@c 2014 by G. David Garson and Statistical Associates Publishing. All rights reserved worldwide in all media. No permission is granted to any user to copy or post this work in any format or any media.


The author and publisher of this eBook and accompanying materials make no representation or warranties with respect to the accuracy, applicability, fitness, or completeness of the contents of this eBook or accompanying materials. The author and publisher disclaim any warranties (express or implied), merchantability, or fitness for any particular purpose. The author and publisher shall in no event be held liable to any party for any direct, indirect, punitive, special, incidental or other consequential damages arising directly or indirectly from any use of this material, which is provided “as is”, and without warranties. Further, the author and publisher do not warrant the performance, effectiveness or applicability of any sites listed or linked to in this eBook or accompanying materials. All links are for information purposes only and are not warranted for content, accuracy or any other implied or explicit purpose. This eBook and accompanying materials is © copyrighted by G. David Garson and Statistical Associates Publishing. No part of this may be copied, or changed in any format, sold, or used in any way under any circumstances other than reading by the downloading individual.

Contact:

G. David Garson, President
Statistical Publishing Associates
274 Glenn Drive
Asheboro, NC 27205 USA

Email: gdavidgarson@gmail.com
Web: www.statisticalassociates.com
## Table of Contents

Overview ....................................................................................................................................... 10
Data examples in this volume ....................................................................................................... 10
Key Concepts and Terms ............................................................................................................... 12
  Terminology ............................................................................................................................. 12
    Distances (proximities) ........................................................................................................ 12
    Cluster formation ................................................................................................................. 12
    Cluster validity ..................................................................................................................... 12
  Types of cluster analysis ........................................................................................................... 14
  Types of cluster analysis by software package .................................................................... 14
    Disjoint clustering ................................................................................................................ 15
    Hierarchical clustering ........................................................................................................ 15
    Overlapping clustering ........................................................................................................ 16
    Fuzzy clustering ................................................................................................................... 16
Hierarchical cluster analysis in SPSS ............................................................................................. 16
  SPSS Input for hierarchical clustering ...................................................................................... 16
    Example ............................................................................................................................... 16
    The main “Hierarchical Cluster Analysis” dialog ................................................................. 17
    Statistics button................................................................................................................... 18
    Plots button ......................................................................................................................... 19
    Methods button ................................................................................................................... 20
Hierarchical cluster analysis in SAS ............................................................................................. Error! Bookmark not defined.
  SAS input for hierarchical cluster analysis\........................................................................... Error! Bookmark not defined.
    Example .................................................................................................................................. Error! Bookmark not defined.
    Data setup ............................................................................................................................. Error! Bookmark not defined.
    SAS syntax ........................................................................................................................... Error! Bookmark not defined.
Hierarchical cluster analysis in SAS ............................................................................................. Error! Bookmark not defined.
  SAS output for hierarchical cluster analysis ........................................................................... Error! Bookmark not defined.
    Simple statistics table ......................................................................................................... Error! Bookmark not defined.
    Eigenvalues of the covariance matrix table ......................................................................... Error! Bookmark not defined.
    Root mean square coefficients............................................................................................ Error! Bookmark not defined.
Cluster history table ............................................................ Error! Bookmark not defined.
Dendogram ............................................................................ Error! Bookmark not defined.
Icicle Plots ............................................................................... Error! Bookmark not defined.
Cluster membership table ...................................................... Error! Bookmark not defined.
Saving data to file ................................................................... Error! Bookmark not defined.
Hierarchical cluster analysis in Stata ......................................... Error! Bookmark not defined.
Stata input for hierarchical cluster analysis ................................ Error! Bookmark not defined.
Stata output for hierarchical cluster analysis ............................ Error! Bookmark not defined.
Agglomeration coefficients .................................................. Error! Bookmark not defined.
Dendogram ............................................................................ Error! Bookmark not defined.
Saving cluster membership values ........................................ Error! Bookmark not defined.
Cluster membership table ...................................................... Error! Bookmark not defined.
K-means cluster analysis ........................................................ Error! Bookmark not defined.
Overview .................................................................................. Error! Bookmark not defined.
Example ................................................................................... Error! Bookmark not defined.
K-means cluster analysis in SPSS ................................................ Error! Bookmark not defined.
SPSS input ................................................................................ Error! Bookmark not defined.
   Main K-means dialog .......................................................... Error! Bookmark not defined.
   The Iterate button ............................................................... Error! Bookmark not defined.
   The Save button .................................................................. Error! Bookmark not defined.
   The Options button ............................................................ Error! Bookmark not defined.
SPSS Output for K-Means cluster analysis ................................ Error! Bookmark not defined.
   The Anova table .................................................................. Error! Bookmark not defined.
   Number of cases in each cluster ........................................ Error! Bookmark not defined.
   Getting different clusters .................................................. Error! Bookmark not defined.
   Cluster membership table .................................................. Error! Bookmark not defined.
Overview .................................................................................. Error! Bookmark not defined.
Example ................................................................................... Error! Bookmark not defined.
SAS input for k-means cluster analysis ...................................... Error! Bookmark not defined.
SAS output for k-means cluster analysis ..................................... Error! Bookmark not defined.
   The “Statistics for Variables” table ....................................... Error! Bookmark not defined.
   Criteria for determining k .................................................. Error! Bookmark not defined.
   The “Cluster Summary” table ............................................. Error! Bookmark not defined.
   Cluster membership and distance values ................................ Error! Bookmark not defined.
   Crosstabulation tables ....................................................... Error! Bookmark not defined.
   Cluster separation plots .................................................... Error! Bookmark not defined.
K-Means cluster analysis in Stata ................................................. Error! Bookmark not defined.
Focal cases ................................................................. Error! Bookmark not defined.
Case labels ............................................................... Error! Bookmark not defined.
Partitions and cross-validation ................................ Error! Bookmark not defined.
Example ...................................................................... Error! Bookmark not defined.
SPSS input .................................................................. Error! Bookmark not defined.
  The user interface .................................................. Error! Bookmark not defined.
  The “Variables” tab ................................................. Error! Bookmark not defined.
  The “Neighbors” tab ............................................... Error! Bookmark not defined.
  The “Features” tab .................................................. Error! Bookmark not defined.
  The “Partitions” tab ............................................... Error! Bookmark not defined.
  The “Save” tab ........................................................ Error! Bookmark not defined.
  The “Output” tab .................................................... Error! Bookmark not defined.
  The “Options” tab .................................................. Error! Bookmark not defined.
SPSS output ............................................................. Error! Bookmark not defined.
  Overview ............................................................... Error! Bookmark not defined.
  The “Case Processing Summary “ table .................. Error! Bookmark not defined.
  The “Predictor Space” plot ...................................... Error! Bookmark not defined.
  The “Peers Chart” ................................................... Error! Bookmark not defined.
  The “k Nearest Neighbors and Distances” table ....... Error! Bookmark not defined.
  “k and Predictor Selection” plots ............................. Error! Bookmark not defined.
  “Quadrant Map” maps ............................................. Error! Bookmark not defined.
  The “Error Summary” table ..................................... Error! Bookmark not defined.
SAS PROC ACECLUS: Pre-processing for elliptical clusters .... Error! Bookmark not defined.
  Overview ............................................................... Error! Bookmark not defined.
  Example .................................................................. Error! Bookmark not defined.
SAS input .................................................................. Error! Bookmark not defined.
  Overview ............................................................... Error! Bookmark not defined.
  Set-up ..................................................................... Error! Bookmark not defined.
  Plots of original data .............................................. Error! Bookmark not defined.
  Using PROC ACECLUS to transform the data ......... Error! Bookmark not defined.
  Plot of transformed data ....................................... Error! Bookmark not defined.
  K-means clustering of transformed data .................. Error! Bookmark not defined.
  K-means clustering of original data ......................... Error! Bookmark not defined.
SAS output ............................................................. Error! Bookmark not defined.
  Plot of untransformed data ...................................... Error! Bookmark not defined.
  Data transformation with PROC ACECLUS .............. Error! Bookmark not defined.
  Plot of transformed data ....................................... Error! Bookmark not defined.
K-means (PROC FASTCLUS) results with original vs. transformed data

SAS PROC VARCLUS: Oblique principal components cluster analysis

Overview
The PROC VARCLUS default method
PROC VARCLUS variations
Example
SAS input
SAS output
The dendogram from PROC TREE
The cluster summary table
The R-squared table
The standardized scoring coefficients table
The cluster structure table
The table of inter-cluster correlations
The cluster history summary statistics table
Cluster membership
Cluster scores

SAS PROC MODECLUS: Nonparametric density cluster analysis
Overview
Interpreting p-values
Example
SAS input
PROC MODECLUS specifications
PROC MODECLUS command syntax
SAS output
First pass: Selecting the optimal radius
Second pass: Generating main output
PROC MODECLUS: Nearest neighbor analysis
SAS syntax for nearest neighbor lists/distances
SAS output for nearest neighbor analysis
Kohonen clustering in SAS Enterprise Miner
Overview of Kohonen clustering
Kohonen Clustering in SAS Enterprise Miner: Setup
Kohonen Clustering in SAS Enterprise Miner: Modeling
Overview
The flow chart model
Node overview ................................................................. Error! Bookmark not defined.
The “Input Data” node ....................................................... Error! Bookmark not defined.
The “SOM/Kohonen” node ............................................... Error! Bookmark not defined.
The “Segment Profile” node ............................................. Error! Bookmark not defined.
Kohonen Clustering in SAS Enterprise Miner: Output .......... Error! Bookmark not defined.
Results of the “Data Input” node ..................................... Error! Bookmark not defined.
Results of the “SOM/Kohonen” node .............................. Error! Bookmark not defined.
Results of the “Segment Profile” node ............................. Error! Bookmark not defined.
Other Forms of Cluster Analysis ........................................ Error! Bookmark not defined.
Expectation maximization (EM) clustering ..................... Error! Bookmark not defined.
Cross-classification to determine k ................................. Error! Bookmark not defined.
Distributional characteristics ......................................... Error! Bookmark not defined.
Classification probabilities .......................................... Error! Bookmark not defined.
Q-mode factor analysis .................................................. Error! Bookmark not defined.
Multidimensional scaling .............................................. Error! Bookmark not defined.
Discriminant function analysis ....................................... Error! Bookmark not defined.
F-ratio methods ............................................................. Error! Bookmark not defined.
Assumptions .................................................................. Error! Bookmark not defined.
Randomization ............................................................... Error! Bookmark not defined.
Data level ...................................................................... Error! Bookmark not defined.
Independence of observations ....................................... Error! Bookmark not defined.
Data distribution .......................................................... Error! Bookmark not defined.
Comparable scaling ..................................................... Error! Bookmark not defined.
GLM assumptions ......................................................... Error! Bookmark not defined.
Sample size ................................................................. Error! Bookmark not defined.
Outliers .......................................................................... Error! Bookmark not defined.
Frequently Asked Questions .......................................... Error! Bookmark not defined.
Should data be standardized prior to running cluster analysis? Error! Bookmark not defined.
What are alternative linkage methods? ............................ Error! Bookmark not defined.

SPSS ........................................................................ Error! Bookmark not defined.
SAS .............................................................................. Error! Bookmark not defined.
Stata ............................................................................ Error! Bookmark not defined.
What are alternative distance measures? .......................... Error! Bookmark not defined.

SPSS ........................................................................ Error! Bookmark not defined.
SAS .............................................................................. Error! Bookmark not defined.
Stata ............................................................................ Error! Bookmark not defined.
It is acknowledged that k-means and hierarchical clustering are inefficient and inaccurate for large datasets, but what is the evidence that two-step clustering does better?............Error! Bookmark not defined.

Can I cluster variables instead of cases?................................. Error! Bookmark not defined.
Can I cluster repeated measures data?................................. Error! Bookmark not defined.
Isn't discriminant analysis the same as cluster analysis?......... Error! Bookmark not defined.
What is the ratio of distance measure used in autoclustering in two-step cluster analysis?  
.................................................................................................... Error! Bookmark not defined.
How does SAS’s PROC MODECLUS work?............................... Error! Bookmark not defined.
How does joining and dissolving work in SAS PROC MODECLUS?.......... Error! Bookmark not defined.
What is the rationale for the stability value criterion in SAS PROC MODECLUS? ..........Error! Bookmark not defined.
What does the content of OUTSTAT= files look like for PROC VARCLUS? . Error! Bookmark not defined.
What is BIRCH clustering? ...................................................... Error! Bookmark not defined.
What is ClustanGraphics? ...................................................... Error! Bookmark not defined.
What is SaTScan? ................................................................. Error! Bookmark not defined.
Where can I find cluster software for R? ................................. Error! Bookmark not defined.
How does cluster analysis compare with factor analysis and multidimensional scaling?..Error! Bookmark not defined.

Acknowledgments........................................................................ Error! Bookmark not defined.

Bibliography .................................................................................. Error! Bookmark not defined.
Cluster Analysis

Overview

Cluster analysis, also called segmentation analysis or taxonomy analysis, seeks to identify homogeneous subgroups of cases in a population. That is, cluster analysis is used when the researcher does not know the number of groups in advance but wishes to establish groups and then analyze group membership. Contrast, for instance, discriminant function analysis, which analyzes group membership for known groups pre-specified by the researcher. Cluster analysis implements this by seeking to identify a set of groups which both minimize within-group variation and maximize between-group variation. Later, group membership values may be saved as a case-level variable and used in other procedures such as crosstabulation.

While sometimes described as a method of clustering observations rather than variables, it is always possible to transpose the data matrix so that variables are clustered instead. Some software options allow the researcher to select whether clustering of observations or of variables is desired, without need for data transposition.

Other related techniques, such as factor analysis, multidimensional scaling, and latent class analysis also perform clustering and are discussed in separate volumes of the Statistical Associates "Blue Book" series.

Data examples in this volume

The example datasets used in this volume are listed below in order of use, with versions for SPSS (.sav), SAS (.sas7bdat), and Stata (.dta).

The judges dataset, drawn from SPSS data samples, is a hypothetical data file focusing on the scores given by trained judges plus one "enthusiast" to 300 gymnastic performances. Each row represents a separate performance. All judges viewed and rated the same performances.

- Click here to download judges.sav for SPSS.
- Click here to download judges.sas7bdat for SAS.
- Click here to download judges.dta for Stata.
For SAS PROC CLUSTER, a reformatted dataset labeled "judges_flipped.sav" is used below.

- Click here to download judges_flipped.sas7bdat for SAS.

Two-step clustering in SPSS and PROC MODECLUS in SAS use the “cars” dataset, also drawn from SPSS data samples. This dataset contains variables dealing with engine size, number of cylinders, and other attributes of automobiles from selected countries, for 406 automobile models.

- Click here to download cars.sav for SPSS.
- Click here to download cars.sas7bdat for SAS.
- Click here to download cars.dta for Stata.

Nearest neighbor analysis in SPSS uses the auto.sav data file as an example. It is also used in the section on SOM/Kohonen clustering with SAS Enterprise Miner. This is not the same dataset as cars.sav above Variables are described below.

- Click here to download auto.sav for SPSS
- Click here to download auto.sas7bdat for SAS

The PROC VARCLUS example for SAS, below, uses the subset.sas7bdat file. This is a modified version of the GSS93subset.sav General Social Survey data file supplied in the SPSS Samples directory.

- Click here to download subset.sav for SPSS.
- Click here to download subset.sas7bdat for SAS.
- Click here to download subset.dta for Stata.

The PROC ACECLUS example for SAS, below, uses a version of the “Iris” sample file supplied with SPSS Amos and widely used elsewhere for instruction. Variables are described below.

- Click here to download iris.sas7bdat for SAS
Key Concepts and Terms

Terminology

Distances (proximities)

The first step in cluster analysis is establishment of the similarity or distance matrix. This matrix is a table in which both the rows and columns are the units of analysis and the cell entries are a measure of similarity or distance for any pair of observations (the usual design) or variables (for transposed data). Depending on software, the similarity or distance matrix may be constructed “behind the scenes” from raw data by the statistics package rather than being required as input. Alternative distance measures vary by software package but typical alternatives are discussed below in the FAQ section as well as throughout this volume.

Cluster formation

Cluster formation is the selection of the procedure for determining how clusters are created, and how the calculations are done. In agglomerative hierarchical clustering every case is initially considered a cluster, then the two cases with the lowest distance (or highest similarity) are combined into a cluster. The case with the lowest distance to either of the first two is considered next. If that third case is closer to a fourth case than it is to either of the first two, the third and fourth cases become the second two-case cluster; if not, the third case is added to the first cluster. The process is repeated, adding cases to existing clusters, creating new clusters, or combining clusters to get to the desired final number of clusters. There is also divisive clustering, which works in the opposite direction, starting with all cases in one large cluster. Hierarchical cluster analysis, discussed below, can use either agglomerative or divisive clustering strategies.

Cluster validity

By whatever method the researcher forms clusters, the utility of clusters must be assessed by multiple criteria:

1. Meaningfulness
As in factor analysis, ideally the meaning of each cluster should be readily intuited from the constituent observations or variables used to create the clusters. Variable importance plots, discussed below, are one method of making this assessment.

2. **Separation**
Clusters are more meaningful if they are distinct from each other. Cluster separation plots, discussed below, are one method of assessing separation.

3. **Size**
All clusters should have enough cases to be meaningful. One or more very small clusters indicates that the researcher has requested too many clusters. Analysis resulting in a very large, dominant cluster may indicate too few clusters have been requested.

4. **Criterion validity**
The crosstabulation of the cluster membership (id) numbers by other variables known from theory or prior research to correlate with the concept which clustering is supposed to reflect, should in fact reveal the expected direction and level of association.

5. **Cross-validation and reliability**
Using one set of data to develop the clustering model and then using another set to validate it is recommended. This is done by computing the centroids of the clusters and comparing them for significant differences using one-way Anova or an independent samples t-test. If the validation sample is a randomly held-back portion of the same overall sample from which the development dataset was drawn, this is referred to as “cross-validation.” If the validation dataset is a wholly new sample, this is referred to as “reliability.”

Failure to meet these criteria may indicate the researcher has requested too many or too few clusters, or possibly that an inappropriate distance measure has been selected. It is also possible that the hypothesized basis for clustering does not exist, resulting in arbitrary clusters.
Types of cluster analysis

Types of cluster analysis by software package

SPSS offers three general approaches to cluster analysis:

1. **Hierarchical clustering** allows users to select a definition of distance, then select a linking method for forming clusters, then determine how many clusters best suit the data. Hierarchical clustering generates representation of clusters in icicle plots and dendograms.

2. **K-means clustering** has the researcher specify the number of clusters in advance (though some coefficients from k-means clustering help with selecting the optimal number of clusters: see below), then the algorithm calculates how to assign cases to the K clusters. K-means clustering is much less computer-intensive and is therefore sometimes preferred when datasets are large (ex., > 1,000). K-means clustering generates an ANOVA table showing mean-square error.

3. **Two-step clustering** creates pre-clusters, then it clusters the pre-clusters using hierarchical methods. Two step clustering handles very large datasets, is the method chosen when data are categorical (it supports continuous variables also), and has the largest array of output options, including variable importance plots.

SAS offers four approaches to cluster analysis:

1. PROC CLUSTER implements hierarchical clustering.
2. PROC FASTCLUS implements k-means clustering.
3. PROC VARCLUS implements disjoint clustering as well as hierarchical clustering (for a definitions of disjoint clustering, see below).
4. PROC MODECLUS implements nonparametric density clustering, in which probability values are computed for clusters.
5. In addition, SAS Enterprise Miner offers a “Cluster node” (for k-means clustering) and a “SOM/Kohonen Node (for Kohonen clustering, discussed below).

Stata supports the following cluster analysis commands:
1. **cluster**: hierarchical cluster analysis, using any of several forms of linkage (single, average, complete, weighted-average, median, centroid, or Ward’s linkage)

2. **cluster kmeans**: k-means clustering

3. **cluster kmedians**: similar to k-means cluster analysis, but using medians

### Disjoint clustering

In disjoint clustering, each object is classified in only one cluster. Clusters are not clustered. K-means clustering and two-step cluster analysis, both discussed below, are of this type.

### Hierarchical clustering

In hierarchical clustering, each object is classified in only one bottom-level cluster but clusters may be clustered. A given object may be in multiple clusters, one per level of clustering. As its name clearly implies, hierarchical cluster analysis creates this type of clustering.

Hierarchical clustering is appropriate for smaller samples (typically < 250). When sample size is large, the algorithm may be very slow to reach a solution and when very large may exceed the capacity of some desktop computers. To accomplish hierarchical clustering, the researcher must specify how similarity or distance is defined and how clusters are aggregated (or divided). Hierarchical clustering generates all possible clusters of sizes 1...K. In hierarchical clustering, the clusters are nested rather than being mutually exclusive, as is the usual case. That is, in hierarchical clustering, larger clusters created at later stages may contain smaller clusters created at earlier stages of agglomeration.

The researcher may wish to use the hierarchical cluster procedure on a sample of cases (ex., 200) to inspect results for different numbers of clusters. The optimum number of clusters depends on the research purpose. Identifying "typical" types may call for few clusters and identifying "exceptional" types may call for many clusters, and in either case the resulting clusters must be meaningful. After using hierarchical clustering to determine the desired number of clusters, the researcher may wish then to analyze the entire dataset with k-means clustering, specifying that number of clusters.
Overlapping clustering

In overlapping clustering, objects may be in more than one cluster, even at the same level.

Fuzzy clustering

In fuzzy clustering, objects may be assigned membership in disjoint, hierarchical, or overlapping clusters on a probabilistic basis. Objects have a probability of membership in each cluster. Factor analysis, discussed in a separate Statistical Associates "Blue Book" volume, yields fuzzy clusters. PROC VARCLUS in SAS is a method of converting the fuzzy clusters emerging from factor analysis into non-fuzzy disjoint clusters.

Hierarchical cluster analysis in SPSS

SPSS Input for hierarchical clustering

Example

This example uses the SPSS example file judges.sav (see access above), where columns (variables) are judges from eight countries and rows are 300 fictional cases of gymnasts being rated on a 0-10 scale, illustrated below.
The main “Hierarchical Cluster Analysis” dialog

From the SPSS menus, select Analyze->Classify>Hierarchical Cluster to bring up the “Hierarchical Cluster Analysis” dialog shown below. Initially, all possible variables will be listed in the box on the left. Move variables desired to be used as the basis of clustering to the box on the right. For this example, judges are variables and all country judges have been moved to the right-hand box in the figure below.

For this example, the primary purpose is to cluster judges, to better understand which country judges are similar to which other judges. Therefore we wish to cluster variables (judges are the column variables). The “Variables” radio button is therefore checked.

Alternatively, if the “Cases” radio button is checked, the cases (rows) will be clustered. For the example data, cases are the sports events being rated.
Statistics button

Under the Statistics button, the dialog for which is shown below, the researcher may request the agglomeration schedule and the proximity matrix, described below in the section on output. The researcher may also specify the minimum and maximum number of clusters (3 to 6 is common) for which to seek solutions, or the researcher may ask for a specific number, or none. The agglomeration schedule, the proximity matrix, and other outputs are discussed further below.
Plots button

Under the “Plots” button dialog, the researcher may request dendograms and icicle plots, also described below in the section on output. Also, the orientation (vertical or horizontal) of icicle plots may be specified.
Methods button

A critical specification for cluster analysis is the selection of the similarity or distance measure used as a basis for clustering. In SPSS, various selections are made in the “Methods” dialog

- In the “Cluster Method” pane, the linkage algorithm for clustering is selected. “Between groups” linkage is the most common choice. Also called UPGMA linkage (unweighted pair-group method using averages), this method uses a form of averaged distances for clustering. Alternative linkage methods are discussed in the FAQ section below.

- In the “Measures” pane, similarity/distance measures are selected. There are three measure pull-down menus, for interval, binary, and count data respectively. The most common interval measure is squared Euclidean
distance. For count data, the most common is chi-square distance. For binary data, squared Euclidean distance is perhaps the most common among a large number of alternatives. Alternative similarity/distance measures are discussed in the FAQ section below.

- It is also possible in the “Transform Values” and “Transform Measure” panes to modify the data used for clustering. While it is possible to standardize and transform variables, in the current example that is not needed as all variables are of the same 0 - 10 scale. When variables are measured on unequal scales, standardization is recommended.

**SPSS output for hierarchical cluster analysis**

**Proximity table**

This table shows the distance from each case to each other case. The type of distance was determined by the researchers selection under the “Method” button discussed above. In this case the default, squared Euclidean distance, is used. The table can be very large but for this example, variables were clustered and judges, eight in number, were the variables, resulting in the small table shown below. The distances show how far apart the row judge is from the column judge, with larger numbers representing greater distances. The “Enthusiast” judge can be seen to be further from other judges than any other judge, with few exceptions (one exception is that China is further from France than is the Enthusiast).

<table>
<thead>
<tr>
<th>Case</th>
<th>Italy</th>
<th>South Korea</th>
<th>Romania</th>
<th>France</th>
<th>China</th>
<th>United States</th>
<th>Russia</th>
<th>Enthusiast</th>
</tr>
</thead>
<tbody>
<tr>
<td>Italy</td>
<td>.000</td>
<td>91.180</td>
<td>79.900</td>
<td>106.290</td>
<td>96.870</td>
<td>89.630</td>
<td>86.970</td>
<td>198.580</td>
</tr>
<tr>
<td>South Korea</td>
<td>91.190</td>
<td>.000</td>
<td>241.750</td>
<td>32.020</td>
<td>271.290</td>
<td>40.420</td>
<td>247.340</td>
<td>260.050</td>
</tr>
<tr>
<td>Romania</td>
<td>79.900</td>
<td>241.750</td>
<td>.000</td>
<td>286.530</td>
<td>28.030</td>
<td>234.910</td>
<td>45.190</td>
<td>243.040</td>
</tr>
<tr>
<td>France</td>
<td>106.290</td>
<td>32.020</td>
<td>286.530</td>
<td>.000</td>
<td>287.340</td>
<td>61.600</td>
<td>279.200</td>
<td>246.070</td>
</tr>
<tr>
<td>China</td>
<td>96.870</td>
<td>241.750</td>
<td>28.030</td>
<td>287.340</td>
<td>.000</td>
<td>270.800</td>
<td>64.180</td>
<td>246.050</td>
</tr>
<tr>
<td>United States</td>
<td>89.630</td>
<td>40.420</td>
<td>234.910</td>
<td>270.800</td>
<td>.000</td>
<td>.000</td>
<td>228.540</td>
<td>257.990</td>
</tr>
<tr>
<td>Russia</td>
<td>86.970</td>
<td>247.340</td>
<td>45.190</td>
<td>279.200</td>
<td>64.180</td>
<td>228.540</td>
<td>.000</td>
<td>256.370</td>
</tr>
<tr>
<td>Enthusiast</td>
<td>198.580</td>
<td>260.050</td>
<td>243.040</td>
<td>246.970</td>
<td>245.050</td>
<td>257.990</td>
<td>256.370</td>
<td>.000</td>
</tr>
</tbody>
</table>
Cluster membership table

The cluster membership table shows variables as rows (this example clusters variables, not cases, where variables were country judges) and columns are alternative numbers of clusters in the solution (as specified in the "Range of Solution" option under the Statistics button, here 3 - 6).

Cell entries show the number of the cluster to which the case belongs in the 3-cluster solution through the 6-cluster solution. From this table, the researcher can see which variables (judges in this example) are in which cluster, depending on the number of clusters in the solution. In each of the four solutions, the Enthusiast judge is in a unique cluster not shared by any country judge.

![Cluster Membership Table]

In SPSS, the “Save “ button allows the researcher to save the cluster membership number to file for use as a variable in future analyses only when clustering observations (cases). It does not support saving cluster membership number when clustering variables (here, judges) as in the current example.

Agglomeration Schedule

The agglomeration schedule shows the sequence of clustering as the algorithm unfolds. The agglomeration schedule is a choice under the “Statistics” button of the SPSS hierarchical cluster analysis procedure (see above). In this table, the rows are stages of clustering, numbered from 1 to (n - 1). Given 8 judges, this
example has 7 stages. The \((n - 1)\)th stage (here Stage 7) includes all the cases in one cluster.

There are also two "Cluster Combined" columns, giving the case or cluster numbers for combination at each stage. In agglomerative clustering using a distance measure like Euclidean distance, stage 1 combines the two cases which have lowest proximity (distance) score. The cluster number goes by the lower of the cases or clusters combined, where cases are initially numbered 1 to \(n\).

<table>
<thead>
<tr>
<th>Stage</th>
<th>Cluster Combined</th>
<th>Coefficients</th>
<th>Stage Cluster First Appears</th>
<th>Next Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cluster 1</td>
<td>Cluster 2</td>
<td>Stage Cluster First Appears</td>
<td>Cluster 1</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>5</td>
<td>28.090</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>4</td>
<td>32.020</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>6</td>
<td>51.110</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>7</td>
<td>54.685</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>3</td>
<td>87.913</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>217.950</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>8</td>
<td>242.579</td>
<td>6</td>
</tr>
</tbody>
</table>

The figure above reflects 8 judges rating 300 objects. The agglomeration schedule shows, for instance, that in Stage 1, judges 3 and 5 are combined in a cluster (the cluster is labeled 3). Then judges 2 and 4 become cluster 2. Then judge 6 is added to cluster 2. Then at Stage 4, the new cluster 3 formed at stage 1 is combined with judge 7 to form a larger cluster, also now labeled 3. Then cluster 3 is joined to judge 1 and is labeled cluster 1. Then cluster 2 is joined to cluster 1 and is labeled cluster 1. Finally, judge 8 (the "enthusiast" judge, who is most different from others) is joined to cluster 1, which then is the only remaining cluster.

The proximity/distance/agglomeration coefficient in the "Coefficients" column is an indicator of how far the agglomeration algorithm has to reach to combine an existing cluster with the next closest cluster or variable (judge). For this example the researcher can see that there is a large jump between stages 5 and 6, corresponding to combining cluster 1 (judges 2,5,7, and 1) with cluster 2 (judges 2, 4, and 6) from stage 5. That is, the algorithm has to reach a long distance to move from a 5-cluster solution to a 6-cluster solution. Reaching a long distance means combining relatively unlike objects.
A large agglomeration coefficient will correspond to a long distance in the dendogram discussed below. When there are relatively few cases, icicle plots or dendograms provide the same linkage information in a visual format.

**Dendogram**

Also called hierarchical tree diagrams or plots, dendograms are one of two types of linkage plots output by SPSS (the other is icicle plots). Dendograms show the relative size of the proximity coefficients at which cases were combined. The bigger the distance coefficient or the smaller the similarity coefficient, the more clustering involved combining unlike entities, which may be undesirable. Trees are usually depicted horizontally, not vertically, with each row representing a case on the Y axis, while the X axis is a rescaled version of the proximity coefficients.
When the number of variables (when clustering variables) or the number of cases (when clustering observations) is large, dendograms can become hard to read.

The figure above shows 8 judges who rated 300 objects. The inset showing the labels for judges 1 – 8 is not part of dendogram output but was lifted from the main hierarchical cluster analysis dialog, where the researcher entered the variables (judges). The dendogram shows judges 3 & 5 (Romania and China) to be in one of the two earliest clusters, with judge 7 (Russia) affiliated with cluster 3 & 5, only at a greater distance.

In general, the dendogram shows the pattern of clustering among the judges, with connecting lines further to the right indicating more distance between judges and clusters. The final linkage to judge 8 ("Enthusiast") shows this judge to be least like the others, but the largest jump occurs a step earlier. If the researcher decided that making that large jump combined objects which were too dissimilar, there would be a three-cluster solution:

1. Judges 3, 5, 7, 1
2. Judges 2, 4, 5
3. Judge 8

In a dendogram, variables or cases with low distance/high similarity are close together. Those showing low distance are close, with a line linking them a short distance from the left of the dendogram, indicating that they are agglomerated into a cluster at a low distance coefficient, indicating similarity. When, on the other hand, the linking line is to the right of the dendogram, the linkage occurs a high distance coefficient, indicating the cases/clusters were agglomerated even though much less alike. If a similarity measure is used rather than a distance measure, the rescaling of the X axis still produces a diagram with linkages involving high alikeness to the left and low alikeness to the right.

The researcher may also cluster cases by so selecting in the main “Hierarchical Cluster Analysis” dialog shown above. The dendogram below is for the clustering of 50 performances (objects) by the 8 judges, with performances 10, 38, 17, 16, 18, 43, 2, 46, and 27 forming one of the first clusters:
CLUSTER ANALYSIS Overview

An illustrated tutorial and introduction to cluster analysis using SPSS, SAS, SAS Enterprise Miner, and Stata for examples. Suitable for introductory graduate-level study.

The 2014 edition is a major update to the 2012 edition. Among the new features are these:

- Was 89 pages, now book length (207 pages total)
- Had 58 figures, now has over 170 illustrations
- Now covers Stata as well as SPSS and SAS
- Totally revised sections on hierarchical, k-means, and two-step clustering
- New coverage of nearest neighbor analysis
- New coverage of oblique principal components cluster analysis
- New coverage of nonparametric density cluster analysis
- New coverage of Kohonen self-organizing map (SOM) clustering
- Links to all datasets used in the text.

The full content is now available from Statistical Associates Publishers. Click here.

Below is the unformatted table of contents.

Table of Contents
CLUSTER ANALYSIS 1
Overview 10
Data examples in this volume 10
Key Concepts and Terms 12
Terminology 12
Distances (proximities) 12
Cluster formation 12
Cluster validity 12
<table>
<thead>
<tr>
<th>Types of cluster analysis</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Types of cluster analysis by software package</td>
<td>14</td>
</tr>
<tr>
<td>Disjoint clustering</td>
<td>15</td>
</tr>
<tr>
<td>Hierarchical clustering</td>
<td>15</td>
</tr>
<tr>
<td>Overlapping clustering</td>
<td>16</td>
</tr>
<tr>
<td>Fuzzy clustering</td>
<td>16</td>
</tr>
<tr>
<td>Hierarchical cluster analysis in SPSS</td>
<td>16</td>
</tr>
<tr>
<td>SPSS Input for hierarchical clustering</td>
<td>16</td>
</tr>
<tr>
<td>Example</td>
<td>16</td>
</tr>
<tr>
<td>The main &quot;Hierarchical Cluster Analysis&quot; dialog</td>
<td>17</td>
</tr>
<tr>
<td>Statistics button</td>
<td>18</td>
</tr>
<tr>
<td>Plots button</td>
<td>19</td>
</tr>
<tr>
<td>Methods button</td>
<td>20</td>
</tr>
<tr>
<td>SPSS output for hierarchical cluster analysis</td>
<td>21</td>
</tr>
<tr>
<td>Proximity table</td>
<td>21</td>
</tr>
<tr>
<td>Cluster membership table</td>
<td>22</td>
</tr>
<tr>
<td>Agglomerate Schedule</td>
<td>22</td>
</tr>
<tr>
<td>Dendogram</td>
<td>24</td>
</tr>
<tr>
<td>Icicle plots</td>
<td>27</td>
</tr>
<tr>
<td>Summary measures</td>
<td>28</td>
</tr>
<tr>
<td>Hierarchical cluster analysis in SAS</td>
<td>29</td>
</tr>
<tr>
<td>SAS input for hierarchical cluster analysis</td>
<td>29</td>
</tr>
<tr>
<td>Example</td>
<td>29</td>
</tr>
<tr>
<td>Data setup</td>
<td>29</td>
</tr>
<tr>
<td>SAS syntax</td>
<td>30</td>
</tr>
<tr>
<td>SAS output for hierarchical cluster analysis</td>
<td>31</td>
</tr>
<tr>
<td>Simple statistics table</td>
<td>31</td>
</tr>
<tr>
<td>Eigenvalues of the covariance matrix table</td>
<td>31</td>
</tr>
<tr>
<td>Root mean square coefficients</td>
<td>32</td>
</tr>
<tr>
<td>Cluster history table</td>
<td>33</td>
</tr>
<tr>
<td>Dendogram</td>
<td>34</td>
</tr>
<tr>
<td>Icicle Plots</td>
<td>36</td>
</tr>
<tr>
<td>Cluster membership table</td>
<td>36</td>
</tr>
<tr>
<td>Saving data to file</td>
<td>37</td>
</tr>
<tr>
<td>Hierarchical cluster analysis in Stata</td>
<td>38</td>
</tr>
<tr>
<td>Stata input for hierarchical cluster analysis</td>
<td>38</td>
</tr>
<tr>
<td>Stata output for hierarchical cluster analysis</td>
<td>40</td>
</tr>
<tr>
<td>Agglomeration coefficients</td>
<td>40</td>
</tr>
<tr>
<td>Dendogram</td>
<td>41</td>
</tr>
<tr>
<td>Saving cluster membership values</td>
<td>42</td>
</tr>
<tr>
<td>Cluster membership table</td>
<td>43</td>
</tr>
<tr>
<td>K-means cluster analysis</td>
<td>44</td>
</tr>
<tr>
<td>Overview</td>
<td>44</td>
</tr>
<tr>
<td>Example</td>
<td>45</td>
</tr>
<tr>
<td>K-means cluster analysis in SPSS</td>
<td>45</td>
</tr>
<tr>
<td>SPSS input</td>
<td>45</td>
</tr>
<tr>
<td>Main K-means dialog</td>
<td>45</td>
</tr>
<tr>
<td>The Iterate button</td>
<td>47</td>
</tr>
<tr>
<td>The Save button</td>
<td>48</td>
</tr>
<tr>
<td>The Options button</td>
<td>49</td>
</tr>
<tr>
<td>SPSS Output for K-Means cluster analysis</td>
<td>50</td>
</tr>
<tr>
<td>The Anova table</td>
<td>50</td>
</tr>
<tr>
<td>Number of cases in each cluster</td>
<td>51</td>
</tr>
<tr>
<td>Getting different clusters</td>
<td>52</td>
</tr>
<tr>
<td>Cluster membership table</td>
<td>52</td>
</tr>
<tr>
<td>K-Means cluster analysis in SAS</td>
<td>53</td>
</tr>
</tbody>
</table>
The "Neighbors" tab 91
The "Features" tab 92
The "Partitions" tab 93
The "Save" tab 95
The "Output" tab 96
The "Options" tab 97
SPSS output 97
Overview 97
The "Case Processing Summary " table 98
The "Predictor Space" plot 98
The "Peers Chart" 101
The "k Nearest Neighbors and Distances" table 102
"k and Predictor Selection" plots 103
"Quadrant Map" maps 104
The "Error Summary" table 105
SAS PROC ACECLUS: Pre-processing for elliptical clusters 106
Overview 106
Example 106
SAS input 107
Overview 107
Set-up 107
Plot of original data 108
Using PROC ACECLUS to transform the data 108
Plot of transformed data 109
K-means clustering of transformed data 109
K-means clustering of original data 110
SAS output 110
Plot of untransformed data 110
Data transformation with PROC ACECLUS 111
Plot of transformed data 112
K-means (PROC FASTCLUS) results with original vs. transformed data 113
SAS PROC VARCLUS : Oblique principal components cluster analysis 115
Overview 115
The PROC VARCLUS default method 115
PROC VARCLUS variations 115
Example 116
SAS input 116
SAS output 119
The dendogram from PROC TREE 119
The cluster summary table 119
The R-squared table 121
The standardized scoring coefficients table 122
The cluster structure table 123
The table of inter-cluster correlations 124
The cluster history summary statistics table 125
Cluster membership 126
Cluster scores 127
SAS PROC MODECLUS: Nonparametric density cluster analysis 127
Overview 127
Interpreting p-values 129
Example 129
SAS input 130
PROC MODECLUS specifications 130
PROC MODECLUS command syntax 131
SAS output 133
First pass: Selecting the optimal radius 133
Second pass: Generating main output 136
PROC MODECLUS: Nearest neighbor analysis 141
SAS syntax for nearest neighbor lists/distances 141
SAS output for nearest neighbor analysis 142
Kohonen clustering in SAS Enterprise Miner 144
Overview of Kohonen clustering 144
Kohonen Clustering in SAS Enterprise Miner: Setup 144
Kohonen Clustering in SAS Enterprise Miner: Modeling 153
Overview 153
The flow chart model 154
Node overview 156
The "Input Data" node 156
The "SOM/Kohonen" node 157
The "Segment Profile" node 159
Kohonen Clustering in SAS Enterprise Miner: Output 160
Results of the "Data Input" node 160
Results of the "SOM/Kohonen" node 161
Results of the "Segment Profile" node 165
Other Forms of Cluster Analysis 173
Expectation maximization (EM) clustering 173
Cross-classification to determine k 173
Distributional characteristics 173
Classification probabilities 174
Q-mode factor analysis 174
Multidimensional scaling 175
Discriminant function analysis 175
F-ratio methods 176
Assumptions 176
Randomization 176
Data level 176
Independence of observations 176
Data distribution 177
Comparable scaling 177
GLM assumptions 178
Sample size 178
Outliers 178
Frequently Asked Questions 178
Should data be standardized prior to running cluster analysis? 178
What are alternative linkage methods? 180
SPSS 180
SAS 181
Stata 182
What are alternative distance measures? 183
SPSS 183
SAS 191
Stata 193
It is acknowledged that k-means and hierarchical clustering are inefficient
and inaccurate for large datasets, but what is the evidence that two-step
clustering does better? 194
Can I cluster variables instead of cases? 194
Can I cluster repeated measures data? 194
Isn't discriminant analysis the same as cluster analysis? 195
What is the ratio of distance measure used in autoclustering in two-step
cluster analysis? 195
How does SAS's PROC MODECLUS work? 196
How does joining and dissolving work in SAS PROC MODECLUS? 196
NEW! For use by a single individual, our entire current library is available at Amazon in no-password pdf format on DVD for $120 plus shipping. Click on [http://www.amazon.com/dp/1626380201](http://www.amazon.com/dp/1626380201). Includes one year of free updates when email address is provided.

NEW! For use by a single individual, our "Regression Models" library of 10 titles is available at Amazon in no-password pdf format on DVD for $50 plus shipping. Click on [http://www.amazon.com/dp/1626380252](http://www.amazon.com/dp/1626380252).

NEW! For use by a single individual, our "Qualitative Methods" library of 10 titles is available at Amazon in no-password pdf format on DVD for $50 plus shipping. Click on [http://www.amazon.com/dp/B00JJ2JZYM](http://www.amazon.com/dp/B00JJ2JZYM).

NEW FOR CLASS USE! If you are requesting this for class use, consider recommending site licensing so the ebook is free for everyone at your institution and is always available. For class use, see our new low-cost site license policy for university libraries and others at [http://statisticalassociates.com/FAQ.htm#sales](http://statisticalassociates.com/FAQ.htm#sales). Site license for a university is $100 per title.

*Association, Measures of*  
*Canonical Correlation*  
*Case Studies*  
*Cluster Analysis*  
*Content Analysis*  
*Correlation*  
*Correlation, Partial*  
*Correspondence Analysis*  
*Cox Regression*  
*Creating Simulated Datasets*  
*Crosstabulation*  
*Curve Estimation & Nonlinear Regression*  
*Delphi Method in Quantitative Research*  
*Discriminant Function Analysis*  
*Ethnographic Research*  
*Evaluation Research*  
*Factor Analysis*  
*Focus Group Research*  
*Game Theory*  
*Generalized Linear Models/Generalized Estimating Equations*  
*GLM (Multivariate), MANOVA, and MANCOVA*  
*GLM (Univariate), ANOVA, and ANCOVA*
Grounded Theory
Life Tables & Kaplan-Meier Survival Analysis
Literature Review in Research and Dissertation Writing
Logistic Regression: Binary & Multinomial
Log-linear Models,
Longitudinal Analysis
Missing Values & Data Imputation
Multidimensional Scaling
Multiple Regression
Narrative Analysis
Network Analysis
Neural Network Models
Nonlinear Regression
Ordinal Regression
Parametric Survival Analysis
Partial Correlation
Partial Least Squares Regression
Participant Observation
Path Analysis
Power Analysis
Probability
Probit and Logit Response Models
Research Design
Scales and Measures
Significance Testing
Social Science Theory in Research and Dissertation Writing
Structural Equation Modeling
Survey Research & Sampling
Testing Statistical Assumptions
Two-Stage Least Squares Regression
Validity & Reliability
Variance Components Analysis
Weighted Least Squares Regression

Statistical Associates Publishing
http://www.statisticalassociates.com
sa.publishers@gmail.com