Title

stata.com

cluster dendrogram — Dendrograms for hierarchical cluster analysis

Description

`cluster dendrogram` produces dendrograms (also called cluster trees) for a hierarchical clustering. See [MV] **cluster** for a discussion of cluster analysis, hierarchical clustering, and the available `cluster` commands.

Dendrograms graphically present the information concerning which observations are grouped together at various levels of (dis)similarity. At the bottom of the dendrogram, each observation is considered its own cluster. Vertical lines extend up for each observation, and at various (dis)similarity values, these lines are connected to the lines from other observations with a horizontal line. The observations continue to combine until, at the top of the dendrogram, all observations are grouped together.

The height of the vertical lines and the range of the (dis)similarity axis give visual clues about the strength of the clustering. Long vertical lines indicate more distinct separation between the groups. Long vertical lines at the top of the dendrogram indicate that the groups represented by those lines are well separated from one another. Shorter lines indicate groups that are not as distinct.

Quick start

Dendrogram of most recent cluster analysis

```
cluster dendrogram
```

Same as above

```
cluster tree
```

As above, but orient horizontally instead of vertically

```
cluster tree, horizontal
```

Dendrogram of cluster analysis named `myclus`

```
cluster tree myclus
```

As above, and apply leaf labels from variable `mylabels` instead of observation numbers

```
cluster tree myclus, labels(mylabels)
```

As above, but rotate leaf labels 90 degrees and reduce text size by half

```
cluster tree myclus, labels(mylabels)  ///
     xlabel(, angle(90) labsize(*.5))
```

Show top 20 branches and associated frequencies from most recent cluster analysis

```
cluster tree, cutnumber(20) showcount
```

Menu

Statistics > Multivariate analysis > Cluster analysis > Postclustering > Dendrograms

Syntax

```
cluster dendrogram [clname] [if] [in] [, options]
```

option	Description
Main	
`quick`	do not center parent branches
`labels(`*varname*`)`	name of variable containing leaf labels
`cutnumber(#)`	display top # branches only
`cutvalue(#)`	display branches above # (dis)similarity measure only
`showcount`	display number of observations for each branch
`countprefix(`*string*`)`	prefix the branch count with *string*; default is "`n=`"
`countsuffix(`*string*`)`	suffix the branch count with *string*; default is empty string
`countinline`	put branch count in line with branch label
`vertical`	orient dendrogram vertically (default)
`horizontal`	orient dendrogram horizontally
Plot	
line_options	affect rendition of the plotted lines
Add plots	
`addplot(`*plot*`)`	add other plots to the dendrogram
Y axis, X axis, Titles, Legend, Overall	
twoway_options	any options other than `by()` documented in [G-3] ***twoway_options***

Note: `cluster tree` is a synonym for `cluster dendrogram`.

In addition to the restrictions imposed by `if` and `in`, the observations are automatically restricted to those that were used in the cluster analysis.

Options

Main

`quick` switches to a different style of dendrogram in which the vertical lines go straight up from the observations instead of the default action of being recentered after each merge of observations in the dendrogram hierarchy. Some people prefer this representation, and it is quicker to render.

`labels(`*varname*`)` specifies that *varname* be used in place of observation numbers for labeling the observations at the bottom of the dendrogram.

`cutnumber(#)` displays only the top # branches of the dendrogram. With large dendrograms, the lower levels of the tree can become too crowded. With `cutnumber()`, you can limit your view to the upper portion of the dendrogram. Also see the `cutvalue()` option.

`cutvalue(#)` displays only those branches of the dendrogram that are above the # (dis)similarity measure. With large dendrograms, the lower levels of the tree can become too crowded. With `cutvalue()`, you can limit your view to the upper portion of the dendrogram. Also see the `cutnumber()` option.

`showcount` requests that the number of observations associated with each branch be displayed below the branches. `showcount` is most useful with `cutnumber()` and `cutvalue()` because, otherwise, the number of observations for each branch is one. When this option is specified, a label for each branch is constructed by using a prefix string, the branch count, and a suffix string.

`countprefix(`*string*`)` specifies the prefix string for the branch count label. The default is `countprefix(n=)`. This option implies the use of the `showcount` option.

`countsuffix(`*string*`)` specifies the suffix string for the branch count label. The default is an empty string. This option implies the use of the `showcount` option.

`countinline` requests that the branch count be put in line with the corresponding branch label. The branch count is placed below the branch label by default. This option implies the use of the `showcount` option.

`vertical` and `horizontal` specify whether the x and y coordinates are to be swapped before plotting—`vertical` (the default) does not swap the coordinates, whereas `horizontal` does.

Plot

line_options affect the rendition of the lines; see [G-3] ***line_options***.

Add plots

`addplot(`*plot*`)` allows adding more `graph twoway` plots to the graph; see [G-3] ***addplot_option***.

Y axis, X axis, Titles, Legend, Overall

twoway_options are any of the options documented in [G-3] ***twoway_options***, excluding `by()`. These include options for titling the graph (see [G-3] ***title_options***) and for saving the graph to disk (see [G-3] ***saving_option***).

Remarks and examples

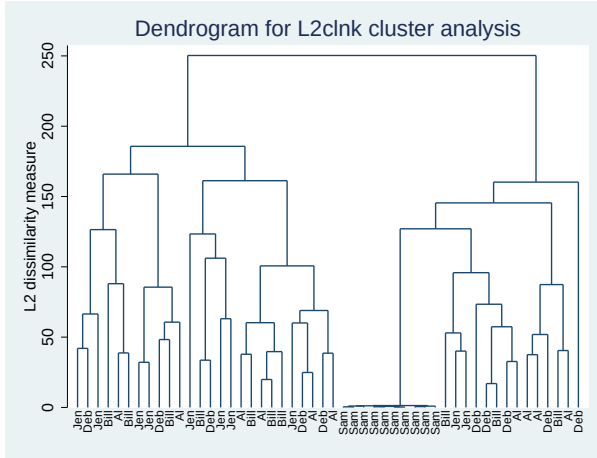
stata.com

Examples of the `cluster dendrogram` command can be found in [MV] **cluster linkage**, [MV] **clustermat**, [MV] **cluster stop**, and [MV] **cluster generate**. Here we illustrate some of the additional options available with `cluster dendrogram`.

Example 1

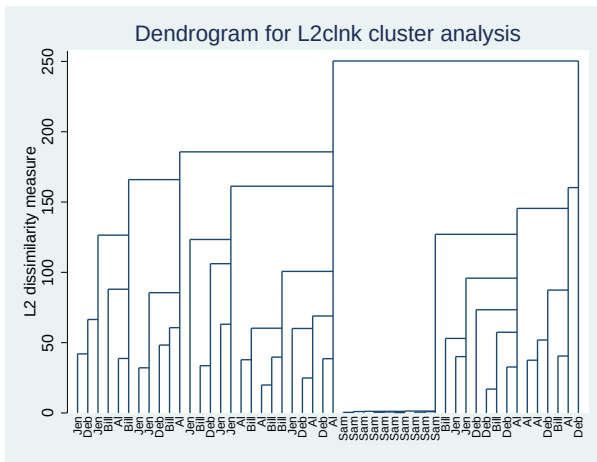
Example 1 of [MV] **cluster linkage** introduces a dataset with 50 observations on four variables. Here we show the dendrogram for a complete-linkage analysis:

```
. use https://www.stata-press.com/data/r17/labtech
. cluster completelinkage x1 x2 x3 x4, name(L2clnk)
. cluster dendrogram L2clnk, labels(labtech) xlabel(, angle(90) labsize(*.75))
```



The same dendrogram can be rendered in a slightly different format by using the `quick` option:

```
. cluster dendrogram L2clnk, quick labels(labtech)
          xlabel(, angle(90) labsize(*.75))
```

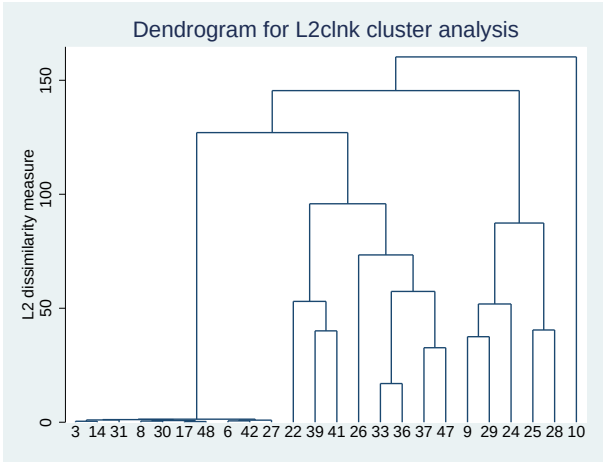


Some people prefer this style of dendrogram. As the name implies, this style of dendrogram is quicker to render.

You can use the `if` and `in` conditions to restrict the dendrogram to the observations for one subgroup. This task is usually accomplished with the `cluster generate` command, which creates a grouping variable; see [MV] **cluster generate**.

Here we show the third of three groups in the dendrogram by first generating the grouping variable for three groups and then using `if` in the command for `cluster dendrogram` to restrict it to the third of those three groups.

```
. cluster generate g3 = group(3)
. cluster tree if g3==3
```

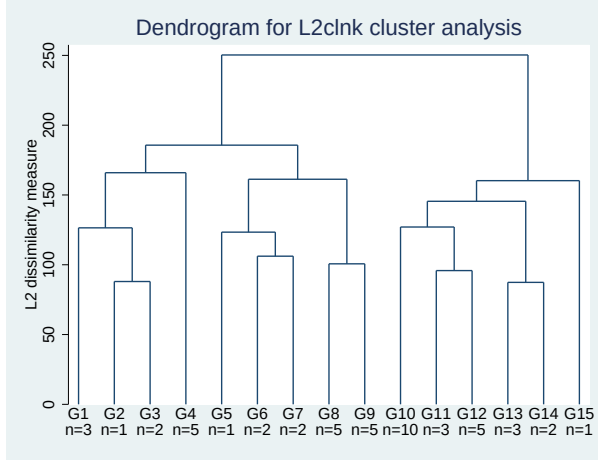


Because we find it easier to type, we used the synonym `tree` instead of `dendrogram`. We did not specify the cluster name, allowing it to default to the most recently performed cluster analysis. We also omitted the `labels()` and `xlabel()` options, which brings us back to the default action of showing, horizontally, the observation numbers.

This example has only 50 observations. When there are many observations, the dendrogram can become too crowded. You will need to limit which part of the dendrogram you display. One way to view only part of the dendrogram is to use `if` and `in` to limit to one particular group, as we did above.

The other way to limit your view of the dendrogram is to specify that you wish to view only the top portion of the tree. The `cutnumber()` and `cutvalue()` options allow you to do this:

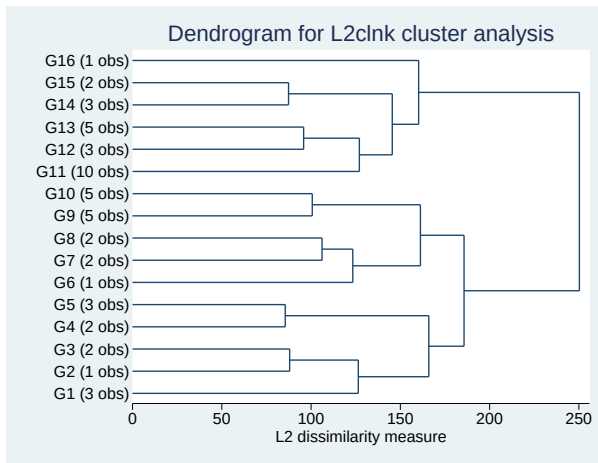
```
. cluster tree, cutn(15) showcount
```



We limited our view to the top 15 branches of the dendrogram with `cutn(15)`. By default, the 15 branches were labeled `G1`–`G15`. The `showcount` option provided, below these branch labels, the number of observations in each of the 15 groups.

The `cutvalue()` option provides another way to limit the view to the top branches of the dendrogram. With this option, you specify the similarity or dissimilarity value at which to trim the tree.

```
. cluster tree, cutvalue(75.3)
          countprefix("(") countsuffix(" obs)") countinline
          ylabel(, angle(0)) horizontal
```



This time, we limited the dendrogram to those branches with dissimilarity greater than 75.3 by using the `cutvalue(75.3)` option. There were 16 branches (groups) that met that restriction. We used the `countprefix()` and `countsuffix()` options to display the number of observations in each branch as "(# obs)" instead of "n=#". The `countinline` option puts the branch counts in line with the branch labels. We specified the `horizontal` option and the `angle(0)` suboption of `ylabel()` to get a horizontal dendrogram with horizontal branch labels.

◁

❑ Technical note

Programmers can control the graphical procedure executed when `cluster dendrogram` is called. This ability will be helpful to programmers adding new hierarchical clustering methods that require a different dendrogram algorithm. See [MV] **cluster programming subroutines** for details.

❑

In systematic zoology, Mayr, Linsley, and Usinger (1953, 312) introduced the term "dendrogram" for "A diagrammatic drawing in the form of a tree designed to indicate degrees of relationship as suggested by degrees of similarity." The first root, "dendron", means "tree": other linked words include "dendrite", "dendritic", and "rhododendron".

But thereby hangs a tale, or two, or three.

The term "dendrogram" was in due course copied from biological systematics and taxonomy into general scientific and statistical literature (for example, Sneath and Sokal [1962]; Hodson, Sneath, and Doran [1966]; Joyce and Channon [1966]). On the way, its meaning became more general, describing tree displays showing the structure of similarity and dissimilarity in nested classifications. The term became widely used in publications on what is now most often called cluster analysis: examples are the books of Sokal and Sneath (1963), Jardine and Sibson (1971), Sneath and Sokal (1973), Everitt (1993), Hartigan (1975), and Gordon (1981), and many others since.

Meanwhile, back in biology, Mayr emerged early as a leading critic of what some biologists, led by Sokal and Sneath, were calling "numerical taxonomy". His objections were evident in a polemic paper (Mayr 1965) and in his lengthy but lively and lucid history of much of biological thought (Mayr 1982). So there is some irony in his term being associated with projects he would not have approved (at least in biological systematics). Those imagining that classification is dull and dreary descriptive work will find ample documentation of scientists red in tooth and claw in Hull (1988), which despite its grand titles is focused on a detailed story of taxonomists' arguments with each other.

Inside biological taxonomy, the distinction is often now between "phenograms", meant to classify resemblance only, and "cladograms", meant to show also evolutionary pedigree.

Naturally, tree diagrams did not spring into existence with the term "dendrogram". Pietsch (2012) and Archibald (2014) give many well-reproduced diagrams from over several centuries showing supposed relationships between different organisms in biology. Lima (2011, 2014) sampled tree imagery even more broadly from many fields and from ancient and modern history to the present.

Ernst Walter Mayr (1904–2005) was a leading evolutionary biologist whose work ranged across systematics, taxonomy, exploration, ornithology, and history and philosophy of biology, especially on and around the concept of species. Mayr was born in Kempten in Germany. He completed his high school education in Dresden and went to university in Greifswald and Berlin. After fieldwork in New Guinea and the Solomon Islands, Mayr joined the American Museum of Natural History in 1931 and Harvard University in 1953, where he remained for the rest of his career. His many honors included membership of the National Academy of Sciences, foreign membership of the Royal Society, Balzan and Crafoord Prizes, and the U.S. National Medal of Science.

Mayr's coauthors, Earle Gorton Linsley (1910–2000) and Robert Leslie Usinger (1912–1968), were distinguished systematic entomologists based at the University of California at Berkeley.

References

Archibald, J. D. 2014. *Aristotle's Ladder, Darwin's Tree: The Evolution of Visual Metaphors for Biological Order*. New York: Columbia University Press.

Everitt, B. S. 1993. *Cluster Analysis*. 3rd ed. London: Arnold.

Falcaro, M., and A. Pickles. 2010. riskplot: A graphical aid to investigate the effect of multiple categorical risk factors. *Stata Journal* 10: 61–68.

Gordon, A. D. 1981. *Classification: Methods for the Exploratory Analysis of Multivariate Data*. London: Chapman & Hall/CRC.

Hartigan, J. A. 1975. Printer graphics for clustering. *Journal of Statistical Computation and Simulation* 4: 187–213. https://doi.org/10.1080/00949657508810123.

Hodson, F. R., P. H. A. Sneath, and J. E. Doran. 1966. Some experiments in the numerical analysis of archaeological data. *Biometrika* 53: 311–324. https://doi.org/10.1093/biomet/53.3-4.311.

Hull, D. L. 1988. *Science as a Process: An Evolutionary Account of the Social and Conceptual Development of Science*. Chicago: University of Chicago Press.

Jardine, N., and R. Sibson. 1971. *Mathematical Taxonomy*. New York: Wiley.

Joyce, T., and C. Channon. 1966. Classifying market survey respondents. *Journal of the Royal Statistical Society, Series C* 15: 191–215. https://doi.org/10.2307/2985300.

Lima, M. 2011. *Visual Complexity: Mapping Patterns of Information*. New York: Princeton Architectural Press.

——. 2014. *The Book of Trees: Visualizing Branches of Knowledge*. New York: Princeton Architectural Press.

Mayr, E. 1965. Numerical phenetics and taxonomic theory. *Systematic Biology* 14: 73–97. https://doi.org/10.2307/2411730.

——. 1982. *The Growth of Biological Thought: Diversity, Evolution, and Inheritance*. Cambridge, MA: Harvard University Press.

Mayr, E., E. G. Linsley, and R. L. Usinger. 1953. *Methods and Principles of Systematic Zoology*. New York: McGraw–Hill.

Pietsch, T. W. 2012. *Trees of Life: A Visual History of Evolution*. Baltimore, MD: Johns Hopkins University Press.

Sneath, P. H. A., and R. R. Sokal. 1962. Numerical taxonomy. *Nature* 193: 855–860. https://doi.org/10.1038/193855a0.

——. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. San Francisco: Freeman.

Sokal, R. R., and P. H. A. Sneath. 1963. *Principles of Numerical Taxonomy*. San Francisco: Freeman.

Also see

[MV] **cluster** — Introduction to cluster-analysis commands

[MV] **clustermat** — Introduction to clustermat commands