

Running an Agglomerative Hierarchical Clustering (AHC) with XLSTAT

[demoCluster.xls](#)

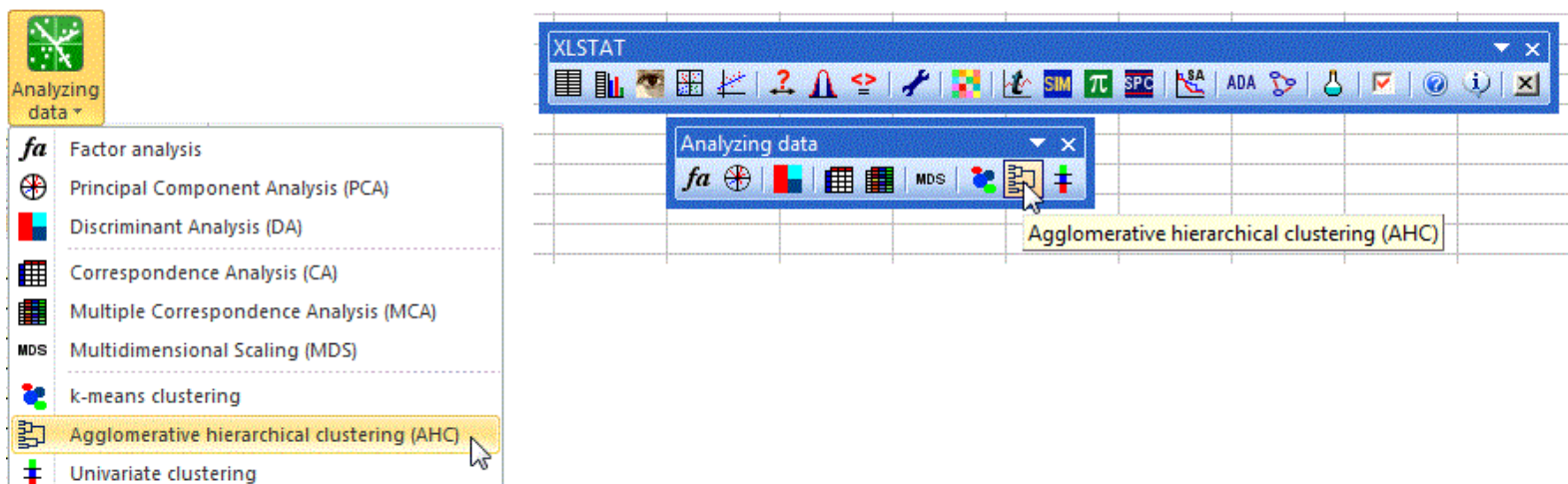
Dataset to run an Agglomerative Hierarchical Clustering in XLSTAT

An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking [here](#).

The data are from the US Census Bureau and describe the changes in the population of 51 states between 2000 and 2001. The initial dataset has been transformed to rates per 1000 inhabitants, with the data for 2001 serving as the focus for the analysis. Our aim is to create homogeneous clusters of states based on the demographic data we have available.

Setting up an Agglomerative Hierarchical Clustering

Once XLSTAT-Pro is activated, select the **XLSTAT / Analyzing data / Agglomerative Hierarchical Clustering** command, or click on the corresponding button of the **Analyzing data** toolbar (see below).

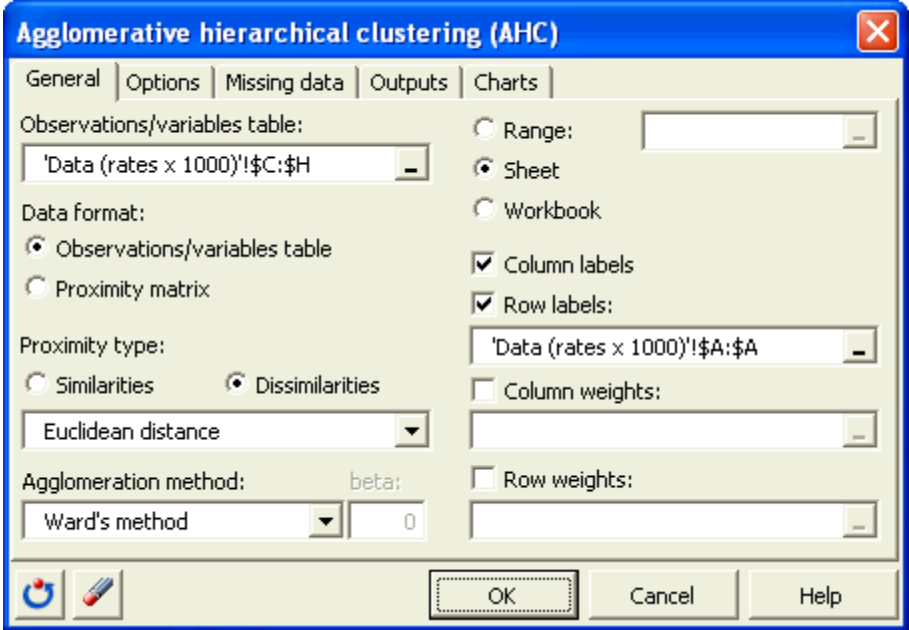


The **Hierarchical Clustering** dialog box will appear. Then select the data on the Excel sheet.

Note: There are several ways of selecting data with XLSTAT - for further information, please check the section on [selecting data](#) in the XLSTAT tutorial.

In this example, the data start from the first row, so it is quicker and easier to use columns selection. This explains why the letters corresponding to the columns are displayed in the selection boxes.

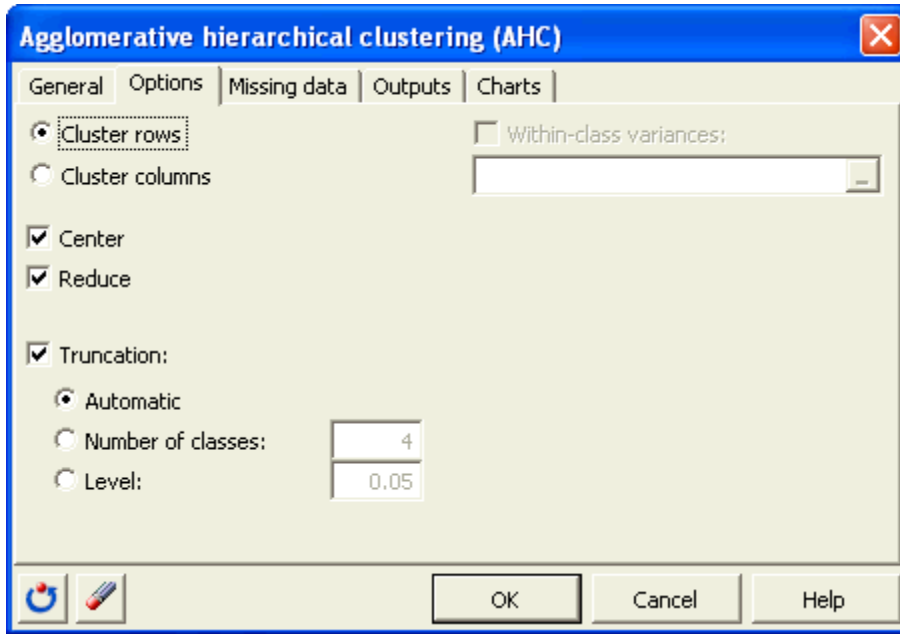
The "Total population" variable was not selected, as we are interested mainly in the demographic dynamics. The last column was not selected as it is fully correlated with the column preceding it.



In the **Options** tab, the **Center/Reduce** options were selected to avoid having group creation influenced by scaling effects.

We selected the **automatic truncation** option, so that the results show the groups to which each observation belongs.

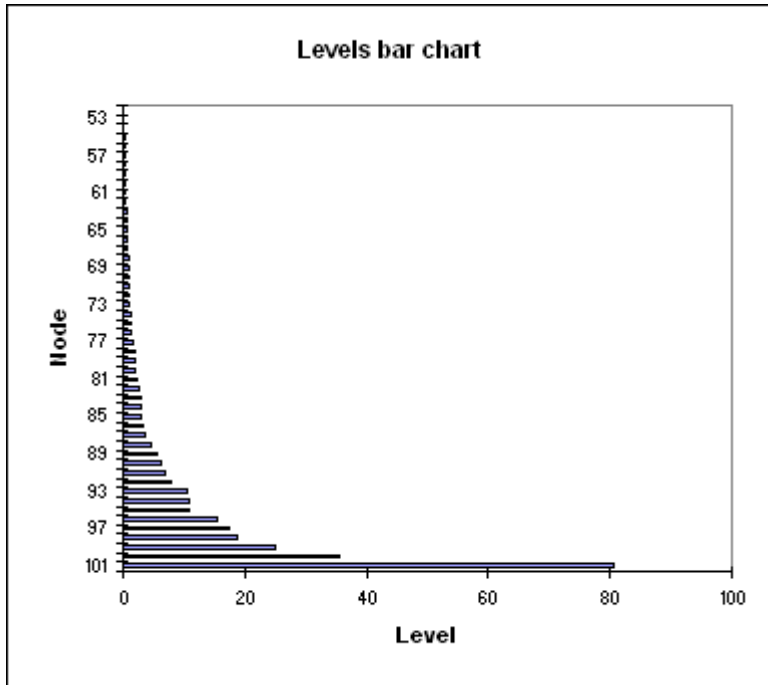
The automatic truncation is based on the entropy and tries to create homogeneous groups. However it should not prevent you from using a different number of groups either because of operational constraints, or because of your prior knowledge.



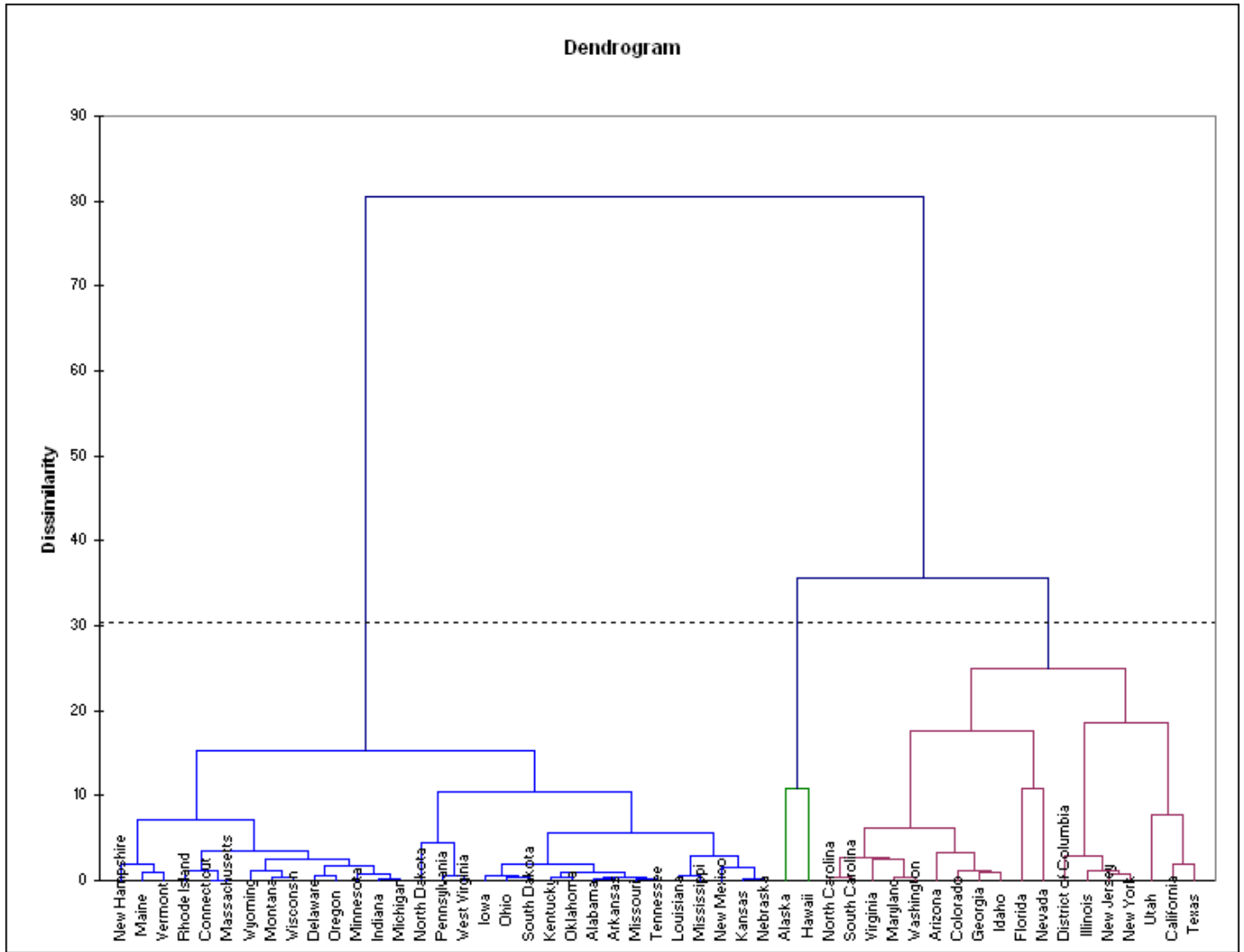
The computations begin once you have clicked on **OK**.

Interpreting the results of an Agglomerative Hierarchical Clustering

The first result to look at is the levels bar chart. The shape reveals a great deal about the structure of the data. When the increase in dissimilarity level is strong, we have reached a level where we are grouping groups that are already homogenous. Automatic truncation uses this criterion to decide when to stop aggregating observations (or groups of observations).



The chart below is the dendrogram. It represents how the algorithm works to group the observations, then the sub groups of observations. As you can see, the algorithm has successfully grouped all the observations. The dotted line represents the automatic truncation, leading to three groups.



Two groups are approximately the same size, and the third one has only two States. The first group (displayed in blue color) is more homogeneous than the third one (it is flatter on the dendrogram). This is confirmed when looking at the Within-class variable. It is a lot higher for the third group than for the first one.

The following table shows the states that have been classified into each cluster.

Results by class:			
Class	1	2	3
Objects	31	2	18
Sum of weights	31	2	18
Within-class	139,112	2876,022	454,045
Minimum class	2,132	37,921	3,222
Average distance	9,945	37,921	16,208
Maximum	23,191	37,921	58,386
	Alabama	Alaska	Arizona
	Arkansas	Hawaii	California
	Connecticut		Colorado
	Delaware	District of Columbia	
	Indiana		Florida
	Iowa		Georgia
	Kansas		Idaho
	Kentucky		Illinois
	Louisiana		Maryland
	Maine		Nevada
	Massachusetts		New Jersey
	Michigan		New York
	Minnesota		North Carolina
	Mississippi		South Carolina
	Missouri		Texas
	Montana		Utah
	Nebraska		Virginia
	New Hampshire		Washington
	New Mexico		
	North Dakota		
	Ohio		
	Oklahoma		
	Oregon		
	Pennsylvania		
	Rhode Island		
	South Dakota		
	Tennessee		
	Vermont		
	West Virginia		
	Wisconsin		
	Wyoming		

A table with the class ID for each State is displayed on the results sheet. A sample is shown below. This table is useful as it can be merged with the initial table for further analyses, for example, discriminant analysis or parallel coordinates plot.

Results by object:	
Observation	Class
Alabama	1
Alaska	2
Arizona	3
Arkansas	1
California	3
Colorado	3
Connecticut	1
Delaware	1
District of Columbia	3
Florida	3
Georgia	3
Hawaii	2
Idaho	3
Illinois	3
Indiana	1
Iowa	1
Kansas	1
Kentucky	1
Louisiana	1
Maine	1
Maryland	3
Massachusetts	1
Michigan	1
Minnesota	1
Mississippi	1
Missouri	1

This video shows how to do this tutorial.

http://www.youtube.com/watch?feature=player_embedded&v=LbAAXgXWo7E