Chapter 37
The CLUSTER Procedure

Contents

Overview: CLUSTER Procedure ........................................ 2426
Getting Started: CLUSTER Procedure ................................ 2427
Syntax: CLUSTER Procedure ........................................... 2434
  PROC CLUSTER Statement ........................................... 2434
  BY Statement ................................................................ 2444
  COPY Statement ........................................................ 2444
  FREQ Statement ........................................................ 2445
  ID Statement ................................................................ 2445
  RMSSTD Statement .................................................... 2445
  VAR Statement .......................................................... 2446
Details: CLUSTER Procedure ............................................ 2446
  Clustering Methods ...................................................... 2446
  Miscellaneous Formulas .............................................. 2453
  Ultrametrics .............................................................. 2454
  Algorithms ............................................................... 2455
  Computational Resources ........................................... 2456
  Missing Values .......................................................... 2456
  Ties ........................................................................... 2457
  Size, Shape, and Correlation ........................................ 2457
  Output Data Set .......................................................... 2458
  Displayed Output ....................................................... 2461
  ODS Table Names ........................................................ 2463
  ODS Graphics ............................................................ 2464
Examples: CLUSTER Procedure ......................................... 2465
  Example 37.1: Cluster Analysis of Flying Mileages between 10 American Cities ... 2465
  Example 37.2: Crude Birth and Death Rates ....................................... 2472
  Example 37.3: Cluster Analysis of Fisher’s Iris Data ............................... 2485
  Example 37.4: Evaluating the Effects of Ties ......................................... 2499
References ........................................................................ 2509
Overview: CLUSTER Procedure

The CLUSTER procedure hierarchically clusters the observations in a SAS data set by using one of 11 methods. The data can be coordinates or distances. If the data are coordinates, PROC CLUSTER computes (possibly squared) Euclidean distances. If you want non-Euclidean distances, use the DISTANCE procedure (see Chapter 40) to compute an appropriate distance data set that can then be used as input to PROC CLUSTER.

The clustering methods 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. Each method is described in the section “Clustering Methods” on page 2446.

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.

The CLUSTER procedure is not practical for very large data sets because the CPU time is roughly proportional to the square or cube of the number of observations. The FASTCLUS procedure (see Chapter 42) requires time proportional to the number of observations and thus can be used with much larger data sets than PROC CLUSTER. If you want to cluster a very large data set hierarchically, use PROC FASTCLUS for a preliminary cluster analysis to produce a large number of clusters. Then use PROC CLUSTER to cluster the preliminary clusters hierarchically. This method is illustrated in Example 37.3.

PROC CLUSTER displays a history of the clustering process, showing statistics useful for estimating the number of clusters in the population from which the data are sampled. It creates a dendrogram when ODS Graphics is enabled. PROC CLUSTER also creates an output data set that can be used by the TREE procedure 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, and then use this output data set as the input data set to the TREE procedure. With PROC TREE, specify the NCLUSTERS=6 and the OUT= options to obtain the six-cluster solution. For an example, see Example 124.1 in Chapter 124, “The TREE Procedure.”

For coordinate data, Euclidean distances are computed from differences between coordinate values. The use of differences has several important consequences:

- For differences to be valid, the variables must have an interval or stronger scale of measurement. Ordinal or ranked data are generally not appropriate for cluster analysis.

- For Euclidean distances to be comparable, equal differences should have equal practical importance. You might need to transform the variables linearly or nonlinearly to satisfy this condition. For example, if one variable is measured in dollars and one in euros, you might need to convert to the same currency. Or, if ratios are more meaningful than differences, take logarithms.
Variables with large variances tend to have more effect on the resulting clusters than variables with small variances. If you consider all variables to be equally important, you can use the STD option in PROC CLUSTER to standardize the variables to mean 0 and standard deviation 1. However, 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 CLUSTER with the STD option unless you specify the TRIM= option. The STDIZE procedure (see Chapter 112) provides additional methods for standardizing variables and imputing missing values.

The ACECLUS procedure (see Chapter 24) is useful for linear transformations of the variables if any of the following conditions hold:

- You have no idea how the variables should be scaled.
- You want to detect natural clusters regardless of whether some variables have more influence than others.
- You want to use a clustering method designed for finding compact clusters, but you want to be able to detect elongated clusters.

Agglomerative hierarchical clustering is discussed in all standard references on cluster analysis, such as Anderberg (1973); Sneath and Sokal (1973); Hartigan (1975); Everitt (1980); 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); Cooper and Milligan (1988).


Getting Started: CLUSTER Procedure

This example shows 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 categorize countries. Previous studies indicate that the clusters computed from this type of data can be elongated and elliptical. Thus, you need to perform a linear transformation on the raw data before the cluster analysis.
Chapter 37: The CLUSTER Procedure

The following data\(^1\) from Rouncefield (1995) are birth rates, death rates, and infant death rates for 97 countries. The DATA step creates the SAS data set Poverty:

```sas
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
... more lines ...
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 by using the SAS data set Poverty. The OUT= option creates an output SAS data set called Ace that contains the canonical variable scores:

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

The P= option specifies that approximately 3% 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:

```sas
ods graphics on;
proc cluster data=Ace method=ward ccc pseudo print=15 out=tree
  plots=den(height=rsq);
  var can1-can3;
  id country;
run;
ods graphics off;
```

---
\(^1\) 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.
The ODS GRAPHICS ON statement enables ODS Graphics. 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. The PRINT=15 option displays only the last 15 generations of the cluster history. By default, when ODS Graphics is enabled, a dendrogram displaying the semipartial R square is displayed on the X axis. The option PLOTS=DEN(HEIGHT=RSQ) requests a dendrogram with R square displayed instead.

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

PROC CLUSTER first displays the table of eigenvalues of the covariance matrix (Figure 37.1). These eigenvalues are used in the computation of the cubic clustering criterion. 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.

**Figure 37.1** Table of Eigenvalues of the Covariance Matrix

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 64.5500051</td>
<td>54.7313223</td>
<td>0.8091</td>
<td>0.8091</td>
</tr>
<tr>
<td>2 9.8186828</td>
<td>4.4038309</td>
<td>0.1231</td>
<td>0.9321</td>
</tr>
<tr>
<td>3 5.4148519</td>
<td>0.0679</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

Root-Mean-Square Total-Sample Standard Deviation 5.156987

Root-Mean-Square Distance Between Observations 12.63199

Figure 37.2 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 CL$n$, where $n$ is the number of the cluster. Next, PROC CLUSTER displays the number of observations in the new cluster and the semipartial R square. The latter value represents the decrease in the proportion of variance accounted for by joining the two clusters.
Figure 37.2 Cluster History

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>Approximate R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 Oman CL37</td>
<td>5</td>
<td>0.0039</td>
<td>.957</td>
<td>.933</td>
<td>6.03</td>
<td>132</td>
<td>12.1</td>
<td></td>
</tr>
<tr>
<td>14 CL31 CL22</td>
<td>13</td>
<td>0.0040</td>
<td>.953</td>
<td>.928</td>
<td>5.81</td>
<td>131</td>
<td>9.7</td>
<td></td>
</tr>
<tr>
<td>13 CL41 CL17</td>
<td>32</td>
<td>0.0041</td>
<td>.949</td>
<td>.922</td>
<td>5.70</td>
<td>131</td>
<td>13.1</td>
<td></td>
</tr>
<tr>
<td>12 CL19 CL21</td>
<td>10</td>
<td>0.0045</td>
<td>.945</td>
<td>.916</td>
<td>5.65</td>
<td>132</td>
<td>6.4</td>
<td></td>
</tr>
<tr>
<td>11 CL39 CL15</td>
<td>9</td>
<td>0.0052</td>
<td>.940</td>
<td>.909</td>
<td>5.60</td>
<td>134</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>10 CL76 CL27</td>
<td>6</td>
<td>0.0075</td>
<td>.932</td>
<td>.900</td>
<td>5.25</td>
<td>133</td>
<td>18.1</td>
<td></td>
</tr>
<tr>
<td>9 CL23 CL11</td>
<td>15</td>
<td>0.0130</td>
<td>.919</td>
<td>.890</td>
<td>4.20</td>
<td>125</td>
<td>12.4</td>
<td></td>
</tr>
<tr>
<td>8 CL10 Afghanistan</td>
<td>7</td>
<td>0.0134</td>
<td>.906</td>
<td>.879</td>
<td>3.55</td>
<td>122</td>
<td>7.3</td>
<td></td>
</tr>
<tr>
<td>7 CL9 CL25</td>
<td>17</td>
<td>0.0217</td>
<td>.884</td>
<td>.864</td>
<td>2.26</td>
<td>114</td>
<td>11.6</td>
<td></td>
</tr>
<tr>
<td>6 CL8 CL20</td>
<td>14</td>
<td>0.0239</td>
<td>.860</td>
<td>.846</td>
<td>1.42</td>
<td>112</td>
<td>10.5</td>
<td></td>
</tr>
<tr>
<td>5 CL14 CL13</td>
<td>45</td>
<td>0.0307</td>
<td>.829</td>
<td>.822</td>
<td>0.65</td>
<td>112</td>
<td>59.2</td>
<td></td>
</tr>
<tr>
<td>4 CL16 CL7</td>
<td>28</td>
<td>0.0323</td>
<td>.797</td>
<td>.788</td>
<td>0.57</td>
<td>122</td>
<td>14.8</td>
<td></td>
</tr>
<tr>
<td>3 CL12 CL6</td>
<td>24</td>
<td>0.0323</td>
<td>.765</td>
<td>.732</td>
<td>1.84</td>
<td>153</td>
<td>11.6</td>
<td></td>
</tr>
<tr>
<td>2 CL3 CL4</td>
<td>52</td>
<td>0.1782</td>
<td>.587</td>
<td>.613</td>
<td>-0.82</td>
<td>135</td>
<td>48.9</td>
<td></td>
</tr>
<tr>
<td>1 CL5 CL2</td>
<td>97</td>
<td>0.5866</td>
<td>.000</td>
<td>.000</td>
<td>0.00</td>
<td>.</td>
<td>135</td>
<td></td>
</tr>
</tbody>
</table>

Next listed is the squared multiple correlation, R square, which is the proportion of variance accounted for by the clusters. Figure 37.2 shows that, when the data are grouped into three clusters, the proportion of variance accounted for by the clusters (R square) is just under 77%. The approximate expected value of R square is given in the ERSq column. This expectation is approximated under the null hypothesis that the data have a uniform distribution instead of forming distinct clusters.

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 for estimating the number of clusters in the data.

The final column in Figure 37.2 lists ties for minimum distance; a blank value indicates the absence of a tie. A tie means that the clusters are indeterminate and that changing the order of the observations might change the clusters. See Example 37.4 for ways to investigate the effects of ties.

Figure 37.3 plots the three statistics for estimating the number of clusters. Peaks in the plot of the cubic clustering criterion with values greater than 2 or 3 indicate good clusters; peaks with values between 0 and 2 indicate possible clusters. Large negative values of the CCC can indicate outliers. In Figure 37.3, there is a local peak of the CCC when the number of clusters is three. The CCC drops at four clusters and then steadily increases, leveling off at eleven 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 good numbers of clusters. In Figure 37.3, the pseudo F statistic suggests three clusters or eleven clusters.
To interpret the values of the pseudo $r^2$ statistic, look down the column or look at the plot from right to left until you find the first value that is markedly larger than the previous value, then move back up the column or to the right in the plot by one step in the cluster history. In Figure 37.3, you can see possibly good clustering levels at eleven clusters, six clusters, three clusters, and two clusters.

Considered together, these statistics suggest that the data can be clustered into eleven clusters or three clusters. The following statements examine the results of clustering the data into three clusters.

Figure 37.4 displays the dendrogram. The figure provides a graphical view of the information in Figure 37.2. As the number of branches grows to the left from the root, the R square approaches 1; the first three clusters (branches of the tree) account for over half of the variation (about 77%, from Figure 37.4). In other words, only three clusters are necessary to explain over three-fourths of the variation.
Figure 37.4  Dendrogram of Clusters versus R-Square Values
You can use PROC TREE and the output data set from PROC CLUSTER to create a new data set that contains information about cluster membership as follows:

```
proc tree data=Tree out=New nclusters=3 noprint;
  height _rsq_;
  copy can1 can2;
  id country;
run;
```

The SAS data set `Tree` is input. The OUT= option creates an output SAS data set named `New` that contains information about cluster membership. The NCLUSTERS= option specifies the number of clusters desired in the data set `New`. The results can be displayed in a scatter plot.

The following statements use the SGПLOT procedure to display the results that are in the SAS data set `New`:

```
proc sgplot data=New;
  scatter y=can2 x=can1 / group=cluster;
run;
```

The SCATTER statement requests a plot of the two canonical variables, using the value of the variable cluster, which is produced by PROC TREE as the identification variable. The results are displayed in Figure 37.5.

**Figure 37.5** Plot of Canonical Variables and Cluster for Three Clusters

![Figure 37.5](image)

The statistics in Figure 37.2 and Figure 37.3, the dendrogram in Figure 37.4, and the plot of the canonical variables in Figure 37.5 assist in the estimation of the number 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: CLUSTER Procedure

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 invokes the CLUSTER procedure. It also specifies a clustering method, and optionally specifies details for clustering methods, data sets, data processing, and displayed output.

Table 37.1 summarizes the options available in the PROC CLUSTER statement.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Specify input and output data sets</td>
<td>DATA= Specifies input data set</td>
</tr>
<tr>
<td>Specify clustering methods</td>
<td>BETA= Specifies beta value for flexible beta method</td>
</tr>
<tr>
<td>Control data processing prior to clustering</td>
<td>NOEIGEN Suppresses computation of eigenvalues</td>
</tr>
<tr>
<td>Control density estimation</td>
<td>K= Specifies number of neighbors for kth-nearest-neighbor density estimation</td>
</tr>
</tbody>
</table>
### PROC CLUSTER Statement

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>R</strong>=</td>
<td>Specifies radius of sphere of support for uniform-kernel density estimation</td>
</tr>
<tr>
<td><strong>Ties</strong></td>
<td>Suppresses checking for ties</td>
</tr>
</tbody>
</table>

**Control display of the cluster history**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CCC</strong></td>
<td>Displays cubic clustering criterion</td>
</tr>
<tr>
<td><strong>NOID</strong></td>
<td>Suppresses display of ID values</td>
</tr>
<tr>
<td><strong>PRINT</strong>=</td>
<td>Specifies number of generations to display</td>
</tr>
<tr>
<td><strong>PSEUDO</strong></td>
<td>Displays pseudo $F$ and $t^2$ statistics</td>
</tr>
<tr>
<td><strong>RMSSTD</strong></td>
<td>Displays root mean square standard deviation</td>
</tr>
<tr>
<td><strong>RSQUARE</strong></td>
<td>Displays $R$ square and semipartial $R$ square</td>
</tr>
</tbody>
</table>

**Control other aspects of output**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NOPRINT</strong></td>
<td>Suppresses display of all output</td>
</tr>
<tr>
<td><strong>PLOTS</strong>=</td>
<td>Specifies ODS graphics details</td>
</tr>
<tr>
<td><strong>SIMPLE</strong></td>
<td>Displays simple summary statistics</td>
</tr>
</tbody>
</table>

**METHOD=** *name*

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 either 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.

- **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 for details. 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 (weighted average linkage, weighted pair-group method using arithmetic averages, WPGMA).

MEDIAN | MED requests Gower’s median method (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 list provides details about the other 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 square under the uniform null hypothesis (Sarle 1983). The statistics associated with the RSQUARE option, R square and semipartial R square, 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. Computation of the CCC requires the eigenvalues of the covariance matrix. If the number of variables is large, computing the eigenvalues requires much computer time and memory.

**DATA=SAS-data-set**

names the input data set that contains 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 about TYPE=DISTANCE data sets, see Chapter A, “Special SAS Data Sets.”

Data set types (such as TYPE=DISTANCE) do not persist when you copy or modify a data set. You must specify the TYPE= data set option for the new data set, as in the following example:

```sas
   data dist2(type=distance);
      set dist;
   run;
```
If you do not specify the TYPE=DISTANCE data set option, the new data set is the default TYPE=DATA. If you use the new data set in a procedure that accepts both TYPE=DATA or TYPE=DISTANCE data sets (such as PROC CLUSTER or PROC MODECLUS), the results will be incorrect.

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=TWO-STAGE 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=TWO-STAGE. 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 the Wong (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 an _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=TWO-STAGE 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 \(k\)th-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.

Density estimation is used with the TRIM=, METHOD=DENSITY, and METHOD=TWO-STAGE options.
**MODE=n**
specifies that, when two clusters are joined, each must have at least \( n \) members in order for either cluster to be designated a modal cluster. If you specify \( \text{MODE}=1 \), each cluster must also have a maximum density greater than the fusion density in order for either cluster to be designated a modal cluster.

Use the \( \text{MODE}= \) option only with \( \text{METHOD=DENSITY} \) or \( \text{METHOD=TWOSTAGE} \). With \( \text{METHOD=TWOSTAGE} \), the \( \text{MODE}= \) option affects the number of modal clusters formed. With \( \text{METHOD=DENSITY} \), the \( \text{MODE}= \) 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 \( \text{MODE}= \) is the same as the value of \( K= \) because the use of \( k\)th-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 \( \text{MODE}=2 \).

If you specify \( \text{MODE}=0 \), the default value is used instead of 0.

If you specify a \( \text{FREQ} \) statement or if a \_FREQ_ variable appears in the input data set, the \( \text{MODE}= \) 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 the eigenvalues of the covariance matrix and substitutes the variances of the variables for the eigenvalues when computing the cubic clustering criterion. The NOEIGEN option saves time if the number of variables is large, but it should be used only if the variables are nearly uncorrelated. If you specify the NOEIGEN option and the variables are highly correlated, the cubic clustering criterion might 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 \( \text{METHOD=WARD} \), 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 \( \text{METHOD=EML} \), 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 20, “Using the Output Delivery System.”

**NOSQUARE**
prevents input distances from being squared with \( \text{METHOD=AVERAGE} \), \( \text{METHOD=CENTROID} \), \( \text{METHOD=_MEDIAN} \), or \( \text{METHOD=WARD} \).

If you specify the NOSQUARE option with distance data, the data are assumed to be squared Euclidean distances for computing \( R^2 \)-square and related statistics defined in a Euclidean coordinate system.

If you specify the NOSQUARE option with coordinate data with \( \text{METHOD=CENTROID} \), \( \text{METHOD=_MEDIAN} \), or \( \text{METHOD=WARD} \), 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 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 2457 for more information.

**OUTTREE=SAS-data-set**

creates an output data set that can be used by the TREE procedure to draw a tree diagram. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see *SAS Language Reference: Concepts*. If you omit the OUTTREE= option, the data set is named by using the DATA_n 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 2446 for more information. Values for p must be greater than zero. By default, PENALTY=2.

**PLOTS < (global-plot-options) > < = plot-request >**

controls the plots produced through ODS Graphics.

ODS Graphics must be enabled before plots can be requested. For example:

```r
ods graphics on;
proc cluster method=ward plots=all;
run;
ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 623 in Chapter 21, “Statistical Graphics Using ODS.”

By default, PROC CLUSTER produces a dendrogram. PROC CLUSTER can also produce plots of the cubic clustering criterion, the pseudo $F$ statistic, and the pseudo $t^2$ statistic from the cluster history table. These statistics are useful for estimating the number of clusters. Each statistic is plotted against the number of clusters. You can request that PROC CLUSTER create these graphs by specifying the CCC or PSEUDO options, or by specifying the statistics in a plot-request in the PLOT option. PROC CLUSTER might be unable to compute the statistics in some cases; for details, see the CCC and PSEUDO options. If a statistic cannot be computed, it cannot be plotted. PROC CLUSTER plots all of these statistics that are computed unless you tell it specifically what to plot using PLOTS=.

PROC CLUSTER has a CCC and PSEUDO option as well as CCC and PSEUDO plot-requests. All four options are illustrated in the following step:
ods graphics on;

proc cluster ccc pseudo plots=(ccc pseudo);
   run;
ods graphics off;

The maximum number of clusters shown in all the plots is the minimum of the following quantities:

- the number of observations
- the value of the PRINT= option, if that option is specified
- the maximum number of clusters for which CCC is computed, if the CCC is plotted

The *global-plot-options* apply to all plots generated by the CLUSTER procedure. The *global-plot-options* are as follows:

**MAXCLUS=n**
right-truncates the CCC, PSF, and PST2 plots at the $n$ value. This prevents these plots from losing resolution when a large number of clusters are plotted. The default is MAXCLUS=200.

**MAXPOINTS=n**

**MAXPTS=n**
suppresses the dendrogram when the number of clusters exceeds the $n$ value. This prevents an unreadable plot from being produced. The default is MAXPOINTS=200.

**UNPACKPANEL**

**UNPACK**
breaks a plot that is otherwise paneled into separate plots for each statistic.

**ONLY**
suppresses the default plots. Only plots specifically requested are displayed.

The following *plot-requests* can be specified:

**ALL**
generates all possible plots. The CCC and PSEUDO options must be specified to obtain the CCC, PSF and PST2 plots in addition to the dendrogram.

**CCC**
implicitly specifies the CCC option and, if possible, plots the cubic clustering criterion against the number of clusters.

**DENDROGRAM < (dendrogram-options)>**
requests a dendrogram and specifies *dendrogram-options*. A dendrogram is created by default unless the ONLY *global-plot-option* is requested.

Unlike most graphs, the size of the dendrogram can vary as a function of the number of objects that appear in the dendrogram. You can specify the following *dendrogram-options* to control the size and appearance of the dendrogram:
COMPUTEHEIGHT=a b

CH=a b

specifies the constants for computing the height of the dendrogram. For \( n \) points being clustered, intercept \( a \), and slope \( b \), the height is based in part on \( a + bn \). For a horizontal dendrogram, the default (given in pixels) is \( \text{COMPUTEHEIGHT}=100 \) \( 12 \), the default height in pixels is \( \max(100 + 12n, 480) \), the default height in inches is \( \max(1.04167 + 0.125n, 5) \), and the default height in centimeters is \( \max(2.64583 + 0.3175n, 12.7) \). For a vertical dendrogram, the default height is 480 pixels. The default unit is pixels, and you can use the \( \text{UNIT=} \ dendrogram-option \) to change the unit to inches or centimeters for this option. Inches equals pixels divided by 96, and centimeters equals inches times 2.54.

COMPUTEWIDTH=a b

CW=a b

specifies the constants for computing the width of the dendrogram. For \( n \) points being clustered, intercept \( a \), and slope \( b \), the width is based in part on \( a + bn \). For a vertical dendrogram, the default (given in pixels) is \( \text{COMPUTEWIDTH}=100 \) \( 12 \), the default width in pixels is \( \max(100 + 12n, 640) \), the default width in inches is \( \max(1.04167 + 0.125n, 6.66667) \), and the default width in centimeters is \( \max(2.64583 + 0.3175n, 16.933) \). For a horizontal dendrogram, the default width is 640 pixels. The default unit is pixels, and you can use the \( \text{UNIT=} \ dendrogram-option \) to change the unit to inches or centimeters for this option. Inches equals pixels divided by 96, and centimeters equals inches times 2.54.

HEIGHT=HEIGHT | MODE | NCL | RSQ

H=H | M | N | R

specifies the method for drawing the height of the dendrogram. \( \text{HEIGHT=} \text{HEIGHT} \) is the default.

\( \text{HEIGHT=} \text{HEIGHT} \) specifies the distance or similarity between the last clusters joined, as defined in the section “Clustering Methods” on page 2446.

\( \text{HEIGHT=} \text{MODE} \) pertains to the modal clusters. With \( \text{METHOD=} \text{DENSITY} \), the mode indicates the number of modal clusters contained by the current cluster. With \( \text{METHOD=} \text{TWOSTAGE} \), the mode gives the maximum density in each modal cluster and the fusion density, \( d^* \), for clusters containing two or more modal clusters; for clusters that contain no modal clusters, that value of the \_MODE_ variable is missing.

\( \text{HEIGHT=} \text{NCL} \) specifies that the number of clusters is used.

\( \text{HEIGHT=} \text{RSQ} \) specifies that the squared multiple correlation is used.

HORIZONTAL | VERTICAL

specifies either a horizontal dendrogram with the objects on the vertical axis (HORIZONTAL) or a vertical dendrogram with the objects on the horizontal axis (VERTICAL). The default is HORIZONTAL.

SETHEIGHT=height

SH=height

specifies the height of the dendrogram. By default, the height is based on the \( \text{COMPUTEHEIGHT=} \) option. The default unit is pixels, and you can use the \( \text{UNIT=} \ dendrogram-option \) to change the unit to inches or centimeters for this \( \text{dendrogram-option} \).
**SETWIDTH=width**

**SW=width**

specifies the width of the dendrogram. By default, the width is based on the \texttt{COMPUTEWIDTH=} option. The default unit is pixels, and you can use the \texttt{UNIT=} \texttt{dendrogram-option} to change the unit to inches or centimeters for this \texttt{dendrogram-option}.

**UNIT=PX | IN | CM**

specifies the unit (pixels, inches, or centimeters) for the \texttt{SETHEIGHT=} \texttt{dendrogram-option}, \texttt{SW=width} \texttt{dendrogram-option}, \texttt{COMPUTEHEIGHT=} \texttt{dendrogram-option}, and \texttt{COMPUTEWIDTH=} \texttt{dendrogram-option}.

**NONE**

suppresses all plots.

**PSD**

implicitly specifies the \texttt{PSEUDO} option and, if possible, plots the pseudo $F$ statistic and the pseudo $t^2$ statistic against the number of clusters.

**PST**

implicitly specifies the \texttt{PSEUDO} option and, if possible, plots the pseudo $F$ statistic against the number of clusters.

You can specify one or more of the \texttt{plot-requests} in the same \texttt{PLOT} option. For example, all of the following are valid:

```plaintext
proc cluster plots=(ccc pst2);
proc cluster plots=(psf);
proc cluster plots=psf;
```

The first statement plots both the cubic clustering criterion and the pseudo $t^2$ statistic, while the second and third statements plot the pseudo $F$ statistic only. When you specify only one plot request, you can omit the parentheses around the plot request. When you specify more than one \texttt{plot-request}, you must specify parentheses. Otherwise the second and subsequent \texttt{plot-requests} are \texttt{options}. Since CCC and PSEUDO are both \texttt{options} as well as \texttt{plot-requests}, the following three statements are valid, but they are not equivalent:

```plaintext
proc cluster plots(only)=ccc pseudo;
proc cluster plots(only)=pseudo ccc;
proc cluster plots(only)=(ccc pseudo);
```

The first two examples have one \texttt{plot-request} and one procedure \texttt{option}. The third example has two \texttt{plot-requests}.

The names of the graphs that PROC CLUSTER generates are listed in \texttt{Table 37.5}, along with the required statements and options.
PRINT=n | P=n
specifies the number of generations of the cluster history to display. The PRINT= option displays the latest n generations; for example, PRINT=5 displays the cluster history from one cluster through five clusters. The value of PRINT= 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 is specified. See the section “Miscellaneous Formulas” on page 2453 for more information. 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.

Density estimation is used with the TRIM=, METHOD=DENSITY, and METHOD=TWOSTAGE 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 is specified.

See the section “Miscellaneous Formulas” on page 2453 for more information.

RSQUARE | RSQ
displays the R square and semipartial R square. This option is effective only when the data are coordinates or when METHOD=AVERAGE or METHOD=CENTROID is specified. The R square and semipartial R square statistics are always displayed with METHOD=WARD. See the section “Miscellaneous Formulas” on page 2453 for more information.

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 2453 for more information.

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% of the points, is a reasonable value for many data sets. Densities are estimated by the $k$th-nearest-neighbor or uniform-kernel method. 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 might 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 can 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 in PROC CLUSTER to obtain separate analyses of observations in groups that are 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 you specify more than one BY statement, only the last one specified is used.

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

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in 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 by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts.* For more information about the DATASETS procedure, see the discussion in the *Base 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 an _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 for details.

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 both of the following:

- 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 squared standard deviation of each cluster

Specify the name of the variable containing root mean squared 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 the root mean squared 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 PROC 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: CLUSTER Procedure**

**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 \( x_i \) \( i \)th observation (row vector if coordinate data)
- \( C_K \) \( K \)th cluster, subset of \( \{1, 2, \ldots, n\} \)
- \( N_K \) number of observations in \( C_K \)
- \( \bar{x} \) sample mean vector
- \( \bar{x}_K \) mean vector for cluster \( C_K \)
- \( \|x\| \) Euclidean length of the vector \( x \)—that is, the square root of the sum of the squares of the elements of \( x \)
- \( T \) \[ \sum_{i=1}^{n} \|x_i - \bar{x}\|^2 \]
- \( W_K \) \[ \sum_{i \in C_K} \|x_i - \bar{x}_K\|^2 \]
- \( P_G \) \[ \sum W_J, \text{ where summation is over the } G \text{ clusters at the } G \text{th level of the hierarchy} \]
- \( B_{KL} \) \[ W_M - W_K - W_L \text{ if } C_M = C_K \cup C_L \]
- \( d(x, y) \) any distance or dissimilarity measure between observations or vectors \( x \) and \( y \)
- \( D_{KL} \) 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, see 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 $d(x, y) = \|x - y\|^2$, then

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

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

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

If $d(x, y) = \|x - y\|^2$, 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 might 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 Sørensen (1948).

Density Linkage

The phrase density linkage is used here to refer to a class of clustering methods that use 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 \( k \)th-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.

\textit{kth-Nearest-Neighbor Method}

The \( k \)th-nearest-neighbor method (Wong and Lane 1983) uses \( k \)th-nearest-neighbor density estimates. Let \( r_k(x) \) be the distance from point \( x \) to the \( k \)th-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

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

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

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

**Wong’s Hybrid Method**

The Wong (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, by using the MEAN= option to create a data set containing cluster means, frequencies, and root mean squared 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, than those with 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 \( \bar{x}_L \) is closer to either \( \bar{x}_K \) or \( \bar{x}_L \) than to any other preliminary cluster mean or, equivalently, if \( d^2(\bar{x}_K, \bar{x}_L) < d^2(\bar{x}_K, \bar{x}_M) + d^2(\bar{x}_L, \bar{x}_M) \) for all other preliminary clusters \( C_M \), \( M \neq K \) or \( L \). The new dissimilarity measure is computed as

\[
d^*(\bar{x}_K, \bar{x}_L) = \begin{cases} \frac{(W_K + W_L + \frac{1}{4} (N_K + N_L) d^2(\bar{x}_K, \bar{x}_L))^{\frac{5}{4}}}{(N_K + N_L)^{1 + \frac{1}{2}}} & \text{if } C_K \text{ and } C_L \text{ are adjacent} \\ \infty & \text{otherwise} \end{cases}
\]

**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 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 Schrack 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

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

where \( s_l^2 \) is the standard deviation of the \( l \)th variable. The estimate for R= can be computed in a DATA step by using the GAMMA function for \( \Gamma \). 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 can be preferable if there are outliers.
If the data are distances, the factor $\sum s_i^2$ 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 37.2 can be
used to obtain an initial guess for the $R=$ option.

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
uses 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.

### Table 37.2

Reasonable First Guess for the $R=$ Option for Standardized Data

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

**EML**

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

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

\[ D_{KL} = n v \ln \left( 1 + \frac{B_{KL}}{P} \right) - p (N_M \ln(N_M) - N_K \ln(N_K) - N_L \ln(N_L)) \]

The EML method was derived by W. S. Sarle of SAS Institute 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 \( d(x, y) = \| x - y \|^2 \), 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. (1951b, a) 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 permitted 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. 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

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

If \( d(x, y) = \frac{1}{2}\|x - y\|^2 \), 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 squared standard deviation of a cluster \( C_K \) is

\[ \text{RMSSTD} = \sqrt{\frac{W_K}{\nu(N_K - 1)}} \]

The R-square 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) can 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 square are given in Sarle (1983).

The pseudo \( F \) statistic for a given level is
\[ \text{pseudo } F = \frac{T - P_G}{P_G} \cdot \frac{G - 1}{n - G} \]

The pseudo \( t^2 \) statistic for joining \( C_K \) and \( C_L \) is
\[ \text{pseudo } t^2 = \frac{B_{KL}}{W_K + W_L} \cdot \frac{W_K + W_L}{N_K + N_L - 2} \]

The pseudo \( F \) and \( t^2 \) statistics can 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 Caliński 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 \( x, y \)
- \( d(x, y) \leq \max(d(x, z), d(y, z)) \) 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 \textit{monotonic} or \textit{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 37.3. 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>
<thead>
<tr>
<th>Clustering Method</th>
<th>Algorithm</th>
<th>Stored Data</th>
<th>Stored Distance</th>
<th>Sorted Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVERAGE</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CENTROID</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COMPLETE</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DENSITY</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EML</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLEXIBLE</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCQUITTY</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEDIAN</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SINGLE</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TWOSTAGE</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WARD</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: All of the hierarchical methods accept coordinate data. Methods that require stored or sorted distances automatically calculate distances from the coordinates.
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, when the stored distance algorithm is used, if the distance matrix can be stored in memory as well. To estimate the bytes of memory needed for the data, use the following formula and round up to the nearest multiple of $d$.

$$n(vd + 8d + i) + i$$

if density estimation or the sorted distance algorithm is used

$$+ 3d$$

if stored data algorithm is used

$$+ 3d$$

if density estimation is used

$$+ \text{max}(8, \text{length of ID variable})$$

if ID variable is used

$$+ \text{length of ID variable}$$

if ID variable is used

$$+ \text{sum of lengths of COPY variables}$$

if COPY variables is used

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 permitted in the lower triangle of the distance matrix. The upper triangle is ignored. For more about TYPE=DISTANCE data sets, see Chapter 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 can 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.

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 that occur 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 37.4 for details.

Size, Shape, and Correlation

In some biological applications, the organisms that are being clustered can 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 data item. 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:

```sas
proc transpose data= coordinatedatatype;
run;
proc standard m=0;
run;
proc transpose out=row-centered-coordinate-data;
run;
```

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 observation 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 might need to cube the measures of length or take the cube root of the weights. Various other complications can 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 tend to give higher overall ratings than other raters. Some raters also tend to spread out their ratings over more of the scale than 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 differences both 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 might be necessary in practice to get reasonable results. It is not clear whether iterating the standardization of rows and columns can 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'.

**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
The total number of output observations is usually $2n - 1$, where $n$ is the number of input observations. The density methods can 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 CL$n$, where $n$ is the number of the cluster. If the node is an observation, the name is OB$n$, 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 2446. 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 square under the uniform null hypothesis
- _RATIO_, equal to $(1 - _ERSQ_) / (1 - _RSQ_)$
- _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 squared 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 any of the TRIM=, K= or R= options are used), the output data set contains the following:

- \_DENS\_, the maximum density in the current cluster
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 following:

- 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 Squared 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 Squared Distance Between Observations
- the Mean Distance Between Observations

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

- 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 might be split 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 the data are coordinates, or if you specify METHOD=AVERAGE, METHOD=CENTER, or METHOD=WARD, then PROC CLUSTER displays the root mean squared 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=CENTER.

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

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

- Approximate Expected R-Square or ERSQ, the approximate expected value of R square under the uniform null hypothesis
- the Cubic Clustering Criterion or CCC. The cubic clustering criterion and approximate expected R square 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 the data are coordinates, or if you specify METHOD=AVERAGE, METHOD=CENTER, or METHOD=WARD, then PROC CLUSTER displays the following:

- 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 squared distance between pairs of objects in the two clusters joined with one object from each cluster.

If METHOD=CENTER, 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 the following:
- Normalized Fusion Density or Normalized Fusion Dens, the value of \( d^* \) as defined in the section “Clustering Methods” on page 2446
- 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 the following:

- 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=W ARD, 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=T WOSTAGE 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=T WOSTAGE 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 Table 37.4. For more information about ODS, see Chapter 20, “Using the Output Delivery System.”

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClusterHistory</td>
<td>Observation or clusters joined, frequencies and other cluster statistics</td>
<td>PROC</td>
<td>default</td>
</tr>
<tr>
<td>SimpleStatistics</td>
<td>Simple statistics, before or after trimming</td>
<td>PROC</td>
<td>SIMPLE</td>
</tr>
</tbody>
</table>
Table 37.4  (continued)

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>EigenvalueTable</td>
<td>Eigenvalues of the CORR or COV matrix</td>
<td>PROC</td>
<td>default</td>
</tr>
<tr>
<td>RMSSd</td>
<td>Root mean square total sample standard deviation</td>
<td>PROC</td>
<td>default</td>
</tr>
<tr>
<td>AvDist</td>
<td>Root mean square distance between observations</td>
<td>PROC</td>
<td>default</td>
</tr>
</tbody>
</table>

ODS Graphics

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, “Statistical Graphics Using ODS.”

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 623 in Chapter 21, “Statistical Graphics Using ODS.”

The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “A Primer on ODS Statistical Graphics” on page 622 in Chapter 21, “Statistical Graphics Using ODS.”

PROC CLUSTER can produce plots of the cubic clustering criterion, pseudo $F$, and pseudo $t^2$ statistics, and a dendrogram. To plot a statistic, you must ask for it to be computed via one or more of the CCC, PSEUDO, or PLOT options.

You can reference every graph produced through ODS Graphics with a name. The names of the graphs that PROC CLUSTER generates are listed in Table 37.5, along with the required statements and options.

Table 37.5  Graphs Produced by PROC CLUSTER

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>Statement &amp; Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>CubicClusCritPlot</td>
<td>Cubic clustering criterion for the number of clusters</td>
<td>PROC CLUSTER PLOTS=CCC</td>
</tr>
<tr>
<td>PseudoFPlot</td>
<td>Pseudo $F$ criterion for the number of clusters</td>
<td>PROC CLUSTER PLOTS=PSF</td>
</tr>
<tr>
<td>PseudoTSqPlot</td>
<td>Pseudo $t^2$ criterion for the number of clusters</td>
<td>PROC CLUSTER PLOTS=PST2</td>
</tr>
<tr>
<td>CccAndPsTSqPlot</td>
<td>Cubic clustering criterion and pseudo $t^2$</td>
<td>PROC CLUSTER PLOTS=(CCC PST2)</td>
</tr>
<tr>
<td>CccAndPsFPlot</td>
<td>Cubic clustering criterion and pseudo $F$</td>
<td>PROC CLUSTER PLOTS=(CCC PSF)</td>
</tr>
<tr>
<td>CccPsfAndPsTSqPlot</td>
<td>Cubic clustering criterion, pseudo $F$, and pseudo $t^2$</td>
<td>PROC CLUSTER PLOTS=ALL</td>
</tr>
<tr>
<td>Dendrogram</td>
<td>Dendrogram (tree diagram)</td>
<td>PROC CLUSTER PLOTS=DENDROGRAM</td>
</tr>
</tbody>
</table>
Example 37.1: Cluster Analysis of Flying Mileages between 10 American Cities

This example clusters 10 American cities based on the flying mileages between them. Six clustering methods are shown with corresponding dendrograms. 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 that contains Denver and Houston.

The following step displays the city mileage SAS data set, which is available in the Sashelp library and is designated as a TYPE=DISTANCE data set when it is used by PROC CLUSTER:

```
proc print noobs data=sashelp.mileages;
run;
```

Output 37.1.1 City Mileage Data Set

<table>
<thead>
<tr>
<th>City</th>
<th>Atlanta</th>
<th>Chicago</th>
<th>Denver</th>
<th>Houston</th>
<th>LosAngeles</th>
<th>Miami</th>
<th>NewYork</th>
<th>SanFrancisco</th>
<th>Seattle</th>
<th>WashingtonDC</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>587</td>
<td>0</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1212</td>
<td>920</td>
<td>0</td>
<td></td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>701</td>
<td>940</td>
<td>879</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1936</td>
<td>1745</td>
<td>831</td>
<td>1374</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>604</td>
<td>1188</td>
<td>1726</td>
<td>968</td>
<td>2339</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>748</td>
<td>713</td>
<td>1631</td>
<td>1420</td>
<td>2451</td>
<td>1092</td>
<td>0</td>
<td>.</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2139</td>
<td>1858</td>
<td>949</td>
<td>1645</td>
<td>347</td>
<td>2594</td>
<td>2571</td>
<td>0</td>
<td></td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2182</td>
<td>1737</td>
<td>1021</td>
<td>1891</td>
<td>959</td>
<td>2734</td>
<td>2408</td>
<td>678</td>
<td>0</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>543</td>
<td>597</td>
<td>1494</td>
<td>1220</td>
<td>2300</td>
<td>923</td>
<td>205</td>
<td>2442</td>
<td>2329</td>
<td>0</td>
<td>Washington D.C.</td>
</tr>
</tbody>
</table>

A partial listing from the following statements include Output 37.1.2 and Output 37.1.3:

```
title 'Cluster Analysis of Flying Mileages Between 10 American Cities';
ods graphics on;

title2 'Using METHOD=AVERAGE';
proc cluster data=sashelp.mileages(type=distance) method=average pseudo;
    id City;
run;
```
Output 37.1.2 Cluster History Using METHOD=AVERAGE

Cluster Analysis of Flying Mileages Between 10 American Cities
Using METHOD=AVERAGE

The CLUSTER Procedure
Average Linkage Cluster Analysis

Cluster History

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Norm RMS Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>New York Washington D.C.</td>
<td>2</td>
<td>66.7</td>
<td>.</td>
<td>0.1297</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Los Angeles San Francisco</td>
<td>2</td>
<td>39.2</td>
<td>.</td>
<td>0.2196</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Atlanta Chicago</td>
<td>2</td>
<td>21.7</td>
<td>.</td>
<td>0.3715</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL7 CL9</td>
<td>4</td>
<td>14.5</td>
<td>3.4</td>
<td>0.4149</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL8 Seattle</td>
<td>3</td>
<td>12.4</td>
<td>7.3</td>
<td>0.5255</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Denver Houston</td>
<td>2</td>
<td>13.9</td>
<td>.</td>
<td>0.5562</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL6 Miami</td>
<td>5</td>
<td>15.5</td>
<td>3.8</td>
<td>0.6185</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3 CL4</td>
<td>7</td>
<td>16.0</td>
<td>5.3</td>
<td>0.8005</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL2 CL5</td>
<td>10</td>
<td>.</td>
<td>16.0</td>
<td>1.2967</td>
<td></td>
</tr>
</tbody>
</table>

Output 37.1.3 Dendrogram Using METHOD=AVERAGE
Example 37.1: Cluster Analysis of Flying Mileages between 10 American Cities

A partial listing from the following statements include Output 37.1.4 and Output 37.1.5:

```
title2 'Using METHOD=CENTROID';
proc cluster data=sashelp.mileages(type=distance) method=centroid pseudo;
   id City;
run;
```

**Output 37.1.4** Cluster History Using METHOD=CENTROID

Cluster Analysis of Flying Mileages Between 10 American Cities
Using METHOD=CENTROID

The CLUSTER Procedure
Centroid Hierarchical Cluster Analysis

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Norm Centroid Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>New York</td>
<td>2</td>
<td>66.7</td>
<td></td>
<td>0.1297</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Los Angeles</td>
<td>2</td>
<td>39.2</td>
<td></td>
<td>0.2196</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Atlanta</td>
<td>2</td>
<td>21.7</td>
<td></td>
<td>0.3715</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL7</td>
<td>4</td>
<td>14.5</td>
<td>3.4</td>
<td>0.3652</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL8</td>
<td>3</td>
<td>12.4</td>
<td>7.3</td>
<td>0.5139</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Denver</td>
<td>4</td>
<td>12.4</td>
<td>2.1</td>
<td>0.5337</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL6</td>
<td>5</td>
<td>14.2</td>
<td>3.8</td>
<td>0.5743</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3</td>
<td>6</td>
<td>22.1</td>
<td>2.6</td>
<td>0.6091</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL2 CL4</td>
<td>10</td>
<td></td>
<td>22.1</td>
<td>1.173</td>
<td></td>
</tr>
</tbody>
</table>

**Output 37.1.5** Dendrogram Using METHOD=CENTROID

Cluster Analysis

Distance Between Cluster Centroids
A partial listing from the following statements include Output 37.1.6 and Output 37.1.7:

```plaintext
title2 'Using METHOD=DENSITY K=3';
proc cluster data=sashelp.mileages(type=distance) method=density k=3;
   id City;
run;
```

**Output 37.1.6** Cluster History Using METHOD=DENSITY K=3

Cluster Analysis of Flying Mileages Between 10 American Cities
Using METHOD=DENSITY K=3

The CLUSTER Procedure
Density Linkage Cluster Analysis

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Normalized Fusion Density</th>
<th>Lesser</th>
<th>Greater</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Atlanta, Washington D.C.</td>
<td>96.106</td>
<td>92.5043</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CL9 Chicago</td>
<td>95.263</td>
<td>90.9548</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CL8 New York</td>
<td>86.465</td>
<td>76.1571</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL7 Houston</td>
<td>74.079</td>
<td>61.7747</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL6 Miami</td>
<td>74.079</td>
<td>58.8299</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Los Angeles, San Francisco</td>
<td>71.968</td>
<td>65.3430</td>
<td>80.0885</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL4 Seattle</td>
<td>66.341</td>
<td>56.6215</td>
<td>80.0885</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3 Denver</td>
<td>63.509</td>
<td>61.7747</td>
<td>80.0885</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL5, CL2</td>
<td>61.775</td>
<td></td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>

**Output 37.1.7** Dendrogram Using METHOD=DENSITY K=3

Cluster Analysis
A partial listing from the following statements include Output 37.1.8 and Output 37.1.9:

```sas
title2 'Using METHOD=SINGLE';
proc cluster data=sashelp.mileages(type=distance) method=single;
    id City;
run;
```

**Output 37.1.8** Cluster History Using METHOD=SINGLE

Cluster Analysis of Flying Mileages Between 10 American Cities
Using METHOD=SINGLE

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Norm Minimum Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>New York</td>
<td>Washington D.C.</td>
<td>2</td>
<td>0.1447</td>
</tr>
<tr>
<td>8</td>
<td>Los Angeles</td>
<td>San Francisco</td>
<td>2</td>
<td>0.2449</td>
</tr>
<tr>
<td>7</td>
<td>Atlanta</td>
<td>CL9</td>
<td>3</td>
<td>0.3832</td>
</tr>
<tr>
<td>6</td>
<td>CL7</td>
<td>Chicago</td>
<td>4</td>
<td>0.4142</td>
</tr>
<tr>
<td>5</td>
<td>CL6</td>
<td>Miami</td>
<td>5</td>
<td>0.4262</td>
</tr>
<tr>
<td>4</td>
<td>CL8</td>
<td>Seattle</td>
<td>3</td>
<td>0.4784</td>
</tr>
<tr>
<td>3</td>
<td>CL5</td>
<td>Houston</td>
<td>6</td>
<td>0.4947</td>
</tr>
<tr>
<td>2</td>
<td>Denver</td>
<td>CL4</td>
<td>4</td>
<td>0.5864</td>
</tr>
<tr>
<td>1</td>
<td>CL3 CL2</td>
<td>10</td>
<td>0.6203</td>
<td></td>
</tr>
</tbody>
</table>

**Output 37.1.9** Dendrogram Using METHOD=SINGLE
A partial listing from the following statements include Output 37.1.10 and Output 37.1.11:

```sas
title2 'Using METHOD=TWOSTAGE K=3';
proc cluster data=sashelp.mileages(type=distance) method=twostage k=3;
  id City;
run;
```

**Output 37.1.10** Cluster History Using METHOD=TWOSTAGE K=3

**Cluster Analysis of Flying Mileages Between 10 American Cities**
Using METHOD=TWOSTAGE K=3

The CLUSTER Procedure
Two-Stage Density Linkage Clustering

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Normalized Fusion Density</th>
<th>Lesser</th>
<th>Greater</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Atlanta</td>
<td>2</td>
<td>96.106</td>
<td>92.5043</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CL9</td>
<td>3</td>
<td>95.263</td>
<td>90.9548</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CL8</td>
<td>4</td>
<td>86.465</td>
<td>76.1571</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL7</td>
<td>5</td>
<td>74.079</td>
<td>61.7747</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL6</td>
<td>6</td>
<td>74.079</td>
<td>58.8299</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Los Angeles</td>
<td>2</td>
<td>71.968</td>
<td>65.3430</td>
<td>80.0885</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL4</td>
<td>3</td>
<td>66.341</td>
<td>56.6215</td>
<td>80.0885</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3</td>
<td>4</td>
<td>63.509</td>
<td>61.7747</td>
<td>80.0885</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL5 CL2</td>
<td>10</td>
<td>61.775</td>
<td>80.0885</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

**Output 37.1.11** Dendrogram Using METHOD=TWOSTAGE K=3

![Dendrogram Diagram]
Example 37.1: Cluster Analysis of Flying Mileages between 10 American Cities

A partial listing from the following statements include Output 37.1.12 and Output 37.1.13:

```
title2 'Using METHOD=WARD';
proc cluster data=sashelp.mileages(type=distance) method=ward pseudo;
   id City;
run;
```

**Output 37.1.12** Cluster History Using METHOD=WARD

Cluster Analysis of Flying Mileages Between 10 American Cities
Using METHOD=WARD

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>R-Square</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>New York</td>
<td>2</td>
<td>0.0019</td>
<td>.998</td>
<td>66.7</td>
<td></td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>Los Angeles</td>
<td>2</td>
<td>0.0054</td>
<td>.993</td>
<td>39.2</td>
<td></td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>Atlanta</td>
<td>2</td>
<td>0.0153</td>
<td>.977</td>
<td>21.7</td>
<td></td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>CL7, CL9</td>
<td>4</td>
<td>0.0296</td>
<td>.948</td>
<td>14.5</td>
<td>3.4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Denver, Houston</td>
<td>2</td>
<td>0.0344</td>
<td>.913</td>
<td>13.2</td>
<td></td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>CL8, Seattle</td>
<td>3</td>
<td>0.0391</td>
<td>.874</td>
<td>13.9</td>
<td>7.3</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL6, Miami</td>
<td>5</td>
<td>0.0586</td>
<td>.816</td>
<td>15.5</td>
<td>3.8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3, CL5</td>
<td>7</td>
<td>0.1488</td>
<td>.667</td>
<td>16.0</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL2, CL4</td>
<td>10</td>
<td>0.6669</td>
<td>.000</td>
<td>16.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Output 37.1.13** Dendrogram Using METHOD=WARD
Example 37.2: Crude Birth and Death Rates

This example uses the SAS data set Poverty created in the section “Getting Started: CLUSTER Procedure” on page 2427. 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 three and eight 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 SGPLOT 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.

```plaintext
title 'Cluster Analysis of Birth and Death Rates';
ods graphics on;

%macro analyze(method,ncl);
   proc cluster data=poverty outtree=tree method=&method print=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;
   %end;
%end;
```

Example 37.2: Crude Birth and Death Rates

This example uses the SAS data set Poverty created in the section “Getting Started: CLUSTER Procedure” on page 2427. 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 three and eight 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 SGPLOT 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.

```plaintext
title 'Cluster Analysis of Birth and Death Rates';
ods graphics on;

%macro analyze(method,ncl);
   proc cluster data=poverty outtree=tree method=&method print=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;
   %end;
%end;
```
Example 37.2: Crude Birth and Death Rates

```sas
proc sgplot;
  scatter y=death x=birth / group=cluster;
  title2 "Plot of &n Clusters from METHOD=&METHOD";
run;

%let k=%eval(&k+1);
%let n=%scan(&ncl,&k);
%end;
%mend;

The following statement produces Output 37.2.1, Output 37.2.3, and Output 37.2.4:

%analyze(average, 3 8)

For average linkage, the CCC has peaks at three, eight, ten, and twelve clusters, but the three-cluster peak is lower than the eight-cluster peak. The pseudo $F$ statistic has peaks at three, eight, and twelve clusters. The pseudo $t^2$ statistic drops sharply at three clusters, continues to fall at four clusters, and has a particularly low value at twelve clusters. However, there are not enough data to seriously consider as many as twelve clusters. Scatter plots are given for three and eight clusters. The results are shown in Output 37.2.1 through Output 37.2.4. In Output 37.2.4, the eighth cluster consists of the two outlying observations, Mexico and Korea.

Output 37.2.1  Cluster Analysis for Birth and Death Rates: METHOD=AVERAGE

Cluster Analysis of Birth and Death Rates
The CLUSTER Procedure
Average Linkage Cluster Analysis

<table>
<thead>
<tr>
<th>Eigenvalues of the Covariance Matrix</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
<td>1</td>
<td>189.106588</td>
<td>173.101020</td>
<td>0.9220</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>16.005568</td>
<td>0.0780</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Root-Mean-Square Total-Sample Standard Deviation  10.127

Root-Mean-Square Distance Between Observations  20.25399
```
Cluster History

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

**Output 37.2.2 Criteria for the Number of Clusters: METHOD=AVERAGE**

![Graph showing criteria for the number of clusters](image)
Example 37.2: Crude Birth and Death Rates

Output 37.2.3  Plot of Three Clusters: METHOD=AVERAGE

Output 37.2.4  Plot of Eight Clusters: METHOD=AVERAGE
The following statement produces Output 37.2.5 and Output 37.2.7:

```plaintext
%analyze(complete, 3)
```

Complete linkage shows CCC peaks at three, eight and twelve clusters. The pseudo $F$ statistic peaks at three and twelve clusters. The pseudo $t^2$ statistic indicates three clusters.

The scatter plot for three clusters is shown.

**Output 37.2.5**  Cluster History 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

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>189.106588</td>
<td>173.101020</td>
<td>0.9220</td>
</tr>
<tr>
<td>2</td>
<td>16.005568</td>
<td>0.0780</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Root-Mean-Square Total-Sample Standard Deviation 10.127

Mean Distance Between Observations 17.13099

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>R-Square</th>
<th>Approximate Expected R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t Squared</th>
<th>Norm Maximum Distance Ties</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>CL22 CL33</td>
<td>8</td>
<td>0.0015</td>
<td>.983</td>
<td>.975</td>
<td>3.80</td>
<td>329</td>
<td>6.1</td>
<td>0.4092</td>
</tr>
<tr>
<td>14</td>
<td>CL56 CL18</td>
<td>8</td>
<td>0.0014</td>
<td>.981</td>
<td>.972</td>
<td>3.97</td>
<td>331</td>
<td>6.6</td>
<td>0.4255</td>
</tr>
<tr>
<td>13</td>
<td>CL30 CL44</td>
<td>8</td>
<td>0.0019</td>
<td>.979</td>
<td>.969</td>
<td>4.04</td>
<td>330</td>
<td>19.0</td>
<td>0.4332</td>
</tr>
<tr>
<td>12</td>
<td>OB23 OB61</td>
<td>2</td>
<td>0.0014</td>
<td>.978</td>
<td>.966</td>
<td>4.45</td>
<td>340</td>
<td>.175</td>
<td>0.4378</td>
</tr>
<tr>
<td>11</td>
<td>CL19 CL24</td>
<td>24</td>
<td>0.0034</td>
<td>.974</td>
<td>.962</td>
<td>4.17</td>
<td>327</td>
<td>24.1</td>
<td>0.4962</td>
</tr>
<tr>
<td>10</td>
<td>CL17 CL28</td>
<td>12</td>
<td>0.0033</td>
<td>.971</td>
<td>.957</td>
<td>4.18</td>
<td>325</td>
<td>14.8</td>
<td>0.5204</td>
</tr>
<tr>
<td>9</td>
<td>CL20 CL13</td>
<td>16</td>
<td>0.0067</td>
<td>.964</td>
<td>.951</td>
<td>3.38</td>
<td>297</td>
<td>25.2</td>
<td>0.5236</td>
</tr>
<tr>
<td>8</td>
<td>CL11 CL21</td>
<td>32</td>
<td>0.0054</td>
<td>.959</td>
<td>.943</td>
<td>3.44</td>
<td>297</td>
<td>19.7</td>
<td>0.6001</td>
</tr>
<tr>
<td>7</td>
<td>CL26 CL15</td>
<td>13</td>
<td>0.0096</td>
<td>.949</td>
<td>.933</td>
<td>2.93</td>
<td>282</td>
<td>28.9</td>
<td>0.7233</td>
</tr>
<tr>
<td>6</td>
<td>CL14 CL10</td>
<td>20</td>
<td>0.0128</td>
<td>.937</td>
<td>.920</td>
<td>2.46</td>
<td>269</td>
<td>27.7</td>
<td>0.8033</td>
</tr>
<tr>
<td>5</td>
<td>CL9 CL16</td>
<td>30</td>
<td>0.0237</td>
<td>.913</td>
<td>.902</td>
<td>1.29</td>
<td>241</td>
<td>47.1</td>
<td>0.8993</td>
</tr>
<tr>
<td>4</td>
<td>CL6 CL7</td>
<td>33</td>
<td>0.0240</td>
<td>.889</td>
<td>.875</td>
<td>1.38</td>
<td>248</td>
<td>21.7</td>
<td>1.2165</td>
</tr>
<tr>
<td>3</td>
<td>CL5 CL12</td>
<td>32</td>
<td>0.0178</td>
<td>.871</td>
<td>.827</td>
<td>2.56</td>
<td>317</td>
<td>13.6</td>
<td>1.2326</td>
</tr>
<tr>
<td>2</td>
<td>CL3 CL8</td>
<td>64</td>
<td>0.1900</td>
<td>.681</td>
<td>.697</td>
<td>-5.5</td>
<td>203</td>
<td>167</td>
<td>1.5412</td>
</tr>
<tr>
<td>1</td>
<td>CL2 CL4</td>
<td>97</td>
<td>0.6810</td>
<td>.000</td>
<td>.000</td>
<td>0.00</td>
<td>.203</td>
<td>2.5233</td>
<td></td>
</tr>
</tbody>
</table>
Example 37.2: Crude Birth and Death Rates

**Output 37.2.6** Criteria for the Number of Clusters: METHOD=COMPLETE

![Criteria for the Number of Clusters](chart)

**Output 37.2.7** Plot of Clusters for METHOD=COMPLETE

![Cluster Analysis of Birth and Death Rates](chart)
The following statement produces Output 37.2.8 and Output 37.2.10:

```
%analyze(single, 7 10)
```

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 seven or ten clusters.

The scatter plots for seven and ten clusters are shown.

### Output 37.2.8  Cluster History 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

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>189.106588</td>
<td>173.101020</td>
<td>0.9220</td>
</tr>
<tr>
<td>2</td>
<td>16.005568</td>
<td>0.0780</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Root-Mean-Square Total-Sample Standard Deviation 10.127

Mean Distance Between Observations 17.13099

<table>
<thead>
<tr>
<th>Number of Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>R-Square</th>
<th>Approximate Expected R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared Norm Minimum Distance Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 CL37 CL19</td>
<td>8</td>
<td>0.0014</td>
<td>0.968</td>
<td>0.975</td>
<td>-2.3</td>
<td>178</td>
<td>6.6</td>
</tr>
<tr>
<td>14 CL20 CL23</td>
<td>15</td>
<td>0.0059</td>
<td>0.962</td>
<td>0.972</td>
<td>-3.1</td>
<td>162</td>
<td>18.7</td>
</tr>
<tr>
<td>13 CL14 CL16</td>
<td>19</td>
<td>0.0054</td>
<td>0.957</td>
<td>0.969</td>
<td>-3.4</td>
<td>155</td>
<td>8.8</td>
</tr>
<tr>
<td>12 CL26 OB58</td>
<td>31</td>
<td>0.0014</td>
<td>0.955</td>
<td>0.966</td>
<td>-2.7</td>
<td>165</td>
<td>4.0</td>
</tr>
<tr>
<td>11 OB86 CL18</td>
<td>4</td>
<td>0.0003</td>
<td>0.955</td>
<td>0.962</td>
<td>-1.6</td>
<td>183</td>
<td>3.8</td>
</tr>
<tr>
<td>10 CL13 CL11</td>
<td>23</td>
<td>0.0088</td>
<td>0.946</td>
<td>0.957</td>
<td>-2.3</td>
<td>170</td>
<td>11.3</td>
</tr>
<tr>
<td>9 CL22 CL17</td>
<td>30</td>
<td>0.0235</td>
<td>0.923</td>
<td>0.951</td>
<td>-4.7</td>
<td>131</td>
<td>45.7</td>
</tr>
<tr>
<td>8 CL15 CL10</td>
<td>31</td>
<td>0.0210</td>
<td>0.902</td>
<td>0.943</td>
<td>-5.8</td>
<td>117</td>
<td>21.8</td>
</tr>
<tr>
<td>7 CL9 OB75</td>
<td>31</td>
<td>0.0052</td>
<td>0.897</td>
<td>0.933</td>
<td>-4.7</td>
<td>130</td>
<td>4.0</td>
</tr>
<tr>
<td>6 CL7 CL12</td>
<td>62</td>
<td>0.2023</td>
<td>0.694</td>
<td>0.920</td>
<td>-15</td>
<td>41.3</td>
<td>223</td>
</tr>
<tr>
<td>5 CL6 CL8</td>
<td>93</td>
<td>0.6681</td>
<td>0.026</td>
<td>0.902</td>
<td>-26</td>
<td>0.6</td>
<td>199</td>
</tr>
<tr>
<td>4 CL5 OB48</td>
<td>94</td>
<td>0.0056</td>
<td>0.021</td>
<td>0.875</td>
<td>-24</td>
<td>0.7</td>
<td>0.5</td>
</tr>
<tr>
<td>3 CL4 OB67</td>
<td>95</td>
<td>0.0083</td>
<td>0.012</td>
<td>0.827</td>
<td>-15</td>
<td>0.6</td>
<td>0.8</td>
</tr>
<tr>
<td>2 OB23 OB61</td>
<td>2</td>
<td>0.0014</td>
<td>0.011</td>
<td>0.697</td>
<td>-13</td>
<td>1.0</td>
<td>.</td>
</tr>
<tr>
<td>1 CL3 CL2</td>
<td>97</td>
<td>0.0109</td>
<td>0.000</td>
<td>0.000</td>
<td>0.0</td>
<td>1.0</td>
<td>0.5815</td>
</tr>
</tbody>
</table>
Output 37.2.9 Criteria for the Number of Clusters: METHOD=SINGLE

Criteria for the Number of Clusters

<table>
<thead>
<tr>
<th>CCC</th>
<th>Pseudo F</th>
<th>Ps-TSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>-25</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>-20</td>
<td>50</td>
<td>50</td>
</tr>
<tr>
<td>-15</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>-10</td>
<td>150</td>
<td>150</td>
</tr>
<tr>
<td>0</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

Number of Clusters
Output 37.2.10  Plot of Clusters for METHOD=SINGLE

Cluster Analysis of Birth and Death Rates
Plot of 7 Clusters from METHOD=single

Cluster Analysis of Birth and Death Rates
Plot of 10 Clusters from METHOD=single
The following statements produce Output 37.2.11 through Output 37.2.14:

```plaintext
%analyze(two k=10, 3)
%analyze(two k=18, 2)
```

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

<table>
<thead>
<tr>
<th>$k$</th>
<th>Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5-7</td>
<td>4</td>
</tr>
<tr>
<td>8-15</td>
<td>3</td>
</tr>
<tr>
<td>16-21</td>
<td>2</td>
</tr>
<tr>
<td>22+</td>
<td>1</td>
</tr>
</tbody>
</table>

Thus, there is strong evidence of three modes and an indication of the possibility of two modes. Uniform-kernel density linkage gives similar results. For $K=10$ (10th-nearest-neighbor density linkage), the scatter plot for three clusters is shown; and for $K=18$, the scatter plot for two clusters is shown.

**Output 37.2.11** Cluster History for Birth and Death Rates: METHOD=TWO_STAGE K=10

Cluster Analysis of Birth and Death Rates

The CLUSTER Procedure
Two-Stage Density Linkage Clustering

<table>
<thead>
<tr>
<th>Eigenvalues of the Covariance Matrix</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>189.106588</td>
<td>173.101020</td>
<td>0.9220</td>
<td>0.9220</td>
</tr>
<tr>
<td>2</td>
<td>16.005568</td>
<td>0.0780</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

$K = 10$

Root-Mean-Square Total-Sample Standard Deviation 10.127
3 modal clusters have been formed.
### Output 37.2.12  Cluster History 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

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>189.106588</td>
<td>173.101020</td>
<td>0.9220</td>
</tr>
<tr>
<td>2</td>
<td>16.005568</td>
<td>0.0780</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

K = 18

Root-Mean-Square Total-Sample Standard Deviation 10.127

#### Cluster History

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>Approximate R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Normalized Fusion Density</th>
<th>Lesser</th>
<th>Greater</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>CL16 OB72</td>
<td>46</td>
<td>0.0107</td>
<td>.799</td>
<td>-21</td>
<td>23.3</td>
<td>3.0</td>
<td>10.118</td>
<td>7.7445</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>CL15 OB94</td>
<td>47</td>
<td>0.0098</td>
<td>.789</td>
<td>-21</td>
<td>23.9</td>
<td>2.7</td>
<td>9.676</td>
<td>7.1257</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>CL14 OB51</td>
<td>48</td>
<td>0.0037</td>
<td>.786</td>
<td>-20</td>
<td>25.6</td>
<td>1.0</td>
<td>9.409</td>
<td>6.8398</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>CL13 OB96</td>
<td>49</td>
<td>0.0099</td>
<td>.776</td>
<td>-19</td>
<td>26.7</td>
<td>2.6</td>
<td>9.409</td>
<td>6.8398</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>CL12 OB76</td>
<td>50</td>
<td>0.0114</td>
<td>.764</td>
<td>-19</td>
<td>27.9</td>
<td>2.9</td>
<td>8.8136</td>
<td>6.3138</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CL11 OB77</td>
<td>51</td>
<td>0.0021</td>
<td>.762</td>
<td>-18</td>
<td>31.0</td>
<td>0.5</td>
<td>8.6593</td>
<td>6.0751</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>CL10 OB78</td>
<td>52</td>
<td>0.0103</td>
<td>.752</td>
<td>-17</td>
<td>33.3</td>
<td>2.5</td>
<td>8.6007</td>
<td>6.0976</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CL9 OB43</td>
<td>53</td>
<td>0.0034</td>
<td>.748</td>
<td>-16</td>
<td>37.8</td>
<td>0.8</td>
<td>8.4964</td>
<td>5.9160</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CL8 OB93</td>
<td>54</td>
<td>0.0109</td>
<td>.737</td>
<td>-15</td>
<td>42.1</td>
<td>2.6</td>
<td>8.367</td>
<td>5.7913</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL7 OB88</td>
<td>55</td>
<td>0.0110</td>
<td>.726</td>
<td>-13</td>
<td>48.3</td>
<td>2.6</td>
<td>7.916</td>
<td>5.3679</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL6 OB87</td>
<td>56</td>
<td>0.0120</td>
<td>.714</td>
<td>-12</td>
<td>57.5</td>
<td>2.7</td>
<td>6.6917</td>
<td>4.3415</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CL20 OB61</td>
<td>39</td>
<td>0.0077</td>
<td>.707</td>
<td>-9.8</td>
<td>74.7</td>
<td>8.3</td>
<td>6.2578</td>
<td>3.2882</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL5 OB82</td>
<td>57</td>
<td>0.0138</td>
<td>.693</td>
<td>-5.0</td>
<td>106</td>
<td>3.0</td>
<td>5.3605</td>
<td>3.2834</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3 OB23</td>
<td>58</td>
<td>0.0117</td>
<td>.681</td>
<td>-54</td>
<td>203</td>
<td>2.5</td>
<td>3.2687</td>
<td>1.7568</td>
<td>23.457</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL2 CL4</td>
<td>97</td>
<td>0.6812</td>
<td>.000</td>
<td>0.0</td>
<td>203</td>
<td>.2</td>
<td>13.764</td>
<td>23.457</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

2 modal clusters have been formed.
Chapter 37: The CLUSTER Procedure

Output 37.2.13  Plot of Clusters for METHOD=TWOSTAGE K=10

Cluster Analysis of Birth and Death Rates
Plot of 3 Clusters from METHOD=two k=10

Output 37.2.14  Plot of Clusters for METHOD=TWOSTAGE K=18

Cluster Analysis of Birth and Death Rates
Plot of 2 Clusters from METHOD=two k=18
In summary, most of the clustering methods indicate three or eight clusters. Most methods agree at the three-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 37.4.

Example 37.3: Cluster Analysis of Fisher's 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 50 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. The following statements create the data set Iris:

```plaintext
proc format;
  value specname
    1='Setosa '  
    2='Versicolor'
    3='Virginica ';  
run;

data iris;
  title 'Fisher (1936) Iris Data';
  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.';
  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
... more lines ...
  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 
;
```

This example analyzes the iris data by using 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.

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 32, “The CANDISC Procedure,” for a canonical discriminant analysis of the iris species.
/**** Define macro show ****/
%macro show;
  proc freq;
    tables cluster*species / nopercent norow nocol plot=none;
  run;

  proc candisc noprint out=can;
    class cluster;
    var petal: sepal:;
  run;

  proc sgplot data=can;
    scatter y=can2 x=can1 / group=cluster;
  run;
%mend;

The first analysis clusters the iris data by using Ward’s method (see Output 37.3.1) and plots the CCC and pseudo $F$ and $t^2$ statistics (see Output 37.3.2). The CCC has a local peak at three clusters but a higher peak at five clusters. The pseudo $F$ statistic indicates three clusters, while the pseudo $t^2$ statistic suggests three or six clusters.

The TREE procedure creates an output data set containing the three-cluster partition for use by the SHOW macro. The FREQ procedure reveals 16 misclassifications. The results are shown in Output 37.3.3.

title2 'By Ward's Method';
ods graphics on;

proc cluster data=iris method=ward print=15 ccc pseudo;
  var petal: sepal:;
  copy species;
run;

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

%show;

Output 37.3.1 Cluster Analysis of Fisher’s Iris Data: PROC CLUSTER with METHOD=WARD

Fisher (1936) Iris Data
By Ward’s Method

The CLUSTER Procedure
Ward's Minimum Variance Cluster Analysis

<table>
<thead>
<tr>
<th>Eigenvalues of the Covariance Matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eigenvalue</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
</tbody>
</table>
Output 37.3.1 continued

Output 37.3.2 Criteria for the Number of Clusters with METHOD=WARD

### Cluster History

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

### Criteria for the Number of Clusters

![Graph showing CCC, Pseudo F, and P*TSq criteria vs. Number of Clusters]
Output 37.3.3  Crosstabulation of Clusters for METHOD=WARD

Fisher (1936) Iris Data
By Ward's Method

The FREQ Procedure

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Table of CLUSTER by Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>Setosa Versicolor Virginica Total</td>
</tr>
<tr>
<td>CLUSTER</td>
<td>1   0  49  15 64</td>
</tr>
<tr>
<td>1</td>
<td>0   1  35  36</td>
</tr>
<tr>
<td>2</td>
<td>50  0  0   50</td>
</tr>
<tr>
<td>3</td>
<td>50  50 50  150</td>
</tr>
<tr>
<td>Total</td>
<td>50  50 50  150</td>
</tr>
</tbody>
</table>

Output 37.3.4  Scatter Plot of Clusters for METHOD=WARD

Fisher (1936) Iris Data
By Ward's Method

Can2

Can1

CLUSTER 1 2 3
The second analysis uses two-stage density linkage. The raw data suggest two or six modes instead of three:

<table>
<thead>
<tr>
<th>$k$</th>
<th>Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>4-6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9-50</td>
<td>2</td>
</tr>
<tr>
<td>51+</td>
<td>1</td>
</tr>
</tbody>
</table>

The following analysis uses $K=8$ to produce three clusters for comparison with other analyses. There are only six misclassifications. The results are shown in Output 37.3.5 and Output 37.3.6.

```
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 37.3.5** Cluster Analysis of Fisher’s Iris Data: PROC CLUSTER with METHOD=TWO-STAGE

**Fisher (1936) Iris Data**
**By Two-Stage Density Linkage**

**The CLUSTER Procedure**
**Two-Stage Density Linkage Clustering**

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 422.824171</td>
<td>398.557096</td>
<td>0.9246</td>
<td>0.9246</td>
</tr>
<tr>
<td>2 24.267075</td>
<td>16.446125</td>
<td>0.0531</td>
<td>0.9777</td>
</tr>
<tr>
<td>3 7.820950</td>
<td>5.437441</td>
<td>0.0171</td>
<td>0.9948</td>
</tr>
<tr>
<td>4 2.383509</td>
<td>0.0052</td>
<td>1.0000</td>
<td></td>
</tr>
</tbody>
</table>

$K = 8$

Root-Mean-Square Total-Sample Standard Deviation 10.69224
### Output 37.3.5 continued

#### Cluster History

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>Approximate Expected R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Normalized Fusion Density</th>
<th>Lesser</th>
<th>Greater</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 CL17 OB127</td>
<td>43</td>
<td>0.0024</td>
<td>.917</td>
<td>.958</td>
<td>-11</td>
<td>107</td>
<td>3.4</td>
<td>0.3903</td>
<td>0.2066</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>14 CL16 OB137</td>
<td>50</td>
<td>0.0023</td>
<td>.915</td>
<td>.955</td>
<td>-10</td>
<td>113</td>
<td>5.6</td>
<td>0.3637</td>
<td>0.1837</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>13 CL15 OB74</td>
<td>44</td>
<td>0.0029</td>
<td>.912</td>
<td>.953</td>
<td>-9.8</td>
<td>119</td>
<td>3.8</td>
<td>0.3553</td>
<td>0.2130</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>12 CL22 OB49</td>
<td>47</td>
<td>0.0036</td>
<td>.909</td>
<td>.950</td>
<td>-7.7</td>
<td>125</td>
<td>5.2</td>
<td>0.3223</td>
<td>0.1736</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>11 CL12 OB85</td>
<td>48</td>
<td>0.0036</td>
<td>.905</td>
<td>.946</td>
<td>-7.4</td>
<td>132</td>
<td>4.8</td>
<td>0.3223</td>
<td>0.1736</td>
<td>8.3678</td>
<td></td>
</tr>
<tr>
<td>10 CL11 OB98</td>
<td>49</td>
<td>0.0033</td>
<td>.902</td>
<td>.942</td>
<td>-6.8</td>
<td>143</td>
<td>4.1</td>
<td>0.2879</td>
<td>0.1479</td>
<td>8.3678</td>
<td></td>
</tr>
<tr>
<td>9 CL13 OB24</td>
<td>45</td>
<td>0.0036</td>
<td>.898</td>
<td>.936</td>
<td>-6.2</td>
<td>155</td>
<td>4.5</td>
<td>0.2802</td>
<td>0.2005</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>8 CL10 OB25</td>
<td>50</td>
<td>0.0019</td>
<td>.896</td>
<td>.930</td>
<td>-5.2</td>
<td>175</td>
<td>2.2</td>
<td>0.2699</td>
<td>0.1372</td>
<td>8.3678</td>
<td></td>
</tr>
<tr>
<td>7 CL8 OB121</td>
<td>51</td>
<td>0.0035</td>
<td>.893</td>
<td>.921</td>
<td>-4.2</td>
<td>198</td>
<td>4.0</td>
<td>0.2586</td>
<td>0.1372</td>
<td>8.3678</td>
<td></td>
</tr>
<tr>
<td>6 CL9 OB45</td>
<td>46</td>
<td>0.0041</td>
<td>.888</td>
<td>.911</td>
<td>-3.0</td>
<td>229</td>
<td>4.7</td>
<td>0.1412</td>
<td>0.0832</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>5 CL6 OB39</td>
<td>47</td>
<td>0.0048</td>
<td>.884</td>
<td>.895</td>
<td>-1.5</td>
<td>276</td>
<td>5.1</td>
<td>0.107</td>
<td>0.0605</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>4 CL5 OB21</td>
<td>48</td>
<td>0.0048</td>
<td>.879</td>
<td>.872</td>
<td>0.54</td>
<td>353</td>
<td>4.7</td>
<td>0.0969</td>
<td>0.0541</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>3 CL4 OB90</td>
<td>49</td>
<td>0.0046</td>
<td>.874</td>
<td>.827</td>
<td>3.49</td>
<td>511</td>
<td>4.2</td>
<td>0.0715</td>
<td>0.0370</td>
<td>3.5156</td>
<td></td>
</tr>
<tr>
<td>2 CL7 CL3</td>
<td>100</td>
<td>0.1017</td>
<td>.773</td>
<td>.697</td>
<td>3.83</td>
<td>503</td>
<td>96.3</td>
<td>2.6277</td>
<td>3.5156</td>
<td>8.3678</td>
<td></td>
</tr>
</tbody>
</table>

3 modal clusters have been formed.

### Output 37.3.6 Criteria for the Number of Clusters with METHOD=TWOSTAGE

![Criteria for the Number of Clusters](image-url)
Output 37.3.7  Crosstabulation of Clusters for METHOD=TWO STAGE

Fisher (1936) Iris Data
By Two-Stage Density Linkage

The FREQ Procedure

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Table of CLUSTER by Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>Setosa  Versicolor  Virginica Total</td>
</tr>
<tr>
<td>CLUSTER</td>
<td>1     50          0         0  50</td>
</tr>
<tr>
<td></td>
<td>2     0           48        3  51</td>
</tr>
<tr>
<td></td>
<td>3     0           2         47 49</td>
</tr>
<tr>
<td>Total</td>
<td>50    50          50       150</td>
</tr>
</tbody>
</table>

Output 37.3.8  Scatter Plot of Clusters for METHOD=TWO STAGE
The CLUSTER procedure is not practical for very large data sets because, with most methods, the CPU time is roughly proportional to 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 to produce a large number of clusters and then use PROC CLUSTER to hierarchically cluster the preliminary clusters.

FASTCLUS automatically creates the 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 following SAS code uses the iris data 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 37.3.9 and Output 37.3.10.

```sas
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 / nopercent norow nocol plot=none;
run;

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

**Output 37.3.9** Preliminary Analysis of Fisher's Iris Data: FASTCLUS Procedure

**Fisher (1936) Iris Data**
**Preliminary Analysis by FASTCLUS**

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

---

Convergence criterion is satisfied.

**Criterion Based on Final Seeds = 2.1389**
Example 37.3: Cluster Analysis of Fisher’s Iris Data

Output 37.3.9 continued

Cluster Summary

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

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.

Output 37.3.10 Crosstabulation of Species and Cluster From the FASTCLUS Procedure

Fisher (1936) Iris Data
Preliminary Analysis by FASTCLUS

The FREQ Procedure

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Table of preclus by Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>preclus(Cluster)</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>33</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
</tr>
</tbody>
</table>
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 three-cluster partition, which is sorted and merged with the OUT= data set from PROC FASTCLUS to determine which cluster each of the original 150 observations belongs to. 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.

```sas
/***** Define macro clus *****/
%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 out prelim;
        by preclus;
    run;

    %show;
%mend;
```

The following statements produce Output 37.3.11 through Output 37.3.14.

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

### Output 37.3.11

Clustering Clusters by Ward’s Method

Fisher (1936) Iris Data
Clustering Clusters by Ward’s Method

The CLUSTER Procedure
Ward’s Minimum Variance Cluster Analysis

<table>
<thead>
<tr>
<th>Eigenvalues of the Covariance Matrix</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 416.976349</td>
<td>398.666421</td>
<td>0.9501</td>
<td>0.9501</td>
<td></td>
</tr>
<tr>
<td>2 18.309928</td>
<td>14.952922</td>
<td>0.0417</td>
<td>0.9918</td>
<td></td>
</tr>
<tr>
<td>3 3.357006</td>
<td>3.126943</td>
<td>0.0076</td>
<td>0.9995</td>
<td></td>
</tr>
<tr>
<td>4 0.230063</td>
<td>0.230063</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root-Mean-Square Total-Sample Standard Deviation 10.69224

Root-Mean-Square Distance Between Observations 30.24221
### Output 37.3.11  continued

<table>
<thead>
<tr>
<th>Number of Clusters Joined</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>Approximate Expected R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 OB2 OB4</td>
<td></td>
<td>23</td>
<td>0.0019</td>
<td>.958</td>
<td>.932</td>
<td>6.26</td>
<td>400</td>
<td>6.3</td>
</tr>
<tr>
<td>8 OB1 OB5</td>
<td></td>
<td>12</td>
<td>0.0025</td>
<td>.955</td>
<td>.926</td>
<td>6.75</td>
<td>434</td>
<td>5.8</td>
</tr>
<tr>
<td>7 CL9 OB6</td>
<td></td>
<td>30</td>
<td>0.0069</td>
<td>.948</td>
<td>.918</td>
<td>6.28</td>
<td>438</td>
<td>19.5</td>
</tr>
<tr>
<td>6 OB3 OB8</td>
<td></td>
<td>36</td>
<td>0.0074</td>
<td>.941</td>
<td>.907</td>
<td>6.21</td>
<td>459</td>
<td>26.0</td>
</tr>
<tr>
<td>5 OB7 OB10</td>
<td></td>
<td>50</td>
<td>0.0104</td>
<td>.931</td>
<td>.892</td>
<td>6.15</td>
<td>485</td>
<td>42.2</td>
</tr>
<tr>
<td>4 CL8 OB9</td>
<td></td>
<td>34</td>
<td>0.0162</td>
<td>.914</td>
<td>.870</td>
<td>4.28</td>
<td>519</td>
<td>39.3</td>
</tr>
<tr>
<td>3 CL7 CL6</td>
<td></td>
<td>66</td>
<td>0.0318</td>
<td>.883</td>
<td>.824</td>
<td>4.39</td>
<td>552</td>
<td>59.7</td>
</tr>
<tr>
<td>2 CL4 CL3</td>
<td></td>
<td>100</td>
<td>0.1099</td>
<td>.773</td>
<td>.695</td>
<td>3.94</td>
<td>503</td>
<td>113</td>
</tr>
<tr>
<td>1 CL2 CL5</td>
<td></td>
<td>150</td>
<td>0.7726</td>
<td>.000</td>
<td>.000</td>
<td>0.00</td>
<td>503</td>
<td></td>
</tr>
</tbody>
</table>

### Output 37.3.12  Criteria for the Number of Clusters for Clustering Clusters from Ward's Method

![Criteria for the Number of Clusters](chart)

- CCC
- Pseudo F
- Pseudo t-Squared

**Number of Clusters**

- CCC: Continuously increasing from 0 to 6.
- Pseudo F: Decreasing from 550 to 0.
- Pseudo t-Squared: Decreasing from 550 to 0.

**Number of Clusters**

- From 2 to 8.
Output 37.3.13  Crosstabulation for Clustering Clusters from Ward’s Method

Fisher (1936) Iris Data
Clustering Clusters by Ward’s Method

The FREQ Procedure

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Table of CLUSTER by Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>Setosa  Versicolor  Virginica</td>
</tr>
<tr>
<td>CLUSTER</td>
<td>1     0      50  16  66</td>
</tr>
<tr>
<td></td>
<td>2     0      0   34  34</td>
</tr>
<tr>
<td></td>
<td>3     50     0   0   50</td>
</tr>
<tr>
<td>Total</td>
<td>50    50     50  150</td>
</tr>
</tbody>
</table>

Output 37.3.14  Scatter Plot for Clustering Clusters using Ward’s Method
Example 37.3: Cluster Analysis of Fisher’s Iris Data

The following statements produce Output 37.3.15 through Output 37.3.17.

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

**Output 37.3.15**  Clustering Clusters by Wong's Hybrid Method

**Fisher (1936) Iris Data**

**Clustering Clusters by Wong’s Hybrid Method**

The CLUSTER Procedure

Two-Stage Density Linkage Clustering

Eigenvalues of the Covariance Matrix

<table>
<thead>
<tr>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>416.976349</td>
<td>398.666421</td>
<td>0.9501</td>
</tr>
<tr>
<td>2</td>
<td>18.309928</td>
<td>14.952922</td>
<td>0.0417</td>
</tr>
<tr>
<td>3</td>
<td>3.357006</td>
<td>3.126943</td>
<td>0.0076</td>
</tr>
<tr>
<td>4</td>
<td>0.230063</td>
<td>0.0005</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Root-Mean-Square Total-Sample Standard Deviation 10.69224

Cluster History

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>Semipartial R-Square</th>
<th>R-Square</th>
<th>Approximate Expected R-Square</th>
<th>Cubic Clustering Criterion</th>
<th>Pseudo F Statistic</th>
<th>Pseudo t-Squared</th>
<th>Normalized Fusion Density</th>
<th>Lesser</th>
<th>Greater</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>OB10 OB7</td>
<td>50</td>
<td>0.0104</td>
<td>.949</td>
<td>.932</td>
<td>3.81</td>
<td>330</td>
<td>42.2</td>
<td>40.24</td>
<td>58.2179</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>OB3 OB8</td>
<td>36</td>
<td>0.0074</td>
<td>.942</td>
<td>.926</td>
<td>3.22</td>
<td>329</td>
<td>26.0</td>
<td>27.981</td>
<td>39.4511</td>
<td>48.4350</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>OB2 OB4</td>
<td>23</td>
<td>0.0019</td>
<td>.940</td>
<td>.918</td>
<td>4.24</td>
<td>373</td>
<td>6.3</td>
<td>23.775</td>
<td>8.9675</td>
<td>46.3026</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL8 OB9</td>
<td>58</td>
<td>0.0194</td>
<td>.921</td>
<td>.907</td>
<td>2.13</td>
<td>334</td>
<td>46.3</td>
<td>20.724</td>
<td>46.8846</td>
<td>48.4350</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL7 OB6</td>
<td>30</td>
<td>0.0069</td>
<td>.914</td>
<td>.892</td>
<td>3.09</td>
<td>383</td>
<td>19.5</td>
<td>13.303</td>
<td>17.6360</td>
<td>46.3026</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CL6 OB1</td>
<td>67</td>
<td>0.0292</td>
<td>.884</td>
<td>.870</td>
<td>1.21</td>
<td>372</td>
<td>41.0</td>
<td>8.4137</td>
<td>10.8758</td>
<td>48.4350</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL4 OB5</td>
<td>70</td>
<td>0.0138</td>
<td>.871</td>
<td>.824</td>
<td>3.33</td>
<td>494</td>
<td>12.3</td>
<td>5.1855</td>
<td>6.2890</td>
<td>48.4350</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3 CL5</td>
<td>100</td>
<td>0.0979</td>
<td>.773</td>
<td>.695</td>
<td>3.94</td>
<td>503</td>
<td>89.5</td>
<td>19.513</td>
<td>46.3026</td>
<td>48.4350</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL2 CL9</td>
<td>150</td>
<td>0.7726</td>
<td>.000</td>
<td>.000</td>
<td>0.00</td>
<td>.</td>
<td>503</td>
<td>1.3337</td>
<td>48.4350</td>
<td>100.0</td>
<td></td>
</tr>
</tbody>
</table>

3 modal clusters have been formed.
**Output 37.3.16**  Crosstabulation for Clustering Clusters from Wong’s Hybrid Method

**Fisher (1936) Iris Data**

**Clustering Clusters by Wong’s Hybrid Method**

**The FREQ Procedure**

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Table of CLUSTER by Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Species</td>
</tr>
<tr>
<td></td>
<td>Setosa</td>
</tr>
<tr>
<td>CLUSTER</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>50</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
</tr>
</tbody>
</table>

**Output 37.3.17**  Scatter Plot for Clustering Clusters using Wong’s Hybrid Method

**Fisher (1936) Iris Data**

**Clustering Clusters by Wong’s Hybrid Method**
Example 37.4: 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 enable comparison of the results.

The following statements produce Output 37.4.1:

```sas
title 'Hierarchical Cluster Analysis of Mammals' Teeth Data';
title2 'Evaluating the Effects of Ties';
data teeth;
  input Mammal & $16. v1-v8 @@;
  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 2 3 1 1 3 3 3 3 Mole 3 2 1 0 3 3 3 3
  Silver Hair Bat 2 3 1 1 2 3 3 3 Pigmy Bat 2 3 1 1 2 2 3 3
  House Bat 2 3 1 1 2 2 3 3 Red Bat 1 3 1 1 2 2 3 3
  Pika 2 1 0 0 2 2 3 3 Rabbit 2 1 0 0 3 2 3 3
  Beaver 1 1 0 0 2 1 3 3 Groundhog 1 1 0 0 2 1 3 3
  Gray Squirrel 1 1 0 0 1 1 3 3 House Mouse 1 1 0 0 0 0 3 3
  Porcupine 1 1 0 0 1 1 3 3 Wolf 3 3 1 1 4 4 2 3
  Bear 3 3 1 1 4 4 2 3 Raccoon 3 3 1 1 4 4 3 2
  Marten 3 3 1 1 4 4 1 2 Weasel 3 3 1 1 3 3 1 2
  Wolverine 3 3 1 1 4 4 1 2 Badger 3 3 1 1 3 3 1 2
  River Otter 3 3 1 1 4 3 1 2 Sea Otter 3 2 1 1 3 3 1 2
  Jaguar 3 3 1 1 3 2 1 1 Cougar 3 3 1 1 3 2 1 1
  Fur Seal 3 2 1 1 4 4 1 1 Sea Lion 3 2 1 1 4 4 1 1
  Grey Seal 3 2 1 1 3 3 2 2 Elephant Seal 2 1 1 1 4 4 1 1
  Reindeer 0 4 1 0 3 3 3 3 Elk 0 4 1 0 3 3 3 3
  Deer 0 4 0 0 3 3 3 3 Moose 0 4 0 0 3 3 3 3
;
```

The following statements produce Output 37.4.1:

```sas
proc cluster data=teeth method=average nonorm noeigen;
  var v1-v8;
  id mammal;
run;
```
### Output 37.4.1  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**

Root-Mean-Square Total-Sample Standard Deviation 0.898027

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>RMS Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>31 Beaver</td>
<td>Groundhog</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>30 Gray Squirrel</td>
<td>Porcupine</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>29 Wolf</td>
<td>Bear</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>28 Marten</td>
<td>Wolverine</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>27 Weasel</td>
<td>Badger</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>26 Jaguar</td>
<td>Cougar</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>25 Fur Seal</td>
<td>Sea Lion</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>24 Reindeer</td>
<td>Elk</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>23 Deer</td>
<td>Moose</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>22 Brown Bat</td>
<td>Silver Hair Bat</td>
<td>2</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>21 Pigmy Bat</td>
<td>House Bat</td>
<td>2</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>20 Pika</td>
<td>Rabbit</td>
<td>2</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>19 CL31</td>
<td>CL30</td>
<td>4</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>18 CL28</td>
<td>River Otter</td>
<td>3</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>17 CL27</td>
<td>Sea Otter</td>
<td>3</td>
<td>1</td>
<td>T</td>
</tr>
<tr>
<td>16 CL24</td>
<td>CL23</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15 CL21</td>
<td>Red Bat</td>
<td>3</td>
<td>1.2247</td>
<td></td>
</tr>
<tr>
<td>14 CL17</td>
<td>Grey Seal</td>
<td>4</td>
<td>1.291</td>
<td></td>
</tr>
<tr>
<td>13 CL29</td>
<td>Raccoon</td>
<td>3</td>
<td>1.4142</td>
<td>T</td>
</tr>
<tr>
<td>12 CL25</td>
<td>Elephant Seal</td>
<td>3</td>
<td>1.4142</td>
<td></td>
</tr>
<tr>
<td>11 CL18</td>
<td>CL14</td>
<td>7</td>
<td>1.5546</td>
<td></td>
</tr>
<tr>
<td>10 CL22</td>
<td>CL15</td>
<td>5</td>
<td>1.5811</td>
<td></td>
</tr>
<tr>
<td>9  CL20</td>
<td>CL19</td>
<td>6</td>
<td>1.8708</td>
<td>T</td>
</tr>
<tr>
<td>8  CL11</td>
<td>CL26</td>
<td>9</td>
<td>1.9272</td>
<td></td>
</tr>
<tr>
<td>7  CL8</td>
<td>CL12</td>
<td>12</td>
<td>2.2278</td>
<td></td>
</tr>
<tr>
<td>6  Mole</td>
<td>CL13</td>
<td>4</td>
<td>2.2361</td>
<td></td>
</tr>
<tr>
<td>5  CL9</td>
<td>House Mouse</td>
<td>7</td>
<td>2.4833</td>
<td></td>
</tr>
<tr>
<td>4  CL6</td>
<td>CL7</td>
<td>16</td>
<td>2.5658</td>
<td></td>
</tr>
<tr>
<td>3  CL10</td>
<td>CL16</td>
<td>9</td>
<td>2.8107</td>
<td></td>
</tr>
<tr>
<td>2  CL3</td>
<td>CL5</td>
<td>16</td>
<td>3.7054</td>
<td></td>
</tr>
<tr>
<td>1  CL2</td>
<td>CL4</td>
<td>32</td>
<td>4.2939</td>
<td></td>
</tr>
</tbody>
</table>
The following statements produce Output 37.4.2:

```sas
title3 'Standardized Data';
proc cluster data=teeth std method=average nonorm noeigen;
  var v1-v8;
  id mammal;
run;
```

Output 37.4.2  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

The data have been standardized to mean 0 and variance 1

Root-Mean-Square Total-Sample Standard Deviation 1
### Output 37.4.2 continued

<table>
<thead>
<tr>
<th>Number of Clusters</th>
<th>Clusters Joined</th>
<th>Freq</th>
<th>RMS Distance</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>Beaver Groundhog</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>30</td>
<td>Gray Squirrel Porcupine</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>29</td>
<td>Wolf Bear</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>28</td>
<td>Marten Wolverine</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>27</td>
<td>Weasel Badger</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>26</td>
<td>Jaguar Cougar</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>25</td>
<td>Fur Seal Sea Lion</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>24</td>
<td>Reindeer Elk</td>
<td>2</td>
<td>0</td>
<td>T</td>
</tr>
<tr>
<td>23</td>
<td>Deer Moose</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>Pigmy Bat Red Bat</td>
<td>2</td>
<td>0.9157</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>CL28 River Otter</td>
<td>3</td>
<td>0.9169</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>CL31 CL30</td>
<td>4</td>
<td>0.9428</td>
<td>T</td>
</tr>
<tr>
<td>19</td>
<td>Brown Bat Silver Hair Bat</td>
<td>2</td>
<td>0.9428</td>
<td>T</td>
</tr>
<tr>
<td>18</td>
<td>Pika Rabbit</td>
<td>2</td>
<td>0.9428</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>CL27 Sea Otter</td>
<td>3</td>
<td>0.9847</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>CL22 House Bat</td>
<td>3</td>
<td>1.1437</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>CL21 CL17</td>
<td>6</td>
<td>1.3314</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>CL25 Elephant Seal</td>
<td>3</td>
<td>1.3447</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>CL19 CL16</td>
<td>5</td>
<td>1.4688</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>CL15 Grey Seal</td>
<td>7</td>
<td>1.6314</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>CL29 Raccoon</td>
<td>3</td>
<td>1.692</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CL18 CL20</td>
<td>6</td>
<td>1.7357</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>CL12 CL26</td>
<td>9</td>
<td>2.0285</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CL24 CL23</td>
<td>4</td>
<td>2.1891</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CL9 CL14</td>
<td>12</td>
<td>2.2674</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CL10 House Mouse</td>
<td>7</td>
<td>2.317</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CL11 CL7</td>
<td>15</td>
<td>2.6484</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>CL13 Mole</td>
<td>6</td>
<td>2.8624</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>CL4 CL8</td>
<td>10</td>
<td>3.5194</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CL3 CL6</td>
<td>17</td>
<td>4.1265</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>CL2 CL5</td>
<td>32</td>
<td>4.7753</td>
<td></td>
</tr>
</tbody>
</table>

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.
Example 37.4: Evaluating the Effects of Ties

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, as follows.

```sas
/* --------------------------------------------------------- */
/* */
/* 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_(drop=i);
  set &data;
  array _random_ _ran_1-_ran_&nperm;
  do i = 1 to dim(_random_);
    _random_[i]=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;
```
Chapter 37: The CLUSTER Procedure

/* --------------------------------------------------------- */
/* */
/* The macro PLOTPERM plots various cluster statistics */
/* against the number of clusters for each permutation. */
/* The arguments are as follows: */
/* */
/* nclus maximum number of clusters to be plotted */
/* nperm number of random permutations. */
/* */
/* --------------------------------------------------------- */
%macro PLOTPERM(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_ _psf_ _pst2_ _ccc_ _perm_;
run;

/* ---PLOT THE REQUESTED STATISTICS BY NUMBER OF CLUSTERS--- */
proc sgscatter;
  compare y=(_ccc_ _psf_ _pst2_) x=_ncl_ /group=_perm_;
  label _ccc_ = 'CCC' _psf_ = 'Pseudo F' _pst2_ = 'Pseudo T-Squared';
run;
%mend;

/* --------------------------------------------------------- */
/* */
/* The macro TABPERM generates cluster-membership variables */
/* for a specified number of clusters for each permutation. */
/* 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 TABPERM(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;
/* */
Example 37.4: Evaluating the Effects of Ties

```plaintext
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;

/* ------ TABULATE CLUSTER COMBINATIONS------------ */
proc sort;
  by _clus_;;
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 macros, it is first convenient to define a macro variable, VLIST, listing the teeth variables, since
the forms V1-V8 or V cannot be used with the TABULATE procedure in the TABPERM 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 10 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 four-cluster level being suggested. Hence, the macros are used as follows to analyze the results at
the four-cluster level:

title3 'Raw Data';

/* ------CLUSTER RAW DATA WITH AVERAGE LINKAGE------------------ */
%clusperm( teeth, &vlist, mammal, average, 10);
```
The following statements produce **Output 37.4.3**.

```plaintext
/* -----PLOT STATISTICS FOR THE LAST 20 LEVELS---------------- */
%plotperm(20, 10);
```

**Output 37.4.3** Analysis of 10 Random Permutations of Raw Mammals' Teeth Data

The following statements produce **Output 37.4.4**.

```plaintext
/* ------ANALYZE THE 4-CLUSTER LEVEL------------------------ */
%tabperm( &vlist, mammal, 9.1, 4, 10);
```
Output 37.4.4  Raw Mammals' Teeth Data: Indeterminacy at the Four-Cluster Level

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

<table>
<thead>
<tr>
<th>all_clus</th>
<th>1 3 1 1 1 3 3 2 3</th>
<th>4 0.0 4.0 0.5 0.0 3.0 3.0 3.0 3.0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2 2 2 2 2 2 1 2 1 1</td>
<td>15 2.9 2.6 1.0 1.0 3.6 3.4 1.3 1.8</td>
</tr>
<tr>
<td></td>
<td>2 4 2 2 4 2 1 2 1 1</td>
<td>1 3.0 2.0 1.0 0.0 3.0 3.0 3.0 3.0</td>
</tr>
<tr>
<td></td>
<td>3 1 3 3 3 1 2 1 3 2</td>
<td>5 1.0 1.0 0.0 0.0 1.2 0.8 3.0 3.0</td>
</tr>
<tr>
<td></td>
<td>3 4 3 3 4 1 2 1 3 2</td>
<td>2 2.0 1.0 0.0 0.0 2.5 2.0 3.0 3.0</td>
</tr>
<tr>
<td></td>
<td>4 4 4 4 4 4 4 4 4 4 4</td>
<td>5 1.8 3.0 1.0 1.0 2.0 2.4 3.0 3.0</td>
</tr>
</tbody>
</table>

From the TABULATE 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 as shown in the following statements. The pseudo $F$ and $t^2$ statistics indicate three or four clusters, while the cubic clustering criterion shows a sharp rise up to four clusters and then levels off up to six clusters. So the TABPERM macro is used again at the four-cluster level. In this case, there is no indeterminacy, because 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.

title3 'Standardized Data';

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

The following statements produce Output 37.4.5.

/ * ------PLOT STATISTICS FOR THE LAST 20 LEVELS----------- */
%plotperm(20, 10);
Output 37.4.5 Analysis of 10 Random Permutations of Standardized Mammals' Teeth Data

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

- CCC
- Pseudo F
- Pseudo T-Squared

Number of Clusters

permutation number

1 2 3 4 5 6 7 8 9 10
The following statements produce Output 37.4.6.

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

Output 37.4.6  Standardized Mammals’ Teeth Data: No Indeterminacy at the Four-Cluster Level

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

<table>
<thead>
<tr>
<th></th>
<th>Top incisors</th>
<th>Bottom incisors</th>
<th>Top canines</th>
<th>Bottom canines</th>
<th>Top premolars</th>
<th>Bottom premolars</th>
<th>Top molars</th>
<th>Bottom molars</th>
</tr>
</thead>
</table>
| all_clus
| all_{clus}  |   FREQ |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |       |
| 1 3 1 1 1 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 4 | 6     | 2.0   | 2.8   | 1.0   | 0.8   | 2.2   | 2.5   | 3.0   | 3.0   |       |       |       |       |       |       |       |

References


Subject Index

agglomerative hierarchical clustering analysis, 2426
  average linkage
    CLUSTER procedure, 2435, 2447

bimodality coefficient
  CLUSTER procedure, 2443, 2454

centroid method
  CLUSTER procedure, 2435, 2447

chaining, reducing when clustering, 2444
  CLUSTER procedure
    algorithms, 2455
    average linkage, 2426
    centroid method, 2426
    clustering methods, 2426, 2446
    complete linkage, 2426
    computational resources, 2456
    density linkage, 2426, 2435
    Euclidean distances, 2426
    F statistics, 2443, 2454
    FASTCLUS procedure, compared, 2426
    flexible-beta method, 2426, 2435, 2436, 2451
    hierarchical clusters, 2426
    input data sets, 2436
    interval scale, 2457
    kth-nearest-neighbor method, 2426
    maximum likelihood, 2426, 2435
    McQuitty’s similarity analysis, 2426
    median method, 2426
    memory requirements, 2456
    missing values, 2456
    non-Euclidean distances, 2426
    ODS Graph names, 2464
    output data sets, 2439, 2458
    output table names, 2463
    pseudo F and t statistics, 2443
    ratio scale, 2457
    single linkage, 2426
    size, shape, and correlation, 2457
    test statistics, 2436, 2443, 2444
    ties, 2457
    time requirements, 2456
    two-stage density linkage, 2426
    types of data sets, 2426
    using macros for many analyses, 2485
    Ward’s minimum-variance method, 2426
    Wong’s hybrid method, 2426
    clustering, 2425, see also CLUSTER procedure
      average linkage, 2435, 2447
  centroid method, 2435, 2447
  complete linkage method, 2435, 2448
  density linkage methods, 2435–2438, 2443, 2448, 2450, 2452
  Gower’s method, 2436, 2451
  maximum-likelihood method, 2439, 2450, 2451
  McQuitty’s similarity analysis, 2435, 2451
  median method, 2436, 2451
  methods affected by frequencies, 2445
  outliers in, 2426, 2444
  penalty coefficient, 2439
  single linkage, 2436, 2452
  smoothing parameters, 2449
  standardizing variables, 2443
  transforming variables, 2426
  two-stage density linkage, 2436
  Ward’s method, 2436, 2453
  weighted average linkage, 2435, 2451

complete linkage
  CLUSTER procedure, 2435, 2448

computational resources
  CLUSTER procedure, 2456

dendritic method, see single linkage

density linkage
  CLUSTER procedure, 2435–2438, 2443, 2448, 2450, 2452

diameter method, see complete linkage

DISTANCE data sets
  CLUSTER procedure, 2436

elementary linkage analysis, see single linkage

Euclidean distances, 2437, 2438
  clustering, 2426

F statistics
  CLUSTER procedure, 2443, 2454

flexible-beta method
  CLUSTER procedure, 2426, 2435, 2436, 2451

FREQ statement
  and RMSSTD statement (CLUSTER), 2451

furthest neighbor clustering, see complete linkage

Gower’s method
  CLUSTER procedure, 2436, 2451

group average clustering, see average linkage
hierarchical clustering, 2435, 2450
HYBRID option
and FREQ statement (CLUSTER), 2445
and other options (CLUSTER), 2443
PROC CLUSTER statement, 2449
K= option
and other options (CLUSTER), 2437, 2443
kth-nearest neighbor, see also density linkage, see also single linkage
estimation (CLUSTER), 2437, 2443
kth-nearest-neighbor estimation (CLUSTER), 2448
kurtosis
displayed in CLUSTER procedure, 2443
Lance-Williams flexible-beta method, see flexible-beta method
maximum likelihood
hierarchical clustering (CLUSTER), 2435, 2439, 2450, 2451
maximum method, see complete linkage
McQuitty’s similarity analysis
CLUSTER procedure, 2435
means
displayed in CLUSTER procedure, 2443
median
method (CLUSTER), 2436, 2451
memory requirements
CLUSTER procedure, 2456
missing values
CLUSTER procedure, 2456
modal clusters
density estimation (CLUSTER), 2438
nearest neighbor method, see also single linkage
NOSQUARE option
algorithms used (CLUSTER), 2455
ODS Graph names
CLUSTER procedure, 2464
output data sets
CLUSTER procedure, 2439
output table names
CLUSTER procedure, 2463
preliminary clusters
definition (CLUSTER), 2449
using in CLUSTER procedure, 2437
pseudo F and t statistics
CLUSTER procedure, 2443
R-square statistic
CLUSTER procedure, 2443
R= option
and other options (CLUSTER), 2437, 2443
radius of sphere of support, 2443
rank order typal analysis, see complete linkage
RMSSTD statement
and FREQ statement (CLUSTER), 2445
semipartial correlation
formula (CLUSTER), 2453
single linkage
CLUSTER procedure, 2436, 2452
skewness
displayed in CLUSTER procedure, 2443
smoothing parameter
cluster analysis, 2449
squared semipartial correlation
formula (CLUSTER), 2453
standard deviation
CLUSTER procedure, 2443
standardizing
CLUSTER procedure, 2443
stored data algorithm, 2455
stored distance algorithms, 2455
t-square statistic
CLUSTER procedure, 2443, 2454
ties
checking for in CLUSTER procedure, 2439
time requirements
CLUSTER procedure, 2456
transformations
cluster analysis, 2426
TRIM= option
and other options (CLUSTER), 2437, 2443
two-stage density linkage
CLUSTER procedure, 2436, 2452
ultrametric, definition, 2454
uniform-kernel estimation
CLUSTER procedure, 2443, 2449
unsquared Euclidean distances, 2437, 2438
unweighted pair-group clustering, see average linkage, see centroid method
UPGMA, see average linkage
UPGMC, see centroid method
Ward’s minimum-variance method
CLUSTER procedure, 2436, 2453
weighted average linkage
CLUSTER procedure, 2435, 2451
weighted pair-group methods, see McQuitty’s similarity analysis
weighted-group method, see centroid method
Wong’s hybrid method
CLUSTER procedure, 2437, 2449
WPGMA, see McQuitty’s similarity analysis
Syntax Index

BETA= option
   PROC CLUSTER statement, 2436
BY statement
   CLUSTER procedure, 2444

CCC option
   PROC CLUSTER statement, 2436

CLUSTER procedure
   syntax, 2434
   CLUSTER procedure, BY statement, 2444
   CLUSTER procedure, COPY statement, 2444
   CLUSTER procedure, FREQ statement, 2445
   CLUSTER procedure, ID statement, 2445
   CLUSTER procedure, PROC CLUSTER statement, 2434
   BETA= option, 2436
   CCC option, 2436
   DATA= option, 2436
   DIM= option, 2437
   HYBRID option, 2437
   K= option, 2437
   MODE= option, 2438
   NOEIGEN option, 2438
   NOID option, 2438
   NONORM option, 2438
   NOPRINT option, 2438
   NOSQUARE option, 2437, 2438
   NOTIE option, 2439
   OUTTREE= option, 2439
   PENALTY= option, 2439
   PLOTS option, 2439
   PRINT= option, 2443
   PSEUDO= option, 2443
   R= option, 2443
   RMSSTD option, 2443
   RSQUARE option, 2443
   SIMPLE option, 2443
   STANDARD option, 2443
   TRIM= option, 2443
   CLUSTER procedure, RMSSTD statement, 2445
   CLUSTER procedure, VAR statement, 2446

DATA= option
   PROC CLUSTER statement, 2436
DIM= option
   PROC CLUSTER statement, 2437

HYBRID option
   PROC CLUSTER statement, 2437

K= option
   PROC CLUSTER statement, 2437

METHOD= specification
   PROC CLUSTER statement, 2435

MODE= option
   PROC CLUSTER statement, 2438

NOEIGEN option
   PROC CLUSTER statement, 2438

NOID option
   PROC CLUSTER statement, 2438

NONORM option
   PROC CLUSTER statement, 2438

NOPRINT option
   PROC CLUSTER statement, 2438

NOSQUARE option
   PROC CLUSTER statement, 2437, 2438

NOTIE option
   PROC CLUSTER statement, 2439

OUTTREE= option
   PROC CLUSTER statement, 2439

PENALTY= option
   PROC CLUSTER statement, 2439

PLOTS option
   PROC CLUSTER statement, 2439

PRINT= option
   PROC CLUSTER statement, 2443

PSEUDO= option
   PROC CLUSTER statement, 2443

R= option
   PROC CLUSTER statement, 2443

RMSSTD option
   PROC CLUSTER statement, 2443

RSQUARE option
   PROC CLUSTER statement, 2443

SIMPLE option
   PROC CLUSTER statement, 2443

STANDARD option
   PROC CLUSTER statement, 2443

TRIM= option
   PROC CLUSTER statement, 2437, 2443
   and other options, 2437

PROC CLUSTER statement, see CLUSTER procedure