

IBM Software

Clustering Course

Lab 5: Using clustering methods for asset selection

Contents

Using clustering methods for asset selection

1. Data import
2. Hierarchical clustering for asset selection

Summary

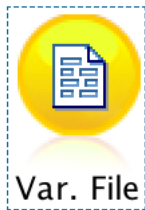
Hello everybody and welcome to another lab session!

In this lab, we are going to learn how to apply hierarchical clustering by using external SPSS Modeler extension.

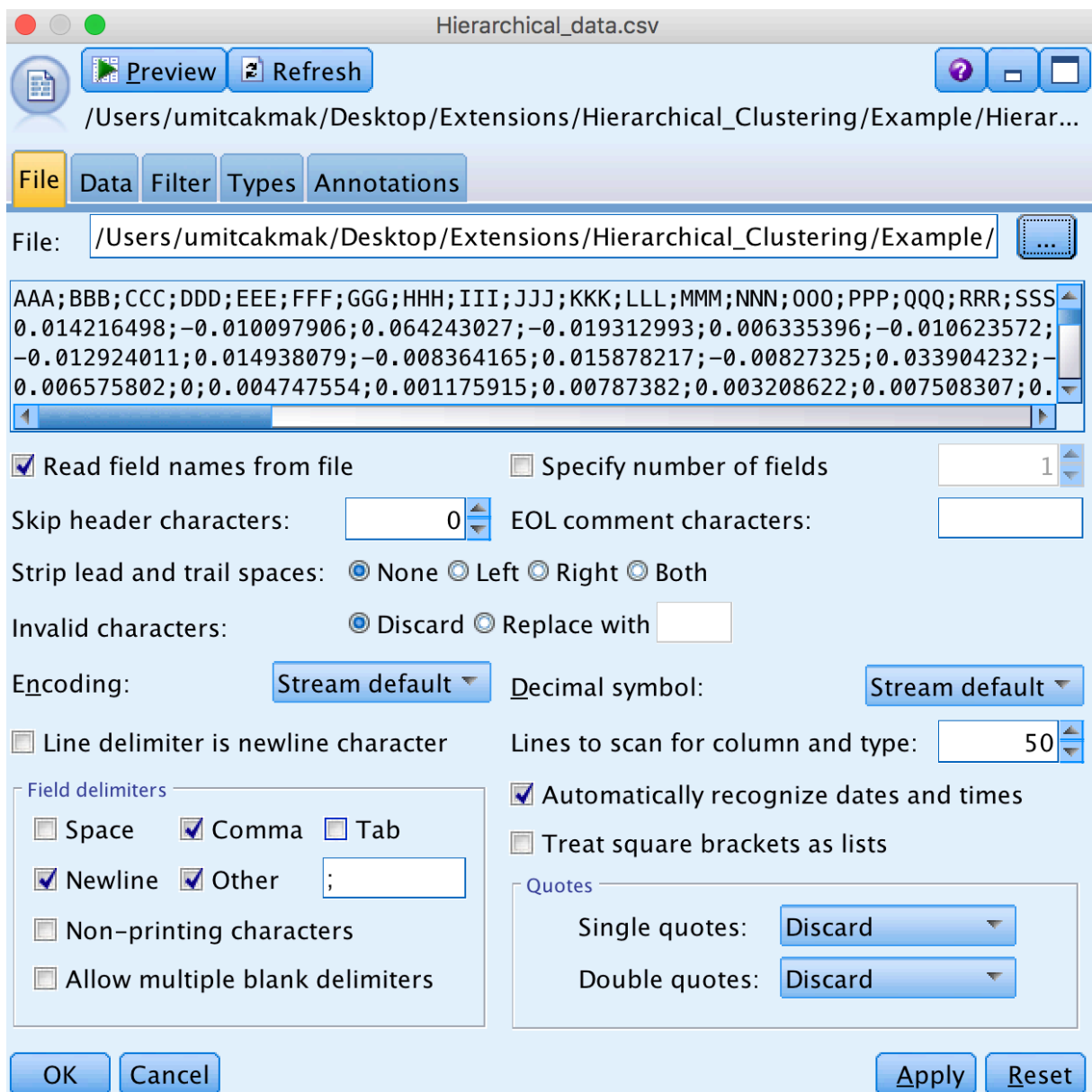
Let's get started!

1. Data import

We will start by opening a new stream. I will add “Var. File” node from sources palette.



I will browse file “Hierarchical_data.csv” which includes return series for different assets.
I will set check the box “Other” in “Field delimiters” section to “;” since that’s field separator.



The screenshot shows a dialog box titled "Hierarchical_data.csv" with a "Preview" button and a "Refresh" button. The file path is "/Users/umitcakmak/Desktop/Extensions/Hierarchical_Clustering/Example/Hierar...". The "File" tab is selected, showing a preview of the CSV data. Below the preview, there are several settings:

- ☒ Read field names from file
- ☐ Specify number of fields: 1
- Skip header characters: 0
- EOL comment characters: (empty)
- Strip lead and trail spaces: ☒ None ☐ Left ☐ Right ☐ Both
- Invalid characters: ☒ Discard ☐ Replace with (empty)
- Encoding: Stream default
- Decimal symbol: Stream default
- ☐ Line delimiter is newline character
- Lines to scan for column and type: 50
- Field delimiters: ☐ Space ☒ Comma ☐ Tab ☒ Newline ☒ Other: ;
- ☐ Non-printing characters
- ☐ Allow multiple blank delimiters
- ☒ Automatically recognize dates and times
- ☐ Treat square brackets as lists
- Quotes: Single quotes: Discard Double quotes: Discard

Buttons at the bottom: OK, Cancel, Apply, Reset.

I can click preview to see preview of the series.

Preview from Hierarchical_data.csv Node (25 fields, 10 records)

	AAA	BBB	CCC	DDD	EEE	FFF	GGG	HHH
1	0.014	-0.010	0.064	-0.019	0.006	-0.011	0.006	-0.017
2	-0.013	0.015	-0.008	0.016	-0.008	0.034	-0.010	0.011
3	0.007	0.000	0.005	0.001	0.008	0.003	0.008	0.003
4	-0.008	0.013	-0.015	0.002	-0.032	0.018	-0.017	0.016
5	0.015	-0.005	0.011	-0.008	0.010	0.004	0.010	-0.005
6	-0.009	0.009	-0.018	0.008	-0.006	0.009	-0.006	0.008
7	0.010	0.002	0.010	-0.009	-0.005	0.001	-0.005	-0.005
8	0.003	-0.001	-0.002	0.005	0.006	0.002	0.009	-0.006
9	0.015	-0.019	0.002	0.002	0.007	-0.004	-0.005	-0.035
10	0.007	-0.017	0.020	-0.008	0.009	-0.020	0.010	-0.019

OK

Once I'm done with importing data, I can add "Hcust" from "Modeling" palette.



Once I double click "Hcust" node, I see 3 parts.

Model Options Data Options Console Output Annotations

Return Series:

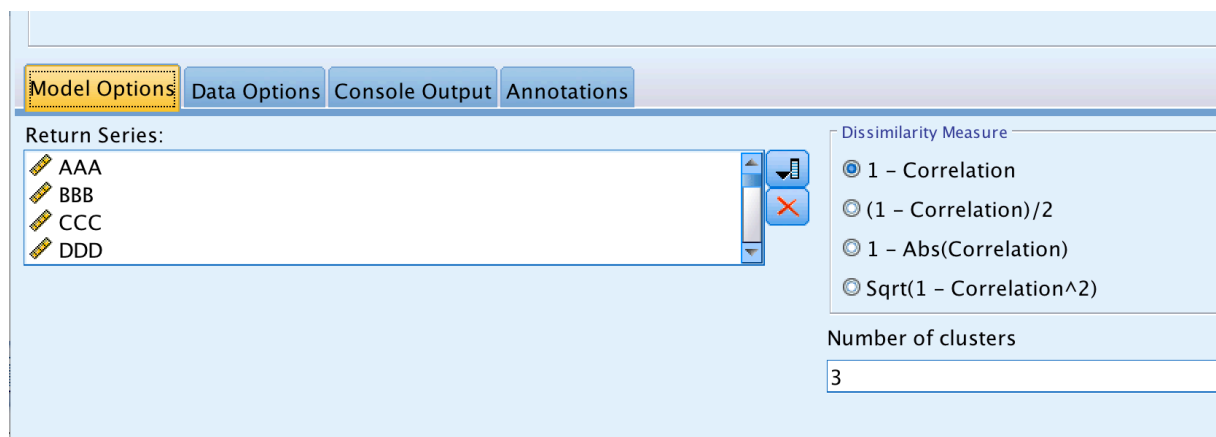
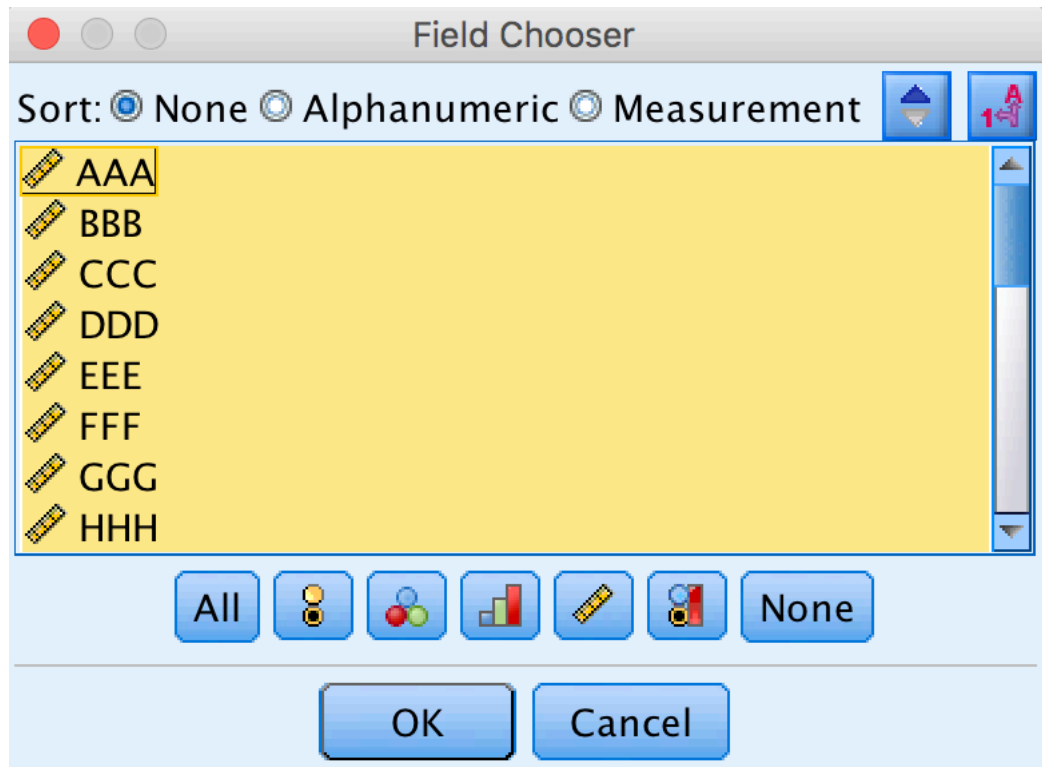
Dissimilarity Measure

- ☒ 1 - Correlation
- ☐ (1 - Correlation)/2
- ☐ 1 - Abs(Correlation)
- ☐ Sqrt(1 - Correlation^2)

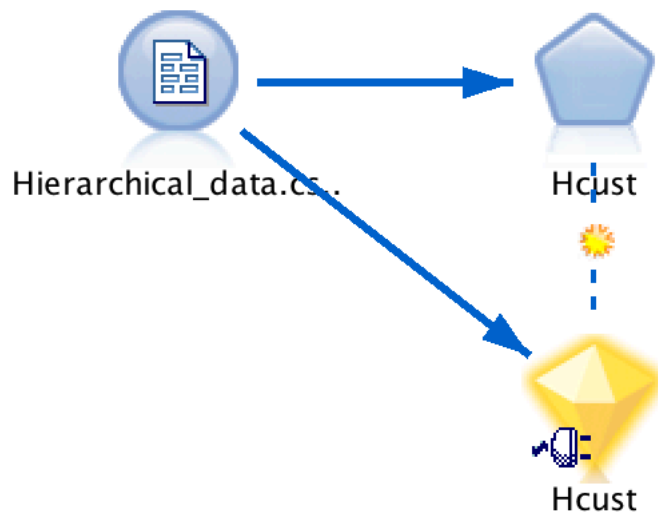
Number of clusters

First part is where I select return series to work with, second is selecting dissimilarity measure, we have 4 different dissimilarity measures here and it's good to experiment with it see if result differs. Last and third section is for defining how many clusters I would like to have as a result.

I will select all return series available and I will stick with first dissimilarity measure “1 – Correlation”, and put “3” clusters to “Number of clusters” section.

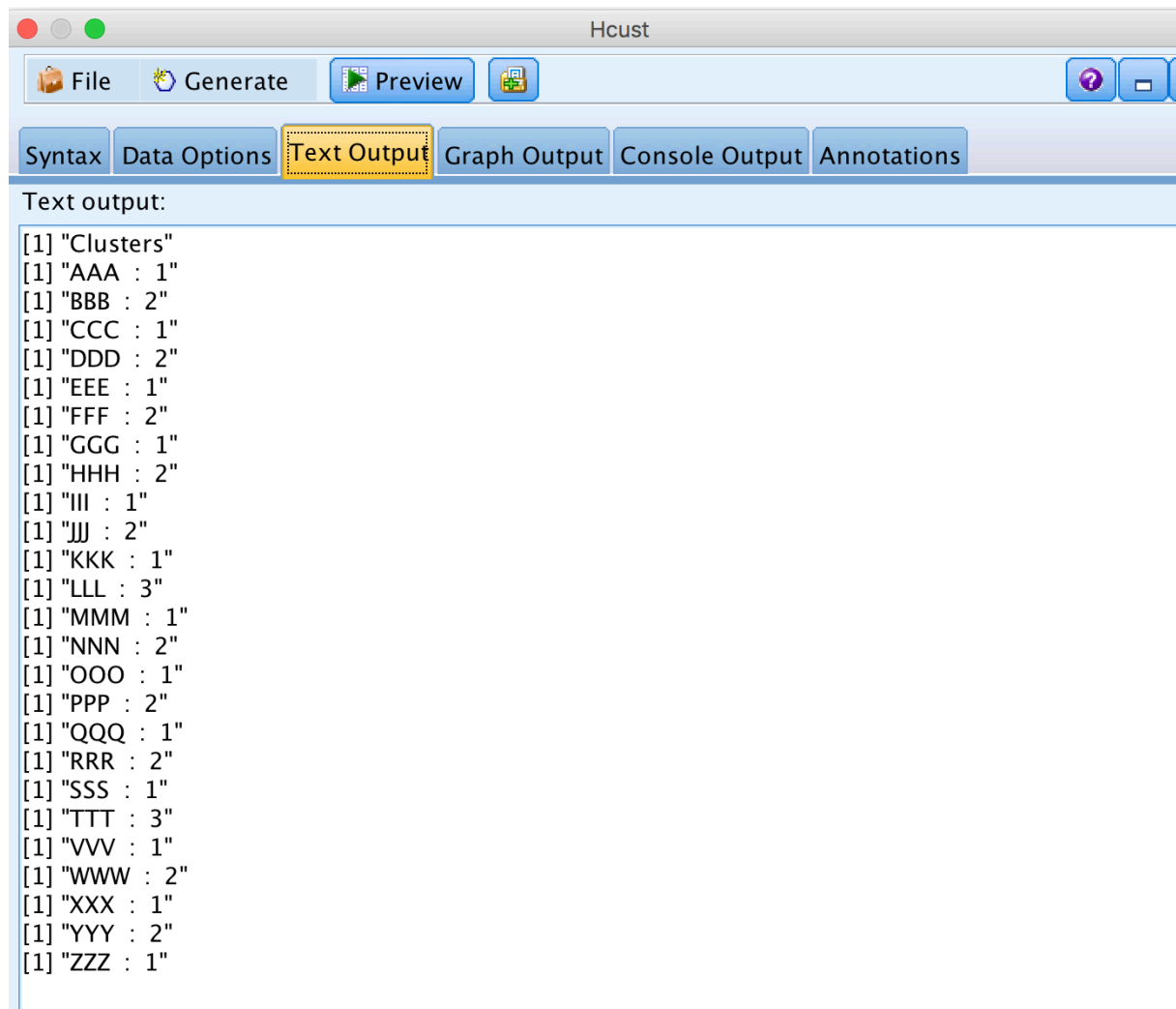


I will click to run the stream.



Resulting node is our model nugget where we have outcome. Let's open that and see hierarchical clustering done based on our dissimilarity measure.

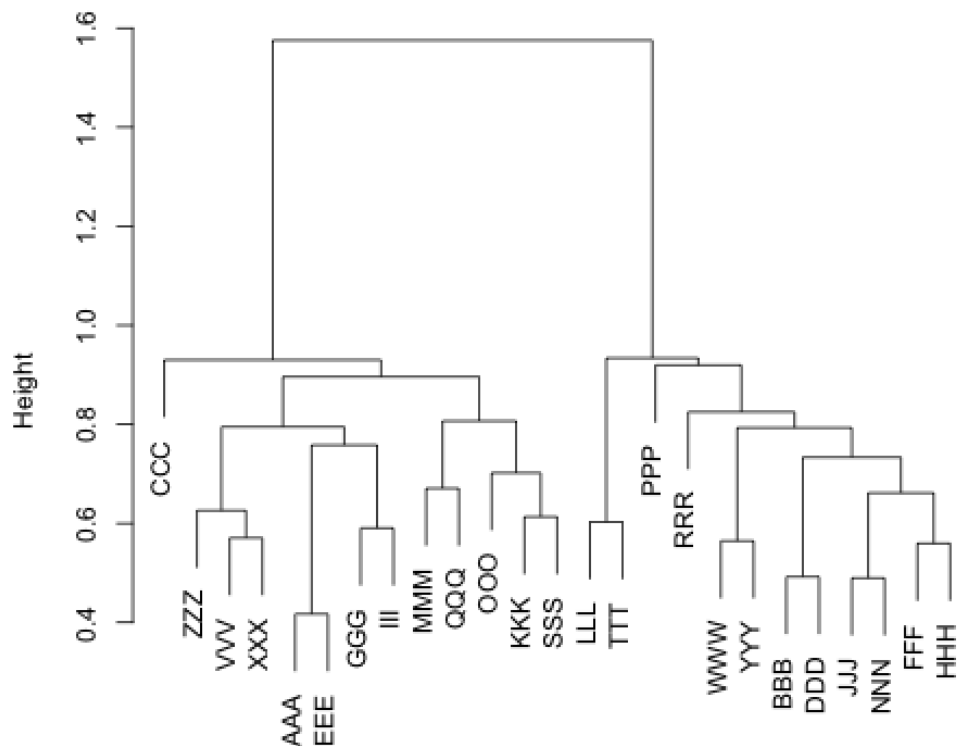
In "Text Output" section, you will see cluster memberships for each records



The screenshot shows the Hcust software interface. At the top is a title bar with the name 'Hcust'. Below it is a menu bar with 'File', 'Generate', 'Preview', and a help icon. A secondary bar contains tabs for 'Syntax', 'Data Options', 'Text Output' (which is selected and highlighted with a yellow border), 'Graph Output', 'Console Output', and 'Annotations'. The main area of the 'Text Output' tab displays a list of clusters under the heading 'Text output:'. Each line represents a cluster with its index in brackets, the cluster name in quotes, and its size.

```
[1] "Clusters"
[1] "AAA : 1"
[1] "BBB : 2"
[1] "CCC : 1"
[1] "DDD : 2"
[1] "EEE : 1"
[1] "FFF : 2"
[1] "GGG : 1"
[1] "HHH : 2"
[1] "III : 1"
[1] "JJJ : 2"
[1] "KKK : 1"
[1] "LLL : 3"
[1] "MMM : 1"
[1] "NNN : 2"
[1] "OOO : 1"
[1] "PPP : 2"
[1] "QQQ : 1"
[1] "RRR : 2"
[1] "SSS : 1"
[1] "TTT : 3"
[1] "VVV : 1"
[1] "WWW : 2"
[1] "XXX : 1"
[1] "YYY : 2"
[1] "ZZZ : 1"
```

We can also navigate to “Graph Output” tab to see dendogram.



We can visualize clusters by using dendrogram.

We can already see logical groupings here starting from 2 big clusters.

Summary

As you can see hierarchical clustering helps us to logically group assets based on their correlations. This could definitely gives us a new and improved perspective in asset selection rather than assuming that grouping based on companies principal business is giving us correct picture.

I hope you enjoyed the course as well as the labs. I suggest you keep experimenting with SPSS Modeler and try and learn new features which will definitely help you in your journey. Thank you very much for your time and hope to see you next time.