQ     CCC     PSF    PST2      Density     Lesser    Greater   e

  9   OB10       OB7            50   0.0104   .949   .932   3.81    330   42.2        40.24    58.2179      100.0
  8   OB3        OB8            36   0.0074   .942   .926   3.22    329   26.0       27.981    39.4511    48.4350
  7   OB2        OB4            23   0.0019   .940   .918   4.24    373    6.3       23.775     8.9675    46.3026
  6   CL8        OB9            58   0.0194   .921   .907   2.13    334   46.3       20.724    46.8846    48.4350
  5   CL7        OB6            30   0.0069   .914   .892   3.09    383   19.5       13.303    17.6360    46.3026
  4   CL6        OB1            67   0.0292   .884   .870   1.21    372   41.0       8.4137    10.8758    48.4350
  3   CL4        OB5            70   0.0138   .871   .824   3.33    494   12.3       5.1855     6.2890    48.4350
  2   CL3        CL5           100   0.0979   .773   .695   3.94    503   89.5       19.513    46.3026    48.4350
  1   CL2        CL9           150   0.7726   .000   .000   0.00     .     503       1.3337    48.4350      100.0
                                        3 modal clusters have been formed.
```

```
Cluster Analysis of Fisher (1936) Iris Data

             The FREQ Procedure

        Table of CLUSTER by Species

CLUSTER     Species

Frequency|
Percent  |
Row Pct  |
Col Pct  |Setosa  |Versicol|Virginic|  Total
         |        |or      |a       |
---------+--------+--------+--------+
       1 |     50 |      0 |      0 |     50
         |  33.33 |   0.00 |   0.00 |  33.33
         | 100.00 |   0.00 |   0.00 |
         | 100.00 |   0.00 |   0.00 |
---------+--------+--------+--------+
       2 |      0 |     21 |     49 |     70
         |   0.00 |  14.00 |  32.67 |  46.67
         |   0.00 |  30.00 |  70.00 |
         |   0.00 |  42.00 |  98.00 |
---------+--------+--------+--------+
       3 |      0 |     29 |      1 |     30
         |   0.00 |  19.33 |   0.67 |  20.00
         |   0.00 |  96.67 |   3.33 |
         |   0.00 |  58.00 |   2.00 |
---------+--------+--------+--------+
Total          50       50       50      150
            33.33    33.33    33.33   100.00
```



Example 23.4. Evaluating the Effects of Ties

If, at some level of the cluster history, there is a tie for minimum distance between clusters, then one or more levels of the sample cluster tree are not uniquely determined. This example shows how the degree of indeterminacy can be assessed.

Mammals have four kinds of teeth: incisors, canines, premolars, and molars. The following data set gives the number of teeth of each kind on one side of the top and bottom jaws for 32 mammals.

Since all eight variables are measured in the same units, it is not strictly necessary to rescale the data. However, the canines have much less variance than the other kinds of teeth and, therefore, have little effect on the analysis if the variables are not standardized. An average linkage cluster analysis is run with and without standardization to allow comparison of the results. The results are shown in Output 23.4.1 and Output 23.4.2.

```
title ’Hierarchical Cluster Analysis of Mammals’’ Teeth Data’;
title2 ’Evaluating the Effects of Ties’;
data teeth;
   input mammal $ 1-16
         @21 (v1-v8) (1.);
   label v1=’Top incisors’
         v2=’Bottom incisors’
         v3=’Top canines’
         v4=’Bottom canines’
         v5=’Top premolars’
         v6=’Bottom premolars’
         v7=’Top molars’
         v8=’Bottom molars’;
   datalines;
BROWN BAT           23113333
MOLE                32103333
SILVER HAIR BAT     23112333
PIGMY BAT           23112233
HOUSE BAT           23111233
RED BAT             13112233
PIKA                21002233
RABBIT              21003233
BEAVER              11002133
GROUNDHOG           11002133
GRAY SQUIRREL       11001133
HOUSE MOUSE         11000033
PORCUPINE           11001133
WOLF                33114423
BEAR                33114423
RACCOON             33114432
MARTEN              33114412
WEASEL              33113312
WOLVERINE           33114412
BADGER              33113312
RIVER OTTER         33114312
SEA OTTER           32113312
JAGUAR              33113211
COUGAR              33113211
FUR SEAL            32114411
SEA LION            32114411
GREY SEAL           32113322
ELEPHANT SEAL       21114411
REINDEER            04103333
ELK                 04103333
DEER                04003333
MOOSE               04003333
;

proc cluster data=teeth method=average nonorm
             outtree=_null_;
   var v1-v8;
   id mammal;
   title3 ’Raw Data’;
run;

proc cluster data=teeth std method=average nonorm
             outtree=_null_;
   var v1-v8;
   id mammal;
   title3 ’Standardized Data’;
run;
```

Output 23.4.1. Average Linkage Analysis of Mammals' Teeth Data: Raw Data

```
                  Hierarchical Cluster Analysis of Mammals’ Teeth Data
                              Evaluating the Effects of Ties
                                         Raw Data

                                  The CLUSTER Procedure
                             Average Linkage Cluster Analysis

                           Eigenvalues of the Covariance Matrix

                      Eigenvalue    Difference    Proportion    Cumulative

                 1    3.76799365    2.33557185        0.5840        0.5840
                 2    1.43242180    0.91781899        0.2220        0.8061
                 3    0.51460281    0.08414950        0.0798        0.8858
                 4    0.43045331    0.30021485        0.0667        0.9525
                 5    0.13023846    0.03814626        0.0202        0.9727
                 6    0.09209220    0.04216914        0.0143        0.9870
                 7    0.04992305    0.01603541        0.0077        0.9947
                 8    0.03388764                      0.0053        1.0000


             Root-Mean-Square Total-Sample Standard Deviation = 0.898027


                                      Cluster History
                                                                                T
                                                                          RMS   i
           NCL    ----------Clusters Joined-----------      FREQ      Dist   e

            31    BEAVER              GROUNDHOG                2         0   T
            30    GRAY SQUIRREL       PORCUPINE                2         0   T
            29    WOLF                BEAR                     2         0   T
            28    MARTEN              WOLVERINE                2         0   T
            27    WEASEL              BADGER                   2         0   T
            26    JAGUAR              COUGAR                   2         0   T
            25    FUR SEAL            SEA LION                 2         0   T
            24    REINDEER            ELK                      2         0   T
            23    DEER                MOOSE                    2         0
            22    BROWN BAT           SILVER HAIR BAT          2         1   T
            21    PIGMY BAT           HOUSE BAT                2         1   T
            20    PIKA                RABBIT                   2         1   T
            19    CL31                CL30                     4         1   T
            18    CL28                RIVER OTTER              3         1   T
            17    CL27                SEA OTTER                3         1   T
            16    CL24                CL23                     4         1
            15    CL21                RED BAT                  3    1.2247
            14    CL17                GREY SEAL                4     1.291
            13    CL29                RACCOON                  3    1.4142   T
            12    CL25                ELEPHANT SEAL            3    1.4142
            11    CL18                CL14                     7    1.5546
            10    CL22                CL15                     5    1.5811
             9    CL20                CL19                     6    1.8708   T
             8    CL11                CL26                     9    1.9272
             7    CL8                 CL12                    12    2.2278
             6    MOLE                CL13                     4    2.2361
             5    CL9                 HOUSE MOUSE              7    2.4833
             4    CL6                 CL7                     16    2.5658
             3    CL10                CL16                     9    2.8107
             2    CL3                 CL5                     16    3.7054
             1    CL2                 CL4                     32    4.2939
```

Output 23.4.2. Average Linkage Analysis of Mammals’ Teeth Data: Standardized Data

```
              Hierarchical Cluster Analysis of Mammals’ Teeth Data
                         Evaluating the Effects of Ties
                               Standardized Data

                             The CLUSTER Procedure
                        Average Linkage Cluster Analysis

                      Eigenvalues of the Correlation Matrix

                  Eigenvalue    Difference    Proportion    Cumulative

             1    4.74153902    3.27458808        0.5927        0.5927
             2    1.46695094    0.70824118        0.1834        0.7761
             3    0.75870977    0.25146252        0.0948        0.8709
             4    0.50724724    0.30264737        0.0634        0.9343
             5    0.20459987    0.05925818        0.0256        0.9599
             6    0.14534169    0.03450100        0.0182        0.9780
             7    0.11084070    0.04606994        0.0139        0.9919
             8    0.06477076                      0.0081        1.0000


          The data have been standardized to mean 0 and variance 1
          Root-Mean-Square Total-Sample Standard Deviation =        1


                                Cluster History
                                                                          T
                                                                   RMS    i
       NCL    ----------Clusters Joined-----------      FREQ      Dist    e

        31    BEAVER              GROUNDHOG                2         0    T
        30    GRAY SQUIRREL       PORCUPINE                2         0    T
        29    WOLF                BEAR                     2         0    T
        28    MARTEN              WOLVERINE                2         0    T
        27    WEASEL              BADGER                   2         0    T
        26    JAGUAR              COUGAR                   2         0    T
        25    FUR SEAL            SEA LION                 2         0    T
        24    REINDEER            ELK                      2         0    T
        23    DEER                MOOSE                    2         0
        22    PIGMY BAT           RED BAT                  2    0.9157
        21    CL28                RIVER OTTER              3    0.9169
        20    CL31                CL30                     4    0.9428    T
        19    BROWN BAT           SILVER HAIR BAT          2    0.9428    T
        18    PIKA                RABBIT                   2    0.9428
        17    CL27                SEA OTTER                3    0.9847
        16    CL22                HOUSE BAT                3    1.1437
        15    CL21                CL17                     6    1.3314
        14    CL25                ELEPHANT SEAL            3    1.3447
        13    CL19                CL16                     5    1.4688
        12    CL15                GREY SEAL                7    1.6314
        11    CL29                RACCOON                  3     1.692
        10    CL18                CL20                     6    1.7357
         9    CL12                CL26                     9    2.0285
         8    CL24                CL23                     4    2.1891
         7    CL9                 CL14                    12    2.2674
         6    CL10                HOUSE MOUSE              7     2.317
         5    CL11                CL7                     15    2.6484
         4    CL13                MOLE                     6    2.8624
         3    CL4                 CL8                     10    3.5194
         2    CL3                 CL6                     17    4.1265
         1    CL2                 CL5                     32    4.7753
```

There are ties at 16 levels for the raw data but at only 10 levels for the standardized data. There are more ties for the raw data because the increments between successive values are the same for all of the raw variables but different for the standardized variables.

One way to assess the importance of the ties in the analysis is to repeat the analysis on several random permutations of the observations and then to see to what extent the results are consistent at the interesting levels of the cluster history. Three macros are presented to facilitate this process.

```
/* --------------------------------------------------------- */
/*                                                           */
/* The macro CLUSPERM randomly permutes observations and     */
/* does a cluster analysis for each permutation.             */
/* The arguments are as follows:                             */
/*                                                           */
/*    data    data set name                                  */
/*    var     list of variables to cluster                   */
/*    id      id variable for proc cluster                   */
/*    method  clustering method (and possibly other options) */
/*    nperm   number of random permutations.                 */
/*                                                           */
/* --------------------------------------------------------- */
%macro CLUSPERM(data,var,id,method,nperm);
/* ------CREATE TEMPORARY DATA SET WITH RANDOM NUMBERS------ */
data _temp_;
   set &data;
   array _random_ _ran_1-_ran_&nperm;
   do over _random_;
     _random_=ranuni(835297461);
   end;
run;
/* ------PERMUTE AND CLUSTER THE DATA----------------------- */
%do n=1 %to &nperm;
    proc sort data=_temp_(keep=_ran_&n &var &id) out=_perm_;
       by _ran_&n;
    run;
    proc cluster method=&method noprint outtree=_tree_&n;
       var &var;
       id &id;
    run;
%end;
%mend;

/* --------------------------------------------------------- */
/*                                                           */
/* The macro PLOTPERM plots various cluster statistics       */
/* against the number of clusters for each permutation.      */
/* The arguments are as follows:                             */
/*                                                           */
/*    stats   names of variables from tree data set          */
/*    nclus   maximum number of clusters to be plotted       */
/*    nperm   number of random permutations.                 */
/*                                                           */
/* --------------------------------------------------------- */
%macro PLOTPERM(stat,nclus,nperm);
/* ---CONCATENATE TREE DATA SETS FOR 20 OR FEWER CLUSTERS--- */
data _plot_;
   set %do n=1 %to &nperm; _tree_&n(in=_in_&n) %end; ;
   if _ncl_<=&nclus;
   %do n=1 %to &nperm;
      if _in_&n then _perm_=&n;
   %end;
   label _perm_='permutation number';
   keep _ncl_ &stat _perm_;
run;
/* ---PLOT THE REQUESTED STATISTICS BY NUMBER OF CLUSTERS--- */

proc plot;
   plot (&stat)*_ncl_=_perm_ /vpos=26;
title2 'Symbol is value of _PERM_';
run;
%mend;

/* -------------------------------------------------------- */
/*                                                          */
/* The macro TREEPERM generates cluster-membership variables */
/* for a specified number of clusters for each permutation.  */
/* PROC PRINT lists the objects in each cluster-combination, */
/* and PROC TABULATE gives the frequencies and means.  The   */
/* arguments are as follows:                                */
/*                                                          */
/*    var     list of variables to cluster                  */
/*            (no "-" or ":" allowed)                       */
/*    id      id variable for proc cluster                  */
/*    meanfmt format for printing means in PROC TABULATE     */
/*    nclus   number of clusters desired                    */
/*    nperm   number of random permutations.                */
/*                                                          */
/* -------------------------------------------------------- */
%macro TREEPERM(var,id,meanfmt,nclus,nperm);
/* ------CREATE DATA SETS GIVING CLUSTER MEMBERSHIP--------- */
%do n=1 %to &nperm;
   proc tree data=_tree_&n noprint n=&nclus
             out=_out_&n(drop=clusname
                          rename=(cluster=_clus_&n));
      copy &var;
      id &id;
   run;
   proc sort;
      by &id &var;
   run;
%end;
/* ------MERGE THE CLUSTER VARIABLES----------------------- */
data _merge_;
   merge
      %do n=1 %to &nperm;
         _out_&n
      %end; ;
   by &id &var;
   length all_clus $ %eval(3*&nperm);
   %do n=1 %to &nperm;
      substr( all_clus, %eval(1+(&n-1)*3), 3) =
         put( _clus_&n, 3.);
   %end;
run;
```

```
/* ------PRINT AND TABULATE CLUSTER COMBINATIONS----------- */
proc sort;
   by _clus_:;
run;
proc print;
   var &var;
   id &id;
   by all_clus notsorted;
run;
proc tabulate order=data formchar='           ';
   class all_clus;
   var &var;
   table all_clus, n='FREQ'*f=5. mean*f=&meanfmt*(&var) /
      rts=%eval(&nperm*3+1);
run;
%mend;
```

To use these, it is first convenient to define a macro, VLIST, listing the teeth variables, since the forms V1-V8 or V: cannot be used with the TABULATE procedure in the TREEPERM macro:

```
/* -TABULATE does not accept hyphens or colons in VAR lists- */
%let vlist=v1 v2 v3 v4 v5 v6 v7 v8;
```

The CLUSPERM macro is then called to analyze ten random permutations. The PLOTPERM macro plots the pseudo F and t^2 statistics and the cubic clustering criterion. Since the data are discrete, the pseudo F statistic and the cubic clustering criterion can be expected to increase as the number of clusters increases, so local maxima or large jumps in these statistics are more relevant than the global maximum in determining the number of clusters. For the raw data, only the pseudo t^2 statistic indicates the possible presence of clusters, with the 4-cluster level being suggested. Hence, the TREEPERM macro is used to analyze the results at the 4-cluster level:

```
title3 'Raw Data';

/* ------CLUSTER RAW DATA WITH AVERAGE LINKAGE------------- */
%clusperm( teeth, &vlist, mammal, average, 10);

/* -----PLOT STATISTICS FOR THE LAST 20 LEVELS------------- */
%plotperm( _psf_ _pst2_ _ccc_, 20, 10);

/* ------ANALYZE THE 4-CLUSTER LEVEL----------------------- */
%treeperm( &vlist, mammal, 9.1, 4, 10);
```

The results are shown in Output 23.4.3.

Output 23.4.3. Analysis of Ten Random Permutations of Raw Mammals’ Teeth Data: Indeterminacy at the 4-Cluster Level

```
                                  Hierarchical Cluster Analysis of Mammals’ Teeth Data
                                               Symbol is value of _PERM_

                                    Plot of _PSF_*_NCL_.  Symbol is value of _perm_.

            |
        100 +
            |
            |
      P     |
      s     |
      e     |                                                                                                 5
      u  80 +
      d     |
      o     |
            |
      F     |
            |                                                                                                 2
      S  60 +
      t     |                                                                                            5    4
      a     |                                                                                            2
      t     |                                                                                       9    9    1
      i     |                                                                                  3    3    1    6
      s     |                                                                        2    2    1    1    4
      t  40 +                                                                   2    4    1
      i     |                                                         1    1    1    1
      c     |                                                    2    3
            |            1                   2         2    1    1
            |       1         1    1    1         1    1
            |            2                   1
         20 +
            ---+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+--
               1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20

                                                      Number of Clusters

NOTE: 10 obs had missing values.  151 obs hidden.
```

```
                                  Hierarchical Cluster Analysis of Mammals’ Teeth Data
                                               Symbol is value of _PERM_

                                   Plot of _PST2_*_NCL_.  Symbol is value of _perm_.

       P    |
       s 30 +
       e    |
       u    |  1
       d    |
       o 25 +
            |
       T    |
       -    |            1
       S 20 +
       q    |       1
       u    |
       a    |
       r 15 +
       e    |
       d    |            2
            |       2                                  1
       S 10 +
       t    |
       a    |                 2                        2                         3
       t    |                 1    2    2    1    1              1
       i  5 +                           1                                       2    5
       s    |                      1                        2    3              4    1
       t    |                                               1
       i    |                                                                   1    2    1
       c  0 +
            ---+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+--
               1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20

                                                      Number of Clusters

NOTE: 69 obs had missing values.  104 obs hidden.
```

```
                                  Hierarchical Cluster Analysis of Mammals’ Teeth Data
                                               Symbol is value of _PERM_

                                    Plot of _CCC_*_NCL_.  Symbol is value of _perm_.

C   |
u 4 +
b   |
i   |
c   |
    |                              2
C   |
l 3 +
u   |                        1     1
s   |
t   |
e   |
r   |                        2
i 2 +
n   |                  1
g   |            1     2
    |
C   |
r   |
i 1 +
t   |
e   |
r   |            2
i   |
o   |      1
n 0 +1
    -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20

                                                     Number of Clusters

NOTE: 140 obs had missing values.  50 obs hidden.
```

```
-------------------------------------- all_clus='  1  3  1  1  1  3  3  3  2  3' --------------------------------------

                                mammal      v1    v2    v3    v4    v5    v6    v7    v8

                                DEER         0     4     0     0     3     3     3     3
                                ELK          0     4     1     0     3     3     3     3
                                MOOSE        0     4     0     0     3     3     3     3
                                REINDEER     0     4     1     0     3     3     3     3


-------------------------------------- all_clus='  2  2  2  2  2  2  1  2  1  1' --------------------------------------

                             mammal           v1    v2    v3    v4    v5    v6    v7    v8

                             BADGER            3     3     1     1     3     3     1     2
                             BEAR              3     3     1     1     4     4     2     3
                             COUGAR            3     3     1     1     3     2     1     1
                             ELEPHANT SEAL     2     1     1     1     4     4     1     1
                             FUR SEAL          3     2     1     1     4     4     1     1
                             GREY SEAL         3     2     1     1     3     3     2     2
                             JAGUAR            3     3     1     1     3     2     1     1
                             MARTEN            3     3     1     1     4     4     1     2
                             RACCOON           3     3     1     1     4     4     3     2
                             RIVER OTTER       3     3     1     1     4     3     1     2
                             SEA LION          3     2     1     1     4     4     1     1
                             SEA OTTER         3     2     1     1     3     3     1     2
                             WEASEL            3     3     1     1     3     3     1     2
                             WOLF              3     3     1     1     4     4     2     3
                             WOLVERINE         3     3     1     1     4     4     1     2


-------------------------------------- all_clus='  2  4  2  2  4  2  1  2  1  1' --------------------------------------

                                 mammal    v1    v2    v3    v4    v5    v6    v7    v8

                                  MOLE      3     2     1     0     3     3     3     3


-------------------------------------- all_clus='  3  1  3  3  3  1  2  1  3  2' --------------------------------------

                             mammal           v1    v2    v3    v4    v5    v6    v7    v8

                             BEAVER            1     1     0     0     2     1     3     3
                             GRAY SQUIRREL     1     1     0     0     1     1     3     3
                             GROUNDHOG         1     1     0     0     2     1     3     3
                             HOUSE MOUSE       1     1     0     0     0     0     3     3
                             PORCUPINE         1     1     0     0     1     1     3     3


-------------------------------------- all_clus='  3  4  3  3  4  1  2  1  3  2' --------------------------------------

                                 mammal    v1    v2    v3    v4    v5    v6    v7    v8

                                 PIKA       2     1     0     0     2     2     3     3
                                 RABBIT     2     1     0     0     3     2     3     3


-------------------------------------- all_clus='  4  4  4  4  4  4  4  4  4  4' --------------------------------------

                            mammal             v1    v2    v3    v4    v5    v6    v7    v8

                            BROWN BAT           2     3     1     1     3     3     3     3
                            HOUSE BAT           2     3     1     1     1     2     3     3
                            PIGMY BAT           2     3     1     1     2     2     3     3
                            RED BAT             1     3     1     1     2     2     3     3
                            SILVER HAIR BAT     2     3     1     1     2     3     3     3
```

```
                                                                    Mean

                                        Top     Bottom       Top     Bottom      Top     Bottom      Top      Bottom
                                FREQ  incisors  incisors   canines   canines  premolars premolars   molars    molars

all_clus

1  3  1  1  1  3  3  3  2  3       4       0.0       4.0       0.5       0.0       3.0       3.0       3.0       3.0

2  2  2  2  2  2  1  2  1  1      15       2.9       2.6       1.0       1.0       3.6       3.4       1.3       1.8

2  4  2  2  4  2  1  2  1  1       1       3.0       2.0       1.0       0.0       3.0       3.0       3.0       3.0

3  1  3  3  3  1  2  1  3  2       5       1.0       1.0       0.0       0.0       1.2       0.8       3.0       3.0

3  4  3  3  4  1  2  1  3  2       2       2.0       1.0       0.0       0.0       2.5       2.0       3.0       3.0

4  4  4  4  4  4  4  4  4  4       5       1.8       3.0       1.0       1.0       2.0       2.4       3.0       3.0
```

From the TABULATE and PRINT output, you can see that two types of clustering are obtained. In one case, the mole is grouped with the carnivores, while the pika and rabbit are grouped with the rodents. In the other case, both the mole and the lagomorphs are grouped with the bats.

Next, the analysis is repeated with the standardized data. The pseudo F and t^2 statistics indicate 3 or 4 clusters, while the cubic clustering criterion shows a sharp rise up to 4 clusters and then levels off up to 6 clusters. So the TREEPERM macro is used again at the 4-cluster level. In this case, there is no indeterminacy, as the same four clusters are obtained with every permutation, although in different orders. It must be emphasized, however, that lack of indeterminacy in no way indicates validity. The results are shown in Output 23.4.4.

```
title3 'Standardized Data';

/*------CLUSTER STANDARDIZED DATA WITH AVERAGE LINKAGE------*/
%clusperm( teeth, &vlist, mammal, average std, 10);

/*------PLOT STATISTICS FOR THE LAST 20 LEVELS--------------*/
%plotperm( _psf_ _pst2_ _ccc_, 20, 10);

/*------ANALYZE THE 4-CLUSTER LEVEL-------------------------*/
%treeperm( &vlist, mammal, 9.1, 4, 10);
```

Output 23.4.4. Analysis of Ten Random Permutations of Standardized Mammals' Teeth Data: No Indeterminacy at the 4-Cluster Level

```
                                  Hierarchical Cluster Analysis of Mammals' Teeth Data
                                               Symbol is value of _PERM_

                                    Plot of _PSF_*_NCL_.  Symbol is value of _perm_.

            |
        100 +
            |
            |
      P     |
      s     |
      e     |                                                                                                 1
      u  80 +
      d     |                                                                                            1
      o     |                                                                                       1
            |
      F     |                                                                                  1
            |                                                                             1
      S  60 +
      t     |
      a     |
      t     |
      i     |                                                                   1    1
      s     |
      t  40 +                                                         1    1
      i     |                                                    1
      c     |                 1                             1
            |            1                        1    1
            |       1              1    1    1
            |
         20 +
            ---+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+--
               1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20

                                                      Number of Clusters

NOTE: 10 obs had missing values.  171 obs hidden.
```

```
                                  Hierarchical Cluster Analysis of Mammals’ Teeth Data
                                               Symbol is value of _PERM_

                                   Plot of _PST2_*_NCL_.  Symbol is value of _perm_.

       P    |
       s    |
       e    |
       u    |
       d    |
       o    |
         30 +
       T    |  1
       -    |
       S    |
       q    |
       u    |
       a 20 +
       r    |
       e    |       1
       d    |
            |            1
       S    |
       t 10 +                                               1
       a    |
       t    |                 1    1         1         1                             1
       i    |
       s    |                            1                             1    1
       t    |                                                                             1
       i  0 +
       c    |
            ---+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+----+--
               1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20

                                                      Number of Clusters

NOTE: 70 obs had missing values.  117 obs hidden.
```

```
                                  Hierarchical Cluster Analysis of Mammals’ Teeth Data
                                               Symbol is value of _PERM_

                                    Plot of _CCC_*_NCL_.  Symbol is value of _perm_.

C   |
u 4 +
b   |                  1           1
i   |
c   |
    |
C   |
l 3 +                        1
u   |
s   |
t   |
e   |
r   |            1
i 2 +
n   |
g   |
    |
C   |
r   |
i 1 +
t   |
e   |
r   |
i   |
o   |      1
n 0 +1
    -+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
     1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16    17    18    19    20

                                                     Number of Clusters

NOTE: 140 obs had missing values.  54 obs hidden.
```

```
------------------------------------- all_clus=' 1  3  1  1  1  3  3  3  2  3' -------------------------------------

                               mammal      v1    v2    v3    v4    v5    v6    v7    v8

                               DEER         0     4     0     0     3     3     3     3
                               ELK          0     4     1     0     3     3     3     3
                               MOOSE        0     4     0     0     3     3     3     3
                               REINDEER     0     4     1     0     3     3     3     3


------------------------------------- all_clus=' 2  2  2  2  2  2  1  2  1  1' -------------------------------------

                            mammal           v1    v2    v3    v4    v5    v6    v7    v8

                            BADGER            3     3     1     1     3     3     1     2
                            BEAR              3     3     1     1     4     4     2     3
                            COUGAR            3     3     1     1     3     2     1     1
                            ELEPHANT SEAL     2     1     1     1     4     4     1     1
                            FUR SEAL          3     2     1     1     4     4     1     1
                            GREY SEAL         3     2     1     1     3     3     2     2
                            JAGUAR            3     3     1     1     3     2     1     1
                            MARTEN            3     3     1     1     4     4     1     2
                            RACCOON           3     3     1     1     4     4     3     2
                            RIVER OTTER       3     3     1     1     4     3     1     2
                            SEA LION          3     2     1     1     4     4     1     1
                            SEA OTTER         3     2     1     1     3     3     1     2
                            WEASEL            3     3     1     1     3     3     1     2
                            WOLF              3     3     1     1     4     4     2     3
                            WOLVERINE         3     3     1     1     4     4     1     2


------------------------------------- all_clus=' 3  1  3  3  3  1  2  1  3  2' -------------------------------------

                            mammal           v1    v2    v3    v4    v5    v6    v7    v8

                            BEAVER            1     1     0     0     2     1     3     3
                            GRAY SQUIRREL     1     1     0     0     1     1     3     3
                            GROUNDHOG         1     1     0     0     2     1     3     3
                            HOUSE MOUSE       1     1     0     0     0     0     3     3
                            PIKA              2     1     0     0     2     2     3     3
                            PORCUPINE         1     1     0     0     1     1     3     3
                            RABBIT            2     1     0     0     3     2     3     3


------------------------------------- all_clus=' 4  4  4  4  4  4  4  4  4  4' -------------------------------------

                           mammal             v1    v2    v3    v4    v5    v6    v7    v8

                           BROWN BAT           2     3     1     1     3     3     3     3
                           HOUSE BAT           2     3     1     1     1     2     3     3
                           MOLE                3     2     1     0     3     3     3     3
                           PIGMY BAT           2     3     1     1     2     2     3     3
                           RED BAT             1     3     1     1     2     2     3     3
                           SILVER HAIR BAT     2     3     1     1     2     3     3     3



                                                                                       Mean

                                              Top     Bottom      Top     Bottom      Top    Bottom      Top     Bottom
                                 FREQ  incisors  incisors    canines   canines  premolars premolars  molars    molars

 all_clus

 1  3  1  1  1  3  3  3  2  3      4       0.0       4.0       0.5       0.0       3.0       3.0       3.0       3.0

 2  2  2  2  2  2  1  2  1  1     15       2.9       2.6       1.0       1.0       3.6       3.4       1.3       1.8

 3  1  3  3  3  1  2  1  3  2      7       1.3       1.0       0.0       0.0       1.6       1.1       3.0       3.0

 4  4  4  4  4  4  4  4  4  4      6       2.0       2.8       1.0       0.8       2.2       2.5       3.0       3.0
```

Example 23.5. Computing a Distance Matrix

A wide variety of distance and similarity measures are used in cluster analysis (Anderberg 1973, Sneath and Sokal 1973). If your data are in coordinate form and you want to use a non-Euclidean distance for clustering, you can compute a distance matrix using a DATA step or the IML procedure.

Similarity measures must be converted to dissimilarities before being used in PROC CLUSTER. Such conversion can be done in a variety of ways, such as taking reciprocals or subtracting from a large value. The choice of conversion method depends on the application and the similarity measure.

In the following example, the observations are states. Binary-valued variables correspond to various grounds for divorce and indicate whether the grounds for divorce apply in each of the states.

The %DISTANCE[*] macro is used to compute the Jaccard coefficient (Anderberg 1973, pp. 89, 115, and 117) between each pair of states. The Jaccard coefficient is defined as the number of variables that are coded as 1 for both states divided by the number of variables that are coded as 1 for either or both states. The Jaccard coefficient is converted to a distance measure by subtracting it from 1.

```
%include  ’<location of  SAS/STAT sample library>/xmacro.sas’;
%include  ’<location of  SAS/STAT sample library>/distnew.sas’;

options ls=120 ps=60;
data divorce;
   title ’Grounds for Divorce’;
   input state $15.
         (incompat cruelty desertn non_supp alcohol
          felony impotenc insanity separate) (1.) @@;
   if mod(_n_,2) then input +4 @@; else input;
   datalines;
ALABAMA        111111111    ALASKA         111011110
ARIZONA        100000000    ARKANSAS       011111111
CALIFORNIA     100000010    COLORADO       100000000
CONNECTICUT    111111011    DELAWARE       100000001
FLORIDA        100000010    GEORGIA        111011110
HAWAII         100000001    IDAHO          111111011
ILLINOIS       011011100    INDIANA        100001110
IOWA           100000000    KANSAS         111011110
KENTUCKY       100000000    LOUISIANA      000001001
MAINE          111110110    MARYLAND       011001111
MASSACHUSETTS  111111101    MICHIGAN       100000000
MINNESOTA      100000000    MISSISSIPPI    111011110
MISSOURI       100000000    MONTANA        100000000
```

[*] The %DISTANCE macro computes various measures of distance, dissimilarity, or similarity between the observations (rows) of a SAS data set. These proximity measures are stored as a lower triangular matrix or a square matrix in an output data set that can then be used as input to the CLUSTER, MDS or MODECLUS procedures. The input data sets may contain numeric or character variables or both, depending on which proximity measure is used. The macro is documented in the macro comments and can be found in the SAS/STAT sample library.

```
NEBRASKA       100000000    NEVADA         100000011
NEW HAMPSHIRE  111111100    NEW JERSEY     011011011
NEW MEXICO     111000000    NEW YORK       011001001
NORTH CAROLINA 000000111    NORTH DAKOTA   111111110
OHIO           111011101    OKLAHOMA       111111110
OREGON         100000000    PENNSYLVANIA   011001110
RHODE ISLAND   111111101    SOUTH CAROLINA 011010001
SOUTH DAKOTA   011111000    TENNESSEE      111111100
TEXAS          111001011    UTAH           011111110
VERMONT        011101011    VIRGINIA       010001001
WASHINGTON     100000001    WEST VIRGINIA  111011011
WISCONSIN      100000001    WYOMING        100000011
;

%distance(data=divorce, id=state, options=nomiss, out=distjacc,
          shape=square, method=djaccard, var=incompat--separate);

proc print data=distjacc(obs=10);
   id state; var alabama--georgia;
   title2 ’First 10 states’;
run;
title2;

proc cluster data=distjacc method=centroid
             pseudo outtree=tree;
   id state;
   var alabama--wyoming;
run;

proc tree data=tree noprint n=9 out=out;
   id state;
run;

proc sort;
   by state;
run;

data clus;
   merge divorce out;
   by state;
run;

proc sort;
   by cluster;
run;

proc print;
   id state;
   var incompat--separate;
   by cluster;
run;
```

Output 23.5.1. Computing a Distance Matrix

```
                                         Grounds for Divorce
                                           First 10 states

state          ALABAMA   ALASKA  ARIZONA  ARKANSAS  CALIFORNIA  COLORADO  CONNECTICUT  DELAWARE  FLORIDA  GEORGIA

ALABAMA        0.00000  0.22222  0.88889   0.11111     0.77778   0.88889      0.11111   0.77778  0.77778  0.22222
ALASKA         0.22222  0.00000  0.85714   0.33333     0.71429   0.85714      0.33333   0.87500  0.71429  0.00000
ARIZONA        0.88889  0.85714  0.00000   1.00000     0.50000   0.00000      0.87500   0.50000  0.50000  0.85714
ARKANSAS       0.11111  0.33333  1.00000   0.00000     0.88889   1.00000      0.22222   0.88889  0.88889  0.33333
CALIFORNIA     0.77778  0.71429  0.50000   0.88889     0.00000   0.50000      0.75000   0.66667  0.00000  0.71429
COLORADO       0.88889  0.85714  0.00000   1.00000     0.50000   0.00000      0.87500   0.50000  0.50000  0.85714
CONNECTICUT    0.11111  0.33333  0.87500   0.22222     0.75000   0.87500      0.00000   0.75000  0.75000  0.33333
DELAWARE       0.77778  0.87500  0.50000   0.88889     0.66667   0.50000      0.75000   0.00000  0.66667  0.87500
FLORIDA        0.77778  0.71429  0.50000   0.88889     0.00000   0.50000      0.75000   0.66667  0.00000  0.71429
GEORGIA        0.22222  0.00000  0.85714   0.33333     0.71429   0.85714      0.33333   0.87500  0.71429  0.00000
```

```
                              Grounds for Divorce

                             The CLUSTER Procedure
                     Centroid Hierarchical Cluster Analysis

           Root-Mean-Square Distance Between Observations   = 0.694873


                                Cluster History
                                                                    Norm    T
                                                                    Cent    i
    NCL    ---------Clusters Joined----------      FREQ     PSF    PST2      Dist    e

     49    ARIZONA           COLORADO                 2       .       .         0    T
     48    CALIFORNIA        FLORIDA                  2       .       .         0    T
     47    ALASKA            GEORGIA                  2       .       .         0    T
     46    DELAWARE          HAWAII                   2       .       .         0    T
     45    CONNECTICUT       IDAHO                    2       .       .         0    T
     44    CL49              IOWA                     3       .       .         0    T
     43    CL47              KANSAS                   3       .       .         0    T
     42    CL44              KENTUCKY                 4       .       .         0    T
     41    CL42              MICHIGAN                 5       .       .         0    T
     40    CL41              MINNESOTA                6       .       .         0    T
     39    CL43              MISSISSIPPI              4       .       .         0    T
     38    CL40              MISSOURI                 7       .       .         0    T
     37    CL38              MONTANA                  8       .       .         0    T
     36    CL37              NEBRASKA                 9       .       .         0    T
     35    NORTH DAKOTA      OKLAHOMA                 2       .       .         0    T
     34    CL36              OREGON                  10       .       .         0    T
     33    MASSACHUSETTS     RHODE ISLAND             2       .       .         0    T
     32    NEW HAMPSHIRE     TENNESSEE                2       .       .         0    T
     31    CL46              WASHINGTON               3       .       .         0    T
     30    CL31              WISCONSIN                4       .       .         0    T
     29    NEVADA            WYOMING                  2       .       .         0
     28    ALABAMA           ARKANSAS                 2    1561       .    0.1599    T
     27    CL33              CL32                     4     479       .    0.1799    T
     26    CL39              CL35                     6     265       .    0.1799    T
     25    CL45              WEST VIRGINIA            3     231       .    0.1799
     24    MARYLAND          PENNSYLVANIA             2     199       .    0.2399
     23    CL28              UTAH                     3     167     3.2    0.2468
     22    CL27              OHIO                     5     136     5.4    0.2698
     21    CL26              MAINE                    7     111     8.9    0.2998
     20    CL23              CL21                    10    75.2     8.7    0.3004
     19    CL25              NEW JERSEY               4    71.8     6.5    0.3053    T
     18    CL19              TEXAS                    5    69.1     2.5    0.3077
     17    CL20              CL22                    15    48.7     9.9    0.3219
     16    NEW YORK          VIRGINIA                 2    50.1       .    0.3598
     15    CL18              VERMONT                  6    49.4     2.9    0.3797
     14    CL17              ILLINOIS                16    47.0     3.2    0.4425
     13    CL14              CL15                    22    29.2    15.3    0.4722
     12    CL48              CL29                     4    29.5       .    0.4797    T
     11    CL13              CL24                    24    27.6     4.5    0.5042
     10    CL11              SOUTH DAKOTA            25    28.4     2.4    0.5449
      9    LOUISIANA         CL16                     3    30.3     3.5    0.5844
      8    CL34              CL30                    14    23.3       .    0.7196
      7    CL8               CL12                    18    19.3    15.0    0.7175
      6    CL10              SOUTH CAROLINA          26    21.4     4.2    0.7384
      5    CL6               NEW MEXICO              27    24.0     4.7    0.8303
      4    CL5               INDIANA                 28    28.9     4.1    0.8343
      3    CL4               CL9                     31    31.7    10.9    0.8472
      2    CL3               NORTH CAROLINA          32    55.1     4.1    1.0017
      1    CL2               CL7                     50       .    55.1    1.0663
```

```
                                           Grounds for Divorce

------------------------------------------------------ CLUSTER=1 -------------------------------------------------------

   state         incompat     cruelty    desertn    non_supp     alcohol    felony    impotenc    insanity    separate

   ARIZONA           1           0          0           0           0          0          0           0           0
   COLORADO          1           0          0           0           0          0          0           0           0
   IOWA              1           0          0           0           0          0          0           0           0
   KENTUCKY          1           0          0           0           0          0          0           0           0
   MICHIGAN          1           0          0           0           0          0          0           0           0
   MINNESOTA         1           0          0           0           0          0          0           0           0
   MISSOURI          1           0          0           0           0          0          0           0           0
   MONTANA           1           0          0           0           0          0          0           0           0
   NEBRASKA          1           0          0           0           0          0          0           0           0
   OREGON            1           0          0           0           0          0          0           0           0


------------------------------------------------------ CLUSTER=2 -------------------------------------------------------

  state          incompat     cruelty    desertn    non_supp     alcohol    felony    impotenc    insanity    separate

  CALIFORNIA         1           0          0           0           0          0          0           1           0
  FLORIDA            1           0          0           0           0          0          0           1           0
  NEVADA             1           0          0           0           0          0          0           1           1
  WYOMING            1           0          0           0           0          0          0           1           1
```

```
------------------------------------------------------ CLUSTER=3 -------------------------------------------------------

 state              incompat     cruelty    desertn    non_supp     alcohol    felony    impotenc    insanity    separate

 ALABAMA                1           1          1           1           1          1          1           1           1
 ALASKA                 1           1          1           0           1          1          1           1           0
 ARKANSAS               0           1          1           1           1          1          1           1           1
 CONNECTICUT            1           1          1           1           1          1          0           1           1
 GEORGIA                1           1          1           0           1          1          1           1           0
 IDAHO                  1           1          1           1           1          1          0           1           1
 ILLINOIS               0           1          1           0           1          1          1           0           0
 KANSAS                 1           1          1           0           1          1          1           1           0
 MAINE                  1           1          1           1           1          0          1           1           0
 MARYLAND               0           1          1           0           0          1          1           1           1
 MASSACHUSETTS          1           1          1           1           1          1          1           0           1
 MISSISSIPPI            1           1          1           0           1          1          1           1           0
 NEW HAMPSHIRE          1           1          1           1           1          1          1           0           0
 NEW JERSEY             0           1          1           0           1          1          0           1           1
 NORTH DAKOTA           1           1          1           1           1          1          1           1           0
 OHIO                   1           1          1           0           1          1          1           0           1
 OKLAHOMA               1           1          1           1           1          1          1           1           0
 PENNSYLVANIA           0           1          1           0           0          1          1           1           0
 RHODE ISLAND           1           1          1           1           1          1          1           0           1
 SOUTH DAKOTA           0           1          1           1           1          1          0           0           0
 TENNESSEE              1           1          1           1           1          1          1           0           0
 TEXAS                  1           1          1           0           0          1          0           1           1
 UTAH                   0           1          1           1           1          1          1           1           0
 VERMONT                0           1          1           1           0          1          0           1           1
 WEST VIRGINIA          1           1          1           0           1          1          0           1           1


------------------------------------------------------ CLUSTER=4 -------------------------------------------------------

  state          incompat     cruelty    desertn    non_supp     alcohol    felony    impotenc    insanity    separate

  DELAWARE           1           0          0           0           0          0          0           0           1
  HAWAII             1           0          0           0           0          0          0           0           1
  WASHINGTON         1           0          0           0           0          0          0           0           1
  WISCONSIN          1           0          0           0           0          0          0           0           1
```

```
------------------------------------------------------ CLUSTER=5 ------------------------------------------------------

     state      incompat    cruelty    desertn    non_supp    alcohol    felony    impotenc    insanity    separate

   LOUISIANA        0          0          0           0          0          1          0           0           1
   NEW YORK         0          1          1           0          0          1          0           0           1
   VIRGINIA         0          1          0           0          0          1          0           0           1


------------------------------------------------------ CLUSTER=6 ------------------------------------------------------

    state          incompat    cruelty    desertn    non_supp    alcohol    felony    impotenc    insanity    separate

 SOUTH CAROLINA        0          1          1           0          1          0          0           0           1


------------------------------------------------------ CLUSTER=7 ------------------------------------------------------

    state       incompat    cruelty    desertn    non_supp    alcohol    felony    impotenc    insanity    separate

  NEW MEXICO        1          1          1           0          0          0          0           0           0


------------------------------------------------------ CLUSTER=8 ------------------------------------------------------

     state     incompat    cruelty    desertn    non_supp    alcohol    felony    impotenc    insanity    separate

    INDIANA        1          0          0           0          0          1          1           1           0


------------------------------------------------------ CLUSTER=9 ------------------------------------------------------

    state          incompat    cruelty    desertn    non_supp    alcohol    felony    impotenc    insanity    separate

 NORTH CAROLINA        0          0          0           0          0          0          1           1           1
```

Example 23.6. Size, Shape, and Correlation

The following example shows the analysis of a data set in which size information is detrimental to the classification. Imagine that an archaeologist of the future is excavating a 20th century grocery store. The archaeologist has discovered a large number of boxes of various sizes, shapes, and colors and wants to do a preliminary classification based on simple external measurements: height, width, depth, weight, and the predominant color of the box. It is known that a given product may have been sold in packages of different size, so the archaeologist wants to remove the effect of size from the classification. It is not known whether color is relevant to the use of the products, so the analysis should be done both with and without color information.

Unknown to the archaeologist, the boxes actually fall into six general categories according to the use of the product: breakfast cereals, crackers, laundry detergents, Little Debbie snacks, tea, and toothpaste. These categories are shown in the analysis so that you can evaluate the effectiveness of the classification.

Since there is no reason for the archaeologist to assume that the true categories have equal sample sizes or variances, the centroid method is used to avoid undue bias. Each analysis is done with Euclidean distances after suitable transformations of the data. Color is coded as five dummy variables with values of 0 or 1. The DATA step is as follows:

```
options ls=120;
title ’Cluster Analysis of Grocery Boxes’;
data grocery2;
```

```
   length name $35   /* name of product */
          class $16  /* category of product */
          unit $1    /* unit of measurement for weights:
                           g=gram
                           o=ounce
                           l=lb
                        all weights are converted to grams */
          color $8   /* predominant color of box */
          height 8   /* height of box in cm. */
          width 8    /* width of box in cm. */
          depth 8    /* depth of box (front to back) in cm. */
          weight 8   /* weight of box in grams */
          c_white c_yellow c_red c_green c_blue 4;
                     /* dummy variables */
   retain class;
   drop unit;

   /*--- read name with possible embedded blanks ---*/
   input name & @;

   /*--- if name starts with "---",              ---*/
   /*--- it’s really a category value            ---*/
   if substr(name,1,3) = ’---’ then do;
      class = substr(name,4,index(substr(name,4),’-’)-1);
      delete;
      return;
   end;

   /*--- read the rest of the variables ---*/
   input height width depth weight unit color;

   /*--- convert weights to grams ---*/
   select (unit);
      when (’l’) weight = weight * 454;
      when (’o’) weight = weight * 28.3;
      when (’g’) ;
      otherwise put ’Invalid unit ’ unit;
   end;

   /*--- use 0/1 coding for dummy variables for colors ---*/
   c_white  = (color = ’w’);
   c_yellow = (color = ’y’);
   c_red    = (color = ’r’);
   c_green  = (color = ’g’);
   c_blue   = (color = ’b’);

datalines;

---Breakfast cereals---

Cheerios                            32.5 22.4  8.4  567 g y
Cheerios                            30.3 20.4  7.2  425 g y
Cheerios                            27.5 19    6.2  283 g y
Cheerios                            24.1 17.2  5.3  198 g y
```

```
Special K                           30.1 20.5  8.5   18 o w
Special K                           29.6 19.2  6.7   12 o w
Special K                           23.4 16.6  5.7    7 o w
Corn Flakes                         33.7 25.4  8     24 o w
Corn Flakes                         30.2 20.6  8.4   18 o w
Corn Flakes                         30   19.1  6.6   12 o w
Grape Nuts                          21.7 16.3  4.9  680 g w
Shredded Wheat                      19.7 19.9  7.5  283 g y
Shredded Wheat, Spoon Size          26.6 19.6  5.6  510 g r
All-Bran                            21.1 14.3  5.2 13.8 o y
Froot Loops                         30.2 20.8  8.5 19.7 o r
Froot Loops                         25   17.7  6.4   11 o r

---Crackers---

Wheatsworth                         11.1 25.2  5.5  326 g w
Ritz                                23.1 16    5.3  340 g r
Ritz                                23.1 20.7  5.2  454 g r
Premium Saltines                    11   25   10.7  454 g w
Waverly Wafers                      14.4 22.5  6.2  454 g g

---Detergent---

Arm & Hammer Detergent              38.8 30   16.9   25 l y
Arm & Hammer Detergent              39.5 25.8 11   14.2 l y
Arm & Hammer Detergent              33.7 22.8  7      7 l y
Arm & Hammer Detergent              27.8 19.4  6.3    4 l y
Tide                                39.4 24.8 11.3  9.2 l r
Tide                                32.5 23.2  7.3  4.5 l r
Tide                                26.5 19.9  6.3   42 o r
Tide                                19.3 14.6  4.7   17 o r

---Little Debbie---

Figaroos                            13.5 18.6  3.7   12 o y
Swiss Cake Rolls                    10.1 21.8  5.8   13 o w
Fudge Brownies                      11   30.8  2.5   12 o w
Marshmallow Supremes                 9.4 32    7     10 o w
Apple Delights                      11.2 30.1  4.9   15 o w
Snack Cakes                         13.4 32    3.4   13 o b
Nutty Bar                           13.2 18.5  4.2   12 o y
Lemon Stix                          13.2 18.5  4.2    9 o w
Fudge Rounds                         8.1 28.3  5.4  9.5 o w

---Tea---

Celestial Saesonings Mint Magic      7.8 13.8  6.3   49 g b
Celestial Saesonings Cranberry Cove  7.8 13.8  6.3   46 g r
Celestial Saesonings Sleepy Time     7.8 13.8  6.3   37 g g
Celestial Saesonings Lemon Zinger    7.8 13.8  6.3   56 g y
Bigelow Lemon Lift                   7.7 13.4  6.9   40 g y
Bigelow Plantation Mint              7.7 13.4  6.9   35 g g
Bigelow Earl Grey                    7.7 13.4  6.9   35 g b
Luzianne                             8.9 22.8  6.4    6 o r
Luzianne                            18.4 20.2  6.9    8 o r
Luzianne Decaffeinated               8.9 22.8  6.4 5.25 o g
Lipton Tea Bags                     17.1 20    6.7    8 o r
Lipton Tea Bags                     11.5 14.4  6.6 3.75 o r
Lipton Tea Bags                      6.7 10    5.7 1.25 o r
Lipton Family Size Tea Bags         13.7 24    9     12 o r
Lipton Family Size Tea Bags          8.7 20.8  8.2    6 o r
Lipton Family Size Tea Bags          8.9 11.1  8.2    3 o r
Lipton Loose Tea                    12.7 10.9  5.4    8 o r

---Paste, Tooth---

Colgate                              4.4 22    3.5    7 o r
Colgate                              3.6 15.6  3.3    3 o r
Colgate                              4.2 18.3  3.5    5 o r
Crest                                4.3 21.7  3.7  6.4 o w
Crest                                4.3 17.4  3.6  4.6 o w
Crest                                3.5 15.2  3.2  2.7 o w
Crest                                3.0 10.9  2.8  .85 o w
Arm & Hammer                         4.4 17    3.7    5 o w
;


data grocery;
   length name $16;
   set grocery2;
```

The FORMAT procedure is used to define to formats to make the output easier to read. The STARS. format is used for graphical crosstabulations in the TABULATE procedure. The $COLOR format displays the names of the colors instead of just the first letter.

```
    /*------ formats and macros for displaying ------*/
    /*------ cluster results                   ------*/
proc format; value stars
      0=’               ’
      1=’              #’
      2=’             ##’
      3=’            ###’
      4=’           ####’
      5=’          #####’
      6=’         ######’
      7=’        #######’
      8=’       ########’
      9=’      #########’
     10=’     ##########’
     11=’    ###########’
     12=’   ############’
     13=’  #############’
     14=’ ##############’
15-high=’>##############’;
run;
```

```
proc format; value $color
   ’w’=’White’
   ’y’=’Yellow’
   ’r’=’Red’
   ’g’=’Green’
   ’b’=’Blue’;
run;
```

Since a full display of the results of each cluster analysis would be very long, a macro is used with five macro variables to select parts of the output. The macro variables are set to select only the PROC CLUSTER output and the crosstabulation of clusters and true categories for the first two analyses. The example could be run with different settings of the macro variables to show the full output or other selected parts.

```
%let cluster=1;   /* 1=show CLUSTER output, 0=don’t */
%let tree=0;      /* 1=print TREE diagram, 0=don’t */
%let list=0;      /* 1=list clusters, 0=don’t */
%let crosstab=1;  /* 1=crosstabulate clusters and classes,
                     0=don’t                              */
%let crosscol=0;  /* 1=crosstabulate clusters and colors,
                     0=don’t                              */

   /*--- define macro with options for TREE ---*/
%macro treeopt;
   %if &tree %then h page=1;
   %else noprint;
%mend;

   /*--- define macro with options for CLUSTER ---*/
%macro clusopt;
   %if &cluster %then pseudo ccc p=20;
   %else noprint;
%mend;

   /*------ macro for showing cluster results ------*/
%macro show(n); /* n=number of clusters
                   to show results for */

proc tree data=tree %treeopt n=&n out=out;
   id name;
   copy class height width depth weight color;
run;

%if &list %then %do;
   proc sort;
      by cluster;
   run;

   proc print;
      var class name height width depth weight color;
      by cluster clusname;
   run;
%end;
```

```
%if &crosstab %then %do;
   proc tabulate noseps /* formchar='           ' */;
        class class cluster;
        table cluster, class*n='
              '*f=stars./rts=10 misstext=' ';
run;
%end;

%if &crosscol %then %do;
   proc tabulate noseps /* formchar='           ' */;
      class color cluster;
      table cluster, color*n='
            '*f=stars./rts=10 misstext=' ';
      format color $color.;
run;
%end;
%mend;
```

The first analysis uses the variables height, width, depth, and weight in standardized form to show the effect of including size information. The CCC, pseudo F, and pseudo t^2 statistics indicate 10 clusters. Most of the clusters do not correspond closely to the true categories, and four of the clusters have only one or two observations.

```
/**********************************************************/
/*                                                        */
/*       Analysis 1: standardized box measurements        */
/*                                                        */
/**********************************************************/
title2 'Analysis 1: Standardized data';
proc cluster data=grocery m=cen std %clusopt outtree=tree;
   var height width depth weight;
   id name;
   copy class color;
run;

%show(10);
```

Output 23.6.1. Analysis of Standardized Data

```
                         Cluster Analysis of Grocery Boxes
                           Analysis 1: Standardized data

                               The CLUSTER Procedure
                       Centroid Hierarchical Cluster Analysis

                        Eigenvalues of the Correlation Matrix

                 Eigenvalue    Difference    Proportion    Cumulative

            1    2.44512438    1.64456210        0.6113        0.6113
            2    0.80056228    0.33149770        0.2001        0.8114
            3    0.46906458    0.18381582        0.1173        0.9287
            4    0.28524876                      0.0713        1.0000


               The data have been standardized to mean 0 and variance 1
               Root-Mean-Square Total-Sample Standard Deviation =        1
               Root-Mean-Square Distance Between Observations   = 2.828427
```

```
                         Cluster Analysis of Grocery Boxes
                           Analysis 1: Standardized data

                               The CLUSTER Procedure
                       Centroid Hierarchical Cluster Analysis

               The data have been standardized to mean 0 and variance 1
               Root-Mean-Square Total-Sample Standard Deviation =        1
               Root-Mean-Square Distance Between Observations   = 2.828427


                                   Cluster History
                                                                                Norm   T
                                                                                Cent   i
NCL   ----------Clusters Joined----------     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2    Dist   e

 20   CL22               Lipton Family Si       11   0.0028   .974   .        .     85.4    4.5  0.3073
 19   CL36               Corn Flakes             5   0.0026   .972   .        .     83.7   15.3  0.3146
 18   CL24               CL41                   12   0.0080   .964   .        .     70.2   10.0  0.3316
 17   CL18               CL30                   18   0.0144   .949   .        .     53.8   12.7  0.3343
 16   Marshmallow Supr   CL29                    3   0.0024   .947   .        .     55.8    4.7  0.3363
 15   CL50               CL33                    7   0.0055   .941   .        .     55.0   24.4   0.346
 14   CL46               CL15                   10   0.0069   .934   .        .     53.7    8.1  0.3192
 13   CL27               Lipton Family Si        6   0.0035   .931   .        .     56.1    6.3   0.362
 12   CL31               CL16                    5   0.0075   .923   .861   8.03   55.8    6.6  0.4416
 11   CL19               CL23                    7   0.0102   .913   .848   7.59   54.6   12.7  0.4713
 10   Arm & Hammer Det   Tide                    2   0.0037   .909   .835   8.36   59.1     .   0.4781
  9   CL11               CL17                   25   0.0393   .870   .819   4.72   45.2   19.3  0.4918
  8   CL13               CL14                   16   0.0329   .837   .801   2.95   40.4   23.7  0.5215
  7   CL8                CL20                   27   0.0629   .774   .779   -.31   32.0   25.9  0.5467
  6   CL7                Crest                  28   0.0112   .763   .752   0.61   36.7    2.4  0.6003
  5   CL9                CL6                    53   0.1879   .575   .718   -5.9   19.6   43.4  0.6641
  4   CL5                CL21                   55   0.0345   .541   .672   -5.2   23.2    4.5   0.745
  3   CL4                CL12                   60   0.1137   .427   .602   -5.3   22.4   14.5  0.8769
  2   CL3                CL10                   62   0.1511   .276   .471   -4.3   23.2   15.8  1.5559
  1   CL2                Arm & Hammer Det       63   0.2759   .000   .000   0.00     .    23.2   2.948
```

```
----------------------------------------------------------------------------------------------------------
|        |                                             class                                             |
|        |-----------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |               |
|        |    cereal     |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea      |
|--------+---------------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |               |
|1       |               |               |               |               |               |    ###########|
|2       |               |             ##|               |              #|               |            ###|
|3       |          #####|               |             ##|               |               |               |
|4       |               |               |               |            ###|        #######|               |
|5       |    ###########|             ##|            ###|               |               |             ##|
|6       |               |               |               |          #####|               |               |
|7       |               |              #|               |               |               |              #|
|8       |               |               |             ##|               |               |               |
|9       |               |               |               |               |              #|               |
|10      |               |               |              #|               |               |               |
----------------------------------------------------------------------------------------------------------
```

The second analysis uses logarithms of height, width, depth, and the cube root of weight; the cube root is used for consistency with the linear measures. The rows are then centered to remove size information. Finally, the columns are standardized to have a standard deviation of 1. There is no compelling a priori reason to standardize the columns, but if they are not standardized, height dominates the analysis because of its large variance. The STANDARD procedure is used instead of the STD option in PROC CLUSTER so that a subsequent analysis can separately standardize the dummy variables for color.

```
/*********************************************************/
/*                                                       */
/*    Analysis 2: standardized row-centered logarithms   */
/*                                                       */
/*********************************************************/

title2 'Row-centered logarithms';
data shape;
   set grocery;
   array x height width depth weight;
   array l l_height l_width l_depth l_weight;
                          /* logarithms */
   weight=weight**(1/3);  /* take cube root to conform with
                             the other linear measurements */
   do over l;             /* take logarithms */
      l=log(x);
   end;
   mean=mean( of l(*));   /* find row mean of logarithms */
   do over l;
      l=l-mean;           /* center row */
   end;
run;

title2 'Analysis 2: Standardized row-centered logarithms';
proc standard data=shape out=shapstan m=0 s=1;
   var l_height l_width l_depth l_weight;
run;
```

```
proc cluster data=shapstan m=cen %clusopt outtree=tree;
   var l_height l_width l_depth l_weight;
   id name;
   copy class height width depth weight color;
run;

%show(8);
```

The results of the second analysis are shown for eight clusters. Clusters 1 through 4 correspond fairly well to tea, toothpaste, breakfast cereals, and detergents. Crackers and Little Debbie products are scattered among several clusters.

Output 23.6.2. Analysis of Standardized Row-Centered Logarithms

```
                                Cluster Analysis of Grocery Boxes
                          Analysis 2: Standardized row-centered logarithms

                                      The CLUSTER Procedure
                                Centroid Hierarchical Cluster Analysis

                                 Eigenvalues of the Covariance Matrix

                            Eigenvalue    Difference    Proportion    Cumulative

                       1    1.94931049    0.34845395        0.4873        0.4873
                       2    1.60085654    1.15102358        0.4002        0.8875
                       3    0.44983296    0.44983296        0.1125        1.0000
                       4    0.00000000                      0.0000        1.0000


                    Root-Mean-Square Total-Sample Standard Deviation =        1
                    Root-Mean-Square Distance Between Observations   = 2.828427


                                           Cluster History
                                                                                         Norm   T
                                                                                         Cent   i
NCL   ----------Clusters Joined----------     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2    Dist   e

 20   CL29              All-Bran                 4   0.0017   .977   .       .     94.7    2.9   0.2658
 19   CL26              CL27                     8   0.0045   .972   .       .     85.4    8.4   0.3047
 18   Fudge Rounds      Crest                    2   0.0016   .971   .       .     87.2     .    0.3193
 17   Fudge Brownies    Snack Cakes              2   0.0018   .969   .       .     89.1     .    0.3331
 16   Arm & Hammer Det  Lipton Loose Tea         2   0.0019   .967   .       .     91.3     .    0.3434
 15   CL23              CL18                     5   0.0050   .962   .       .     86.5    4.8   0.3587
 14   CL37              CL21                     5   0.0051   .957   .       .     83.5   10.4   0.3613
 13   CL30              CL24                     9   0.0068   .950   .       .     79.2   12.9   0.3682
 12   CL32              CL20                    16   0.0142   .936   .892   5.75   67.6   29.3   0.3826
 11   CL22              Apple Delights           4   0.0037   .932   .881   6.31   71.4    3.2   0.3901
 10   CL11              CL31                     7   0.0090   .923   .869   6.17   70.8    6.3   0.4032
  9   CL33              CL13                    11   0.0092   .914   .853   6.25   71.7    7.6   0.4181
  8   CL19              CL16                    10   0.0131   .901   .835   6.12   71.4   10.9    0.503
  7   CL14              CL9                     16   0.0297   .871   .813   4.63   63.1   15.6   0.5173
  6   CL10              CL15                    12   0.0329   .838   .785   3.69   59.1   13.6   0.5916
  5   CL6               CL28                    19   0.0557   .783   .748   2.01   52.2   15.8   0.6252
  4   CL12              CL8                     26   0.0885   .694   .697   -.16   44.6   48.8   0.6679
  3   CL5               CL17                    21   0.0459   .648   .617   1.21   55.3    7.4   0.8863
  2   CL4               CL7                     42   0.2841   .364   .384   -.56   34.9   60.3   0.9429
  1   CL2               CL3                     63   0.3640   .000   .000   0.00     .    34.9   0.8978
```

```
----------------------------------------------------------------------------------------------------------
|        |                                             class                                             |
|        |-----------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |               |
|        |    cereal     |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea      |
|--------+---------------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |               |
|1       |               |              #|               |               |               |     ##########|
|2       |               |               |               |               |        #######|               |
|3       | ##############|             ##|               |               |               |               |
|4       |              #|               |       ########|               |               |              #|
|5       |               |               |               |             ##|              #|             ##|
|6       |              #|               |               |               |               |           ####|
|7       |               |             ##|               |          #####|               |               |
|8       |               |               |               |             ##|               |               |
----------------------------------------------------------------------------------------------------------
```

The third analysis is similar to the second analysis except that the rows are standardized rather than just centered. There is a clear indication of seven clusters from the CCC, pseudo F, and pseudo t^2 statistics. The clusters are listed as well as crosstabulated with the true categories and colors.

```
/************************************************************/
/*                                                          */
/*  Analysis 3: standardized row-standardized logarithms  */
/*                                                          */
/************************************************************/

%let list=1;
%let crosscol=1;

title2 ’Row-standardized logarithms’;
data std;
   set grocery;
   array x height width depth weight;
   array l l_height l_width l_depth l_weight;
                      /* logarithms */
   weight=weight**(1/3); /* take cube root to conform with
                         the other linear measurements */
   do over l;
      l=log(x);          /* take logarithms */
   end;
   mean=mean( of l(*));  /* find row mean of logarithms */
   std=std( of l(*));    /* find row standard deviation */
   do over l;
      l=(l-mean)/std;    /* standardize row */
   end;
run;

title2 ’Analysis 3: Standardized row-standardized logarithms’;
proc standard data=std out=stdstan m=0 s=1;
   var l_height l_width l_depth l_weight;
run;

proc cluster data=stdstan m=cen %clusopt outtree=tree;
   var l_height l_width l_depth l_weight;
   id name;
```

```
    copy class height width depth weight color;
run;

%show(7);
```

The output from the third analysis shows that cluster 1 contains 9 of the 17 teas. Cluster 2 contains all of the detergents plus Grape Nuts, a very heavy cereal. Cluster 3 includes all of the toothpastes and one Little Debbie product that is of very similar shape, although roughly twice as large. Cluster 4 has most of the cereals, Ritz crackers (which come in a box very similar to most of the cereal boxes), and Lipton Loose Tea (all the other teas in the sample come in tea bags). Clusters 5 and 6 each contain several Luzianne and Lipton teas and one or two miscellaneous items. Cluster 7 includes most of the Little Debbie products and two types of crackers. Thus, the crackers are not identified and the teas are broken up into three clusters, but the other categories correspond to single clusters. This analysis classifies toothpaste and Little Debbie products slightly better than the second analysis,

Output 23.6.3. Analysis of Standardized Row-Standardized Logarithms

```
                                           Cluster Analysis of Grocery Boxes
                                  Analysis 3: Standardized row-standardized logarithms

                                                 The CLUSTER Procedure
                                         Centroid Hierarchical Cluster Analysis

                                          Eigenvalues of the Covariance Matrix

                                      Eigenvalue    Difference    Proportion    Cumulative

                                 1    2.42684848    0.94583675        0.6067        0.6067
                                 2    1.48101173    1.38887193        0.3703        0.9770
                                 3    0.09213980    0.09213980        0.0230        1.0000
                                 4    -.00000000                     -0.0000        1.0000


                              Root-Mean-Square Total-Sample Standard Deviation =        1
                              Root-Mean-Square Distance Between Observations   = 2.828427


                                                     Cluster History
                                                                                                           Norm   T
                                                                                                           Cent   i
        NCL   ----------Clusters Joined----------     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2     Dist   e

         20   CL35               CL33                    8   0.0024   .990   .       .      229   32.0   0.1923
         19   CL22               Ritz                    5   0.0010   .989   .       .      224    2.9   0.2014
         18   CL44               CL27                    6   0.0018   .987   .       .      206   20.5   0.2073
         17   CL18               CL26                    9   0.0025   .985   .       .      187    6.4   0.1956
         16   Fudge Rounds       Crest                   2   0.0009   .984   .       .      192     .      0.24
         15   CL24               CL23                    5   0.0029   .981   .       .      177    7.8   0.2753
         14   CL25               Waverly Wafers          4   0.0021   .979   .       .      175    7.7   0.2917
         13   CL30               CL19                   17   0.0101   .969   .       .      130   41.0   0.2974
         12   CL16               CL31                    9   0.0049   .964   .932   5.49    124   20.5   0.3121
         11   CL21               Lipton Family Si        4   0.0029   .961   .924   5.81    129    8.2   0.3445
         10   CL41               CL11                    6   0.0045   .957   .915   5.94    130    5.0    0.323
          9   CL29               Lipton Tea Bags         4   0.0031   .953   .904   6.52    138   20.3   0.3603
          8   CL14               CL15                    9   0.0101   .943   .890   6.08    131   10.7   0.3761
          7   CL20               Lipton Family Si        9   0.0047   .939   .872   6.89    143   11.7   0.4063
          6   CL13               CL9                    21   0.0272   .911   .848   5.23    117   30.0   0.5101
          5   CL6                CL17                   30   0.0746   .837   .814   1.30   74.3   42.2    0.606
          4   CL10               CL7                    15   0.0440   .793   .764   1.40   75.3   36.4   0.6152
          3   CL8                CL12                   18   0.0642   .729   .681   2.02   80.6   44.0   0.6648
          2   CL3                CL4                    33   0.2580   .471   .470   0.01   54.2   54.4   0.9887
          1   CL5                CL2                    63   0.4707   .000   .000   0.00     .    54.2   0.9636
```

```
------------------------------------------------ CLUSTER=1 CLUSNAME=CL7 ------------------------------------------------

                    Obs    class          name          height    width    depth     weight    color

                      1     Tea     Bigelow Plantati      7.7      13.4     6.9     3.27107      g
                      2     Tea     Bigelow Earl Gre      7.7      13.4     6.9     3.27107      b
                      3     Tea     Celestial Saeson      7.8      13.8     6.3     3.65931      b
                      4     Tea     Celestial Saeson      7.8      13.8     6.3     3.58305      r
                      5     Tea     Bigelow Lemon Li      7.7      13.4     6.9     3.41995      y
                      6     Tea     Celestial Saeson      7.8      13.8     6.3     3.82586      y
                      7     Tea     Celestial Saeson      7.8      13.8     6.3     3.33222      g
                      8     Tea     Lipton Tea Bags       6.7      10.0     5.7     3.28271      r
                      9     Tea     Lipton Family Si      8.9      11.1     8.2     4.39510      r


----------------------------------------------- CLUSTER=2 CLUSNAME=CL17 ------------------------------------------------

              Obs    class               name                height    width    depth     weight    color

               10    Detergent           Tide                  26.5      19.9      6.3    10.5928      r
               11    Detergent           Tide                  19.3      14.6      4.7     7.8357      r
               12    Detergent           Tide                  32.5      23.2      7.3    12.6889      r
               13    Breakfast cereal    Grape Nuts            21.7      16.3      4.9     8.7937      w
               14    Detergent           Arm & Hammer Det      33.7      22.8      7.0    14.7023      y
               15    Detergent           Arm & Hammer Det      27.8      19.4      6.3    12.2003      y
               16    Detergent           Arm & Hammer Det      38.8      30.0     16.9    22.4732      y
               17    Detergent           Tide                  39.4      24.8     11.3    16.1045      r
               18    Detergent           Arm & Hammer Det      39.5      25.8     11.0    18.6115      y


----------------------------------------------- CLUSTER=3 CLUSNAME=CL12 ------------------------------------------------

                  Obs        class        name            height    width    depth     weight    color

                   19    Paste, Tooth     Colgate           3.6      15.6     3.3     4.39510      r
                   20    Paste, Tooth     Crest             3.5      15.2     3.2     4.24343      w
                   21    Paste, Tooth     Crest             4.3      17.4     3.6     5.06813      w
                   22    Paste, Tooth     Arm & Hammer      4.4      17.0     3.7     5.21097      w
                   23    Paste, Tooth     Colgate           4.2      18.3     3.5     5.21097      r
                   24    Paste, Tooth     Crest             4.3      21.7     3.7     5.65790      w
                   25    Paste, Tooth     Colgate           4.4      22.0     3.5     5.82946      r
                   26    Little Debbie    Fudge Rounds      8.1      28.3     5.4     6.45411      w
                   27    Paste, Tooth     Crest             3.0      10.9     2.8     2.88670      w
```

```
----------------------------------------------- CLUSTER=4 CLUSNAME=CL13 -----------------------------------------------

              Obs    class              name                height    width    depth     weight    color

               28    Breakfast cereal   Cheerios             27.5      19.0     6.2     6.56541      y
               29    Breakfast cereal   Froot Loops          25.0      17.7     6.4     6.77735      r
               30    Breakfast cereal   Special K            30.1      20.5     8.5     7.98644      w
               31    Breakfast cereal   Corn Flakes          30.2      20.6     8.4     7.98644      w
               32    Breakfast cereal   Special K            29.6      19.2     6.7     6.97679      w
               33    Breakfast cereal   Corn Flakes          30.0      19.1     6.6     6.97679      w
               34    Breakfast cereal   Froot Loops          30.2      20.8     8.5     8.23034      r
               35    Breakfast cereal   Cheerios             30.3      20.4     7.2     7.51847      y
               36    Breakfast cereal   Cheerios             24.1      17.2     5.3     5.82848      y
               37    Breakfast cereal   Corn Flakes          33.7      25.4     8.0     8.79021      w
               38    Breakfast cereal   Special K            23.4      16.6     5.7     5.82946      w
               39    Breakfast cereal   Cheerios             32.5      22.4     8.4     8.27677      y
               40    Breakfast cereal   Shredded Wheat,      26.6      19.6     5.6     7.98957      r
               41    Crackers           Ritz                 23.1      16.0     5.3     6.97953      r
               42    Breakfast cereal   All-Bran             21.1      14.3     5.2     7.30951      y
               43    Tea                Lipton Loose Tea     12.7      10.9     5.4     6.09479      r
               44    Crackers           Ritz                 23.1      20.7     5.2     7.68573      r


----------------------------------------------- CLUSTER=5 CLUSNAME=CL10 -----------------------------------------------

                Obs    class            name                height    width    depth     weight    color

                 45    Tea              Luzianne              8.9      22.8      6.4    5.53748      r
                 46    Tea              Luzianne Decaffe      8.9      22.8      6.4    5.29641      g
                 47    Crackers         Premium Saltines     11.0      25.0     10.7    7.68573      w
                 48    Tea              Lipton Family Si      8.7      20.8      8.2    5.53748      r
                 49    Little Debbie    Marshmallow Supr      9.4      32.0      7.0    6.56541      w
                 50    Tea              Lipton Family Si     13.7      24.0      9.0    6.97679      r
```

```
------------------------------------------------ CLUSTER=6 CLUSNAME=CL9 -----------------------------------------------

               Obs    class              name                height    width    depth     weight    color

                51    Tea                Luzianne             18.4      20.2     6.9     6.09479      r
                52    Tea                Lipton Tea Bags      17.1      20.0     6.7     6.09479      r
                53    Breakfast cereal   Shredded Wheat       19.7      19.9     7.5     6.56541      y
                54    Tea                Lipton Tea Bags      11.5      14.4     6.6     4.73448      r


------------------------------------------------ CLUSTER=7 CLUSNAME=CL8 -----------------------------------------------

                Obs    class            name                height    width    depth     weight    color

                 55    Crackers         Wheatsworth          11.1      25.2     5.5     6.88239      w
                 56    Little Debbie    Swiss Cake Rolls     10.1      21.8     5.8     7.16545      w
                 57    Little Debbie    Figaroos             13.5      18.6     3.7     6.97679      y
                 58    Little Debbie    Nutty Bar            13.2      18.5     4.2     6.97679      y
                 59    Little Debbie    Apple Delights       11.2      30.1     4.9     7.51552      w
                 60    Little Debbie    Lemon Stix           13.2      18.5     4.2     6.33884      w
                 61    Little Debbie    Fudge Brownies       11.0      30.8     2.5     6.97679      w
                 62    Little Debbie    Snack Cakes          13.4      32.0     3.4     7.16545      b
                 63    Crackers         Waverly Wafers       14.4      22.5     6.2     7.68573      g
```

```
----------------------------------------------------------------------------------------------------------
|        |                                             class                                             |
|        |-----------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |               |
|        |    cereal     |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea      |
|--------+---------------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |               |
|1       |               |               |               |               |               |      #########|
|2       |              #|               |       ########|               |               |               |
|3       |               |               |               |              #|       ########|               |
|4       | ##############|             ##|               |               |               |              #|
|5       |               |              #|               |              #|               |           ####|
|6       |              #|               |               |               |               |            ###|
|7       |               |             ##|               |        #######|               |               |
----------------------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------------------
|        |                                     color                                     |
|        |-------------------------------------------------------------------------------|
|        |     Blue      |     Green     |      Red      |     White     |    Yellow     |
|--------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |
|1       |             ##|             ##|            ###|               |             ##|
|2       |               |               |           ####|              #|           ####|
|3       |               |               |            ###|         ######|               |
|4       |               |               |         ######|         ######|          #####|
|5       |               |              #|            ###|             ##|               |
|6       |               |               |            ###|               |              #|
|7       |              #|              #|               |          #####|             ##|
------------------------------------------------------------------------------------------
```

The last several analyses include color. Obviously, the dummy variables must not be included in calculations to standardize the rows. If the five dummy variables are simply standardized to variance 1.0 and included with the other variables, color dominates the analysis. The dummy variables should be scaled to a smaller variance, which must be determined by trial and error. Four analyses are done using PROC STANDARD to scale the dummy variables to a standard deviation of 0.2, 0.3, 0.4, or 0.8. The cluster listings are suppressed.

Since dummy variables drastically violate the normality assumption on which the CCC depends, the CCC tends to indicate an excessively large number of clusters.

```
/************************************************************/
/*                                                          */
/* Analyses 4-7: standardized row-standardized logs & color */
/*                                                          */
/************************************************************/
%let list=0;
%let crosscol=1;

title2
  'Analysis 4: Standardized row-standardized
               logarithms and color (s=.2)';
proc standard data=stdstan out=stdstan m=0 s=.2;
   var c_:;
run;

proc cluster data=stdstan m=cen %clusopt outtree=tree;
   var l_height l_width l_depth l_weight c_:;
   id name;
   copy class height width depth weight color;
run;

%show(7);

title2
  'Analysis 5: Standardized row-standardized
               logarithms and color (s=.3)';
proc standard data=stdstan out=stdstan m=0 s=.3;
   var c_:;
run;
```

```
proc cluster data=stdstan m=cen %clusopt outtree=tree;
   var l_height l_width l_depth l_weight c_:;
   id name;
   copy class height width depth weight color;
run;

%show(6);

title2
  'Analysis 6: Standardized row-standardized
               logarithms and color (s=.4)';
proc standard data=stdstan out=stdstan m=0 s=.4;
   var c_:;
run;

proc cluster data=stdstan m=cen %clusopt outtree=tree;
   var l_height l_width l_depth l_weight c_:;
   id name;
   copy class height width depth weight color;
run;

%show(3);

title2
  'Analysis 7: Standardized row-standardized
               logarithms and color (s=.8)';
proc standard data=stdstan out=stdstan m=0 s=.8;
   var c_:;
run;

proc cluster data=stdstan m=cen %clusopt outtree=tree;
   var l_height l_width l_depth l_weight c_:;
   id name;
   copy class height width depth weight color;
run;

%show(10);
```

Using PROC STANDARD on the dummy variables with S=0.2 causes four of the Little Debbie products to join the toothpastes. Using S=0.3 causes one of the tea clusters to merge with the breakfast cereals while three cereals defect to the detergents. Using S=0.4 produces three clusters consisting of (1) cereals and detergents, (2) Little Debbie products and toothpaste, and (3) teas, with crackers divided among all three clusters and a few other misclassifications. With S=0.8, ten clusters are indicated, each entirely monochrome. So, S=0.2 or S=0.3 degrades the classification, S=0.4 yields a good but perhaps excessively coarse classification, and higher values of the S= option produce clusters that are determined mainly by color.

Output 23.6.4. Analysis of Standardized Row-Standardized Logarithms and Color

```
                               Cluster Analysis of Grocery Boxes
                 Analysis 4: Standardized row-standardized logarithms and color (s=.2)

                                     The CLUSTER Procedure
                               Centroid Hierarchical Cluster Analysis

                                Eigenvalues of the Covariance Matrix

                            Eigenvalue    Difference    Proportion    Cumulative

                       1    2.43584975    0.94791932        0.5800        0.5800
                       2    1.48793042    1.39363531        0.3543        0.9342
                       3    0.09429511    0.03686218        0.0225        0.9567
                       4    0.05743293    0.01036136        0.0137        0.9704
                       5    0.04707157    0.00489503        0.0112        0.9816
                       6    0.04217654    0.00693298        0.0100        0.9916
                       7    0.03524355    0.03524355        0.0084        1.0000
                       8    -.00000000    0.00000000       -0.0000        1.0000
                       9    -.00000000                     -0.0000        1.0000


                      Root-Mean-Square Total-Sample Standard Deviation =  0.68313
                      Root-Mean-Square Distance Between Observations   = 2.898275


                                          Cluster History
                                                                                               Norm   T
                                                                                               Cent   i
 NCL   ----------Clusters Joined----------     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2    Dist   e

  20   CL46               Lemon Stix              3   0.0016   .968   .        .     67.5   11.9   0.2706
  19   Luzianne           Lipton Family Si        2   0.0014   .966   .        .     69.7     .    0.2995
  18   CL25               CL37                    6   0.0041   .962   .        .     67.1    5.0   0.3081
  17   CL33               CL35                   16   0.0099   .952   .        .     57.2   16.7   0.3196
  16   CL19               Luzianne Decaffe        3   0.0024   .950   .        .     59.2    1.7   0.3357
  15   CL30               CL16                    5   0.0042   .946   .        .     59.5    2.7   0.3299
  14   CL27               CL18                    8   0.0057   .940   .        .     58.9    4.2   0.3429
  13   CL20               Fudge Brownies          4   0.0031   .937   .        .     61.7    3.6   0.3564
  12   CL24               Lipton Tea Bags         4   0.0031   .934   .905   3.23   65.2    4.7    0.359
  11   CL39               CL28                    6   0.0068   .927   .896   3.17   65.9   12.1   0.3743
  10   CL13               Snack Cakes             5   0.0036   .923   .886   3.62   70.8    2.3   0.3755
   9   CL11               CL32                   13   0.0176   .906   .874   2.70   64.8   16.0   0.4107
   8   CL14               Lipton Family Si        9   0.0052   .900   .859   3.29   71.0    2.6   0.4265
   7   Waverly Wafers     CL10                    6   0.0052   .895   .841   4.09   79.8    2.4   0.4378
   6   CL17               CL12                   20   0.0248   .870   .817   3.52   76.6   19.7   0.4898
   5   CL15               CL8                    14   0.0326   .838   .783   3.08   75.0   14.0   0.5607
   4   CL6                CL21                   30   0.0743   .764   .734   1.35   63.5   35.6   0.5877
   3   CL9                CL7                    19   0.0579   .706   .653   2.17   72.0   22.8   0.6611
   2   CL4                CL3                    49   0.3632   .343   .450   -2.6   31.8   73.0   0.9838
   1   CL2                CL5                    63   0.3426   .000   .000   0.00     .    31.8   0.9876
```

```
----------------------------------------------------------------------------------------------------------
|        |                                           class                                               |
|        |-----------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |               |
|        |    cereal     |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea      |
|--------+---------------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |               |
|1       |             ##|               |       ########|               |               |               |
|2       |               |              #|               |           ####|       ########|               |
|3       |  #############|             ##|               |               |               |              #|
|4       |              #|               |               |               |               |            ###|
|5       |               |              #|               |          #####|               |               |
|6       |               |               |               |               |               |      #########|
|7       |               |              #|               |               |               |           ####|
----------------------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------------------
|        |                                     color                                     |
|        |-------------------------------------------------------------------------------|
|        |     Blue      |     Green     |     Red       |     White     |    Yellow     |
|--------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |
|1       |               |               |           ####|              #|          #####|
|2       |               |               |            ###|     ##########|               |
|3       |               |               |         ######|         ######|           ####|
|4       |               |               |            ###|               |              #|
|5       |              #|              #|               |             ##|             ##|
|6       |             ##|             ##|            ###|               |             ##|
|7       |               |              #|            ###|              #|               |
------------------------------------------------------------------------------------------
```

```
                                   Cluster Analysis of Grocery Boxes
                 Analysis 5: Standardized row-standardized logarithms and color (s=.3)

                                         The CLUSTER Procedure
                                 Centroid Hierarchical Cluster Analysis

                                  Eigenvalues of the Covariance Matrix

                              Eigenvalue    Difference    Proportion    Cumulative

                         1    2.44752302    0.95026671        0.5500        0.5500
                         2    1.49725632    1.36701945        0.3365        0.8865
                         3    0.13023687    0.02135049        0.0293        0.9157
                         4    0.10888637    0.00867367        0.0245        0.9402
                         5    0.10021271    0.00628821        0.0225        0.9627
                         6    0.09392449    0.02196469        0.0211        0.9838
                         7    0.07195981    0.07195981        0.0162        1.0000
                         8    0.00000000    0.00000000        0.0000        1.0000
                         9    -.00000000                     -0.0000        1.0000


                      Root-Mean-Square Total-Sample Standard Deviation = 0.703167
                      Root-Mean-Square Distance Between Observations   = 2.983287


                                             Cluster History
                                                                                                   Norm   T
                                                                                                   Cent   i
NCL   ----------Clusters Joined----------     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2     Dist   e

 20   CL24               CL28                    4   0.0038   .953   .       .     45.7    2.7   0.3448
 19   Grape Nuts         CL23                    6   0.0033   .950   .       .     46.0    3.5   0.3477
 18   CL46               Lemon Stix              3   0.0027   .947   .       .     47.1   21.9   0.3558
 17   CL21               Lipton Tea Bags         4   0.0031   .944   .       .     48.2    2.5   0.3577
 16   CL39               CL33                    6   0.0064   .937   .       .     46.9   12.1   0.3637
 15   CL19               CL29                   14   0.0152   .922   .       .     40.6   12.4   0.3707
 14   CL18               Fudge Brownies          4   0.0035   .919   .       .     42.5    2.5   0.3813
 13   CL16               CL25                   13   0.0175   .901   .       .     38.0   13.7   0.4103
 12   CL22               Lipton Family Si        5   0.0049   .896   .875   1.76   40.0    3.2   0.4353
 11   CL12               CL37                    7   0.0089   .887   .865   1.71   40.9    4.6   0.4397
 10   CL20               Luzianne Decaffe        5   0.0056   .882   .854   2.02   43.9    2.5   0.4669
  9   CL26               CL17                   16   0.0222   .859   .841   1.20   41.3   16.6    0.479
  8   CL32               CL11                    9   0.0125   .847   .826   1.31   43.5    4.5   0.4988
  7   CL14               Snack Cakes             5   0.0070   .840   .806   1.95   49.0    3.3    0.519
  6   Waverly Wafers     CL7                     6   0.0077   .832   .782   2.79   56.6    2.3   0.5366
  5   CL9                CL15                   30   0.0716   .761   .749   0.54   46.1   28.3   0.5452
  4   CL10               CL8                    14   0.0318   .729   .700   1.21   52.9    8.6   0.5542
  3   CL5                CL6                    36   0.0685   .660   .622   1.50   58.3   14.2   0.6516
  2   CL13               CL4                    27   0.2008   .460   .427   0.90   51.9   46.6   0.9611
  1   CL3                CL2                    63   0.4595   .000   .000   0.00     .    51.9   0.9609
```

```
----------------------------------------------------------------------------------------------------------
|        |                                             class                                             |
|        |-----------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |               |
|        |    cereal     |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea      |
|--------+---------------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |               |
|1       |            ###|             ##|       ########|               |               |              #|
|2       |               |              #|               |           ####|       ########|               |
|3       |  #############|               |               |               |               |            ###|
|4       |               |              #|               |          #####|               |               |
|5       |               |               |               |               |               |      #########|
|6       |               |              #|               |               |               |           ####|
----------------------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------------------
|        |                                     color                                     |
|        |-------------------------------------------------------------------------------|
|        |     Blue      |     Green     |      Red      |     White     |    Yellow     |
|--------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |
|1       |               |               |       ########|              #|          #####|
|2       |               |               |            ###|     ##########|               |
|3       |               |               |          #####|         ######|          #####|
|4       |              #|              #|               |             ##|             ##|
|5       |             ##|             ##|            ###|               |             ##|
|6       |               |              #|            ###|              #|               |
------------------------------------------------------------------------------------------
```

```
                            Cluster Analysis of Grocery Boxes
                Analysis 6: Standardized row-standardized logarithms and color (s=.4)

                                  The CLUSTER Procedure
                            Centroid Hierarchical Cluster Analysis

                             Eigenvalues of the Covariance Matrix

                           Eigenvalue    Difference    Proportion    Cumulative

                      1    2.46469435    0.95296119        0.5135        0.5135
                      2    1.51173316    1.28149311        0.3149        0.8284
                      3    0.23024005    0.04306536        0.0480        0.8764
                      4    0.18717469    0.01766446        0.0390        0.9154
                      5    0.16951023    0.01827481        0.0353        0.9507
                      6    0.15123542    0.06582379        0.0315        0.9822
                      7    0.08541162    0.08541162        0.0178        1.0000
                      8    -.00000000    0.00000000       -0.0000        1.0000
                      9    -.00000000                     -0.0000        1.0000


                  Root-Mean-Square Total-Sample Standard Deviation = 0.730297
                  Root-Mean-Square Distance Between Observations   = 3.098387


                                       Cluster History
                                                                                    Norm   T
                                                                                    Cent   i
     NCL   ----------Clusters Joined----------     FREQ    SPRSQ    RSQ   ERSQ    CCC    PSF   PST2     Dist   e

      20   CL29              CL44                    10   0.0074   .955   .       .     47.7    8.2   0.3789
      19   CL38              Lipton Family Si         3   0.0031   .952   .       .     48.1    9.3   0.3792
      18   CL25              CL41                    11   0.0155   .936   .       .     38.8   36.7   0.4192
      17   CL23              CL43                    10   0.0120   .924   .       .     35.0   11.6   0.4208
      16   Grape Nuts        CL26                     6   0.0050   .919   .       .     35.6    5.8   0.4321
      15   CL19              CL31                     5   0.0074   .912   .       .     35.4    5.3   0.4362
      14   Premium Saltines  CL27                     4   0.0046   .907   .       .     36.8    2.9   0.4374
      13   CL18              CL20                    21   0.0352   .872   .       .     28.4   19.7   0.4562
      12   CL13              CL16                    27   0.0372   .835  .839   -.37    23.4   12.0   0.4968
      11   CL21              CL17                    15   0.0289   .806  .828   -1.5    21.6   13.6   0.5183
      10   CL14              CL15                     9   0.0200   .786  .815   -1.8    21.6    7.2   0.5281
       9   Waverly Wafers    Luzianne Decaffe         2   0.0047   .781  .801   -1.2    24.1     .    0.5425
       8   CL10              CL24                    12   0.0243   .757  .785   -1.3    24.5    5.8   0.5783
       7   CL12              CL46                    29   0.0224   .735  .765   -1.3    25.8    5.3   0.6105
       6   CL8               CL37                    14   0.0220   .712  .740   -1.1    28.3    4.0   0.6313
       5   CL6               CL32                    16   0.0251   .687  .707   -.78    31.9    3.9   0.6664
       4   CL11              CL9                     17   0.0287   .659  .660   -.04    38.0    7.0   0.7098
       3   CL4               Snack Cakes             18   0.0180   .641  .584   2.21    53.5    3.2   0.7678
       2   CL3               CL5                     34   0.2175   .423  .400   0.67    44.8   31.4   0.8923
       1   CL7               CL2                     63   0.4232   .000  .000   0.00      .    44.8   0.9156
```

```
-----------------------------------------------------------------------------------------------------------
|        |                                         class                                                  |
|        |------------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |                |
|        |     cereal    |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea       |
|--------+---------------+---------------+---------------+---------------+---------------+----------------|
|CLUSTER |               |               |               |               |               |                |
|1       |>##############|             ##|       ########|             ##|               |               #|
|2       |               |             ##|               |        #######|       ########|               #|
|3       |               |              #|               |               |               |>###############|
-----------------------------------------------------------------------------------------------------------
```

```
-------------------------------------------------------------------------------------------
|        |                                   color                                        |
|        |--------------------------------------------------------------------------------|
|        |     Blue      |     Green     |      Red      |     White     |    Yellow      |
|--------+---------------+---------------+---------------+---------------+----------------|
|CLUSTER |               |               |               |               |                |
|1       |               |               |     ##########|        #######|    ############|
|2       |              #|             ##|            ###|   ############|                |
|3       |             ##|             ##|      #########|              #|              ##|
-------------------------------------------------------------------------------------------
```

```
                               Cluster Analysis of Grocery Boxes
                 Analysis 7: Standardized row-standardized logarithms and color (s=.8)

                                     The CLUSTER Procedure
                               Centroid Hierarchical Cluster Analysis

                              Eigenvalues of the Covariance Matrix

                            Eigenvalue    Difference    Proportion    Cumulative

                       1    2.61400794    0.93268930        0.3631        0.3631
                       2    1.68131864    0.77645948        0.2335        0.5966
                       3    0.90485916    0.22547234        0.1257        0.7222
                       4    0.67938683    0.00292216        0.0944        0.8166
                       5    0.67646466    0.12119211        0.0940        0.9106
                       6    0.55527255    0.46658428        0.0771        0.9877
                       7    0.08868827    0.08868827        0.0123        1.0000
                       8    -.00000000    0.00000000       -0.0000        1.0000
                       9    -.00000000                     -0.0000        1.0000


                    Root-Mean-Square Total-Sample Standard Deviation = 0.894427
                    Root-Mean-Square Distance Between Observations   = 3.794733


                                         Cluster History
                                                                                         Norm   T
                                                                                         Cent   i
NCL   ----------Clusters Joined----------     FREQ    SPRSQ     RSQ   ERSQ     CCC     PSF   PST2     Dist   e

 20   CL29               CL44                   10   0.0049    .970   .           .    72.7    8.2   0.3094
 19   CL38               Lipton Family Si        3   0.0021    .968   .           .    73.3    9.3   0.3096
 18   CL21               CL23                   12   0.0153    .952   .           .    53.0   15.0   0.4029
 17   Waverly Wafers     Luzianne Decaffe        2   0.0032    .949   .           .    53.8     .     0.443
 16   CL27               CL24                    6   0.0095    .940   .           .    48.9   10.4    0.444
 15   CL19               CL16                    9   0.0136    .926   .           .    43.0    6.1   0.4587
 14   CL41               Grape Nuts              7   0.0058    .920   .           .    43.6   51.2   0.4591
 13   CL26               CL46                    7   0.0105    .910   .           .    42.1   22.0   0.4769
 12   CL25               CL13                   12   0.0205    .889   .743    16.5    37.3   13.8    0.467
 11   CL18               Premium Saltines       13   0.0093    .880   .726    16.7    38.2    4.0   0.5586
 10   CL17               CL37                    4   0.0134    .867   .706    16.5    38.3    7.9   0.6454
  9   CL14               CL20                   17   0.0567    .810   .684    11.0    28.8   52.6   0.6534
  8   CL12               CL9                    29   0.0828    .727   .659    5.03    20.9   20.7    0.604
  7   CL11               CL43                   16   0.0359    .691   .631    4.25    20.9   14.4   0.6758
  6   CL15               CL31                   11   0.0263    .665   .598    4.24    22.6    8.0   0.7065
  5   CL7                CL6                    27   0.1430    .522   .557    -1.7    15.8   28.2   0.8247
  4   CL8                CL5                    56   0.2692    .253   .507    -9.1     6.6   31.5   0.7726
  3   Snack Cakes        CL32                    3   0.0216    .231   .435    -6.6     9.0   46.0   1.0027
  2   CL4                CL10                   60   0.1228    .108   .289    -5.6     7.4    9.5   1.0096
  1   CL2                CL3                    63   0.1083    .000   .000    0.00      .     7.4   1.0839
```

```
----------------------------------------------------------------------------------------------------------
|        |                                       class                                                   |
|        |-----------------------------------------------------------------------------------------------|
|        |   Breakfast   |               |               |               |               |               |
|        |    cereal     |   Crackers    |   Detergent   | Little Debbie | Paste, Tooth  |      Tea      |
|--------+---------------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |               |
|1       |            ###|             ##|           ####|               |               |              #|
|2       |               |             ##|               |         ######|          #####|               |
|3       |        #######|               |               |               |               |               |
|4       |         ######|               |           ####|             ##|               |               |
|5       |               |               |               |               |            ###|               |
|6       |               |               |               |               |               |      #########|
|7       |               |              #|               |               |               |            ###|
|8       |               |               |               |               |               |             ##|
|9       |               |               |               |               |               |             ##|
|10      |               |               |               |              #|               |               |
----------------------------------------------------------------------------------------------------------
```

```
------------------------------------------------------------------------------------------
|        |                                     color                                     |
|        |-------------------------------------------------------------------------------|
|        |     Blue      |     Green     |      Red      |     White     |    Yellow     |
|--------+---------------+---------------+---------------+---------------+---------------|
|CLUSTER |               |               |               |               |               |
|1       |               |               |     ##########|               |               |
|2       |               |               |               |  #############|               |
|3       |               |               |               |        #######|               |
|4       |               |               |               |               |   ############|
|5       |               |               |            ###|               |               |
|6       |               |               |      #########|               |               |
|7       |               |           ####|               |               |               |
|8       |             ##|               |               |               |               |
|9       |               |               |               |               |             ##|
|10      |              #|               |               |               |               |
------------------------------------------------------------------------------------------
```

References

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

Batagelj, V. (1981), “Note on Ultrametric Hierarchical Clustering Algorithms,” *Psychometrika*, 46, 351–352.

Blackith, R.E. and Reyment, R.A. (1971), *Multivariate Morphometrics*, London: Academic Press.

Blashfield, R.K. and Aldenderfer, M.S. (1978), “The Literature on Cluster Analysis,” *Multivariate Behavioral Research*, 13, 271–295.

Calinski, T. and Harabasz, J. (1974), “A Dendrite Method for Cluster Analysis,” *Communications in Statistics*, 3, 1–27.

Cooper, M.C. and Milligan, G.W. (1988), “The Effect of Error on Determining the Number of Clusters,” *Proceedings of the International Workship on Data Analysis, Decision Support, and Expert Knowledge Representation in Marketing and Related Areas of Research*, 319–328.

Duda, R.O. and Hart, P.E. (1973), *Pattern Classification and Scene Analysis*, New York: John Wiley & Sons, Inc.

Everitt, B.S. (1980), *Cluster Analysis*, Second Edition, London: Heineman Educational Books Ltd.

Fisher, L. and Van Ness, J.W. (1971), “Admissible Clustering Procedures,” *Biometrika*, 58, 91–104.

Fisher, R.A. (1936), “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics*, 7, 179–188.

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951a), “Sur la Liaison et la Division des Points d’un Ensemble Fini,” *Colloquium Mathematicae*, 2, 282–285.

Florek, K., Lukaszewicz, J., Perkal, J., and Zubrzycki, S. (1951b), “Taksonomia Wroclawska,” *Przeglad Antropol.*, 17, 193–211.

Gower, J.C. (1967), “A Comparison of Some Methods of Cluster Analysis,” *Biometrics*, 23, 623–637.

Hamer, R.M. and Cunningham, J.W. (1981), “Cluster analyzing profile data with interrater differences: A comparison of profile association measures,” *Applied Psychological Measurement*, 5, 63–72.

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

Hartigan, J.A. (1977), “Distribution Problems in Clustering,” in *Classification and Clustering*, ed. J. Van Ryzin, New York: Academic Press, Inc.

Hartigan, J.A. (1981), “Consistency of Single Linkage for High-density Clusters,” *Journal of the American Statistical Association*, 76, 388–394.

Hawkins, D.M., Muller, M.W., and ten Krooden, J.A. (1982), “Cluster Analysis,” in *Topics in Applied Multivariate Analysis*, ed. D.M. Hawkins, Cambridge: Cambridge University Press.

Jardine, N. and Sibson, R. (1971), *Mathematical Taxonomy*, New York: John Wiley & Sons, Inc.

Johnson, S.C. (1967), “Hierarchical Clustering Schemes,” *Psychometrika*, 32, 241–254.

Lance, G.N. and Williams, W.T. (1967), “A General Theory of Classificatory Sorting Strategies. I. Hierarchical Systems,” *Computer Journal*, 9, 373–380.

Massart, D.L. and Kaufman, L. (1983), *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, New York: John Wiley & Sons, Inc.

McQuitty, L.L. (1957), “Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies,” *Educational and Psychological Measurement*, 17, 207–229.

McQuitty, L.L. (1966), “Similarity Analysis by Reciprocal Pairs for Discrete and Continuous Data,” *Educational and Psychological Measurement*, 26, 825–831.

Mezzich, J.E and Solomon, H. (1980), *Taxonomy and Behavioral Science*, New York: Academic Press, Inc.

Milligan, G.W. (1979), “Ultrametric Hierarchical Clustering Algorithms,” *Psychometrika*, 44, 343–346.

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. (1987), “A Study of the Beta-Flexible Clustering Method,” *College of Administrative Science Working Paper Series*, 87–61 Columbus, OH: The Ohio State University.

Milligan, G.W. and Cooper, M.C. (1985), “An Examination of Procedures for Determining the Number of Clusters in a Data Set,” *Psychometrika*, 50,159–179.

Milligan, G.W. and Cooper, M.C. (1987), “A Study of Variable Standardization,” *College of Administrative Science Working Paper Series*, 87–63, Columbus, OH: The Ohio State University.

Rouncefield, M. (1995), “The Statistics of Poverty and Inequality,” *Journal of Statistics Education*, 3(2). [Online]: [http://www.stat.ncsu.edu/info/jse], accessed Dec. 19, 1997.

Sarle, W.S. (1983), *Cubic Clustering Criterion*, SAS Technical Report A-108, Cary, NC: SAS Institute Inc.

Silverman, B.W. (1986), *Density Estimation*, New York: Chapman and Hall.

Sneath, P.H.A. (1957), “The Application of Computers to Taxonomy,” *Journal of General Microbiology*, 17, 201–226.

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

Sokal, R.R. and Michener, C.D. (1958), “A Statistical Method for Evaluating Systematic Relationships,” *University of Kansas Science Bulletin*, 38, 1409–1438.

Sorensen, T. (1948), “A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and Its Application to Analyses of the Vegetation on Danish Commons,” *Biologiske Skrifter*, 5, 1–34.

Spath, H. (1980), *Cluster Analysis Algorithms*, Chichester, England: Ellis Horwood.

Symons, M.J. (1981), “Clustering Criteria and Multivariate Normal Mixtures,” *Biometrics*, 37, 35–43.

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

Wishart, D. (1969), “Mode Analysis: A Generalisation of Nearest Neighbour Which Reduces Chaining Effects,” in *Numerical Taxonomy*, ed. A.J. Cole, London: Academic Press.

Wong, M.A. (1982), “A Hybrid Clustering Method for Identifying High-Density Clusters,” *Journal of the American Statistical Association*, 77, 841–847.

Wong, M.A. and Lane, T. (1983), “A kth Nearest Neighbor Clustering Procedure,” *Journal of the Royal Statistical Society*, Series B, 45, 362–368.

Wong, M.A. and Schaack, C. (1982), “Using the kth Nearest Neighbor Clustering Procedure to Determine the Number of Subpopulations,” *American Statistical Association 1982 Proceedings of the Statistical Computing Section*, 40–48.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

Copyright © 1999 by SAS Institute Inc., Cary, NC, USA.

ISBN 1–58025–494–2

All rights reserved. Produced in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

U.S. Government Restricted Rights Notice. Use, duplication, or disclosure of the software and related documentation by the U.S. government is subject to the Agreement with SAS Institute and the restrictions set forth in FAR 52.227–19 Commercial Computer Software-Restricted Rights (June 1987).

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries.® indicates USA registration.

Other brand and product names are registered trademarks or trademarks of their respective companies.

The Institute is a private company devoted to the support and further development of its software and related services.