An overview of most common Statistical packages for data analysis

Antonio Lucadamo

Università del Sannio - Italy
antonio.lucadamo@unisannio.it

Workshop in Methodology of Teaching Statistics

Novi Sad, December, 13 - 2011

- Pearson (1901)
- Spearman (1904)
- Hotelling (1933-1936)

- Pearson (1901)
- Spearman (1904)
- Hotelling (1933-1936)

The absence of adequate tools has been, for many years, an obstacle to the development of these methods. They were studied only in a theoretical context. (Multivariate analysis)

- Pearson (1901)
- Spearman (1904)
- Hotelling (1933-1936)

The absence of adequate tools has been, for many years, an obstacle to the development of these methods. They were studied only in a theoretical context. (Multivariate analysis)

1960-1970: Benzécri - Analyse des données (Multidimensional Data Analysis)

Multivariate Analysis vs. Multidimensional Analysis

Multivariate Analysis vs. Multidimensional Analysis

Distributional hypotheses vs. Structural hypotheses

Multivariate Analysis vs. Multidimensional Analysis

Distributional hypotheses vs. Structural hypotheses

'The model must follow the data and not viceversa'

Multivariate Analysis vs. Multidimensional Analysis

Distributional hypotheses vs. Structural hypotheses

'The model must follow the data and not viceversa'

1980s: trade-off between the two positions.

Multivariate Analysis vs. Multidimensional Analysis

Distributional hypotheses vs. Structural hypotheses

'The model must follow the data and not viceversa'

1980s: trade-off between the two positions.

Multidimensional analysis may be defined as a group of techniques that have the aim to visualize, classify and interprete the data. It try to underline the latent structure of the data, removing the redundant information.

- Principal Component Analysis
- Correspondence Analysis
- Discriminant Analysis
- Canonical Correlation Analysis
- Cluster Analysis

Principal Component Analysis

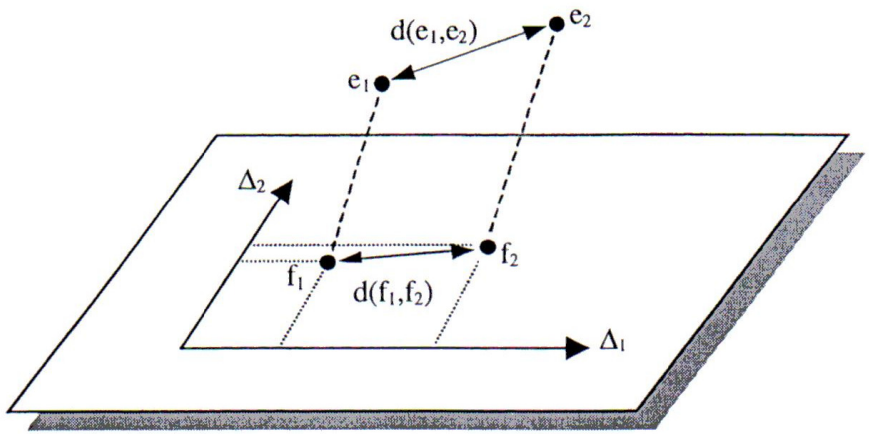
Hypotesis: the new factors are linear combinations of the original variables.

Now few factors are sufficient to explain the most part of the variability in the data.

Principal Component Analysis

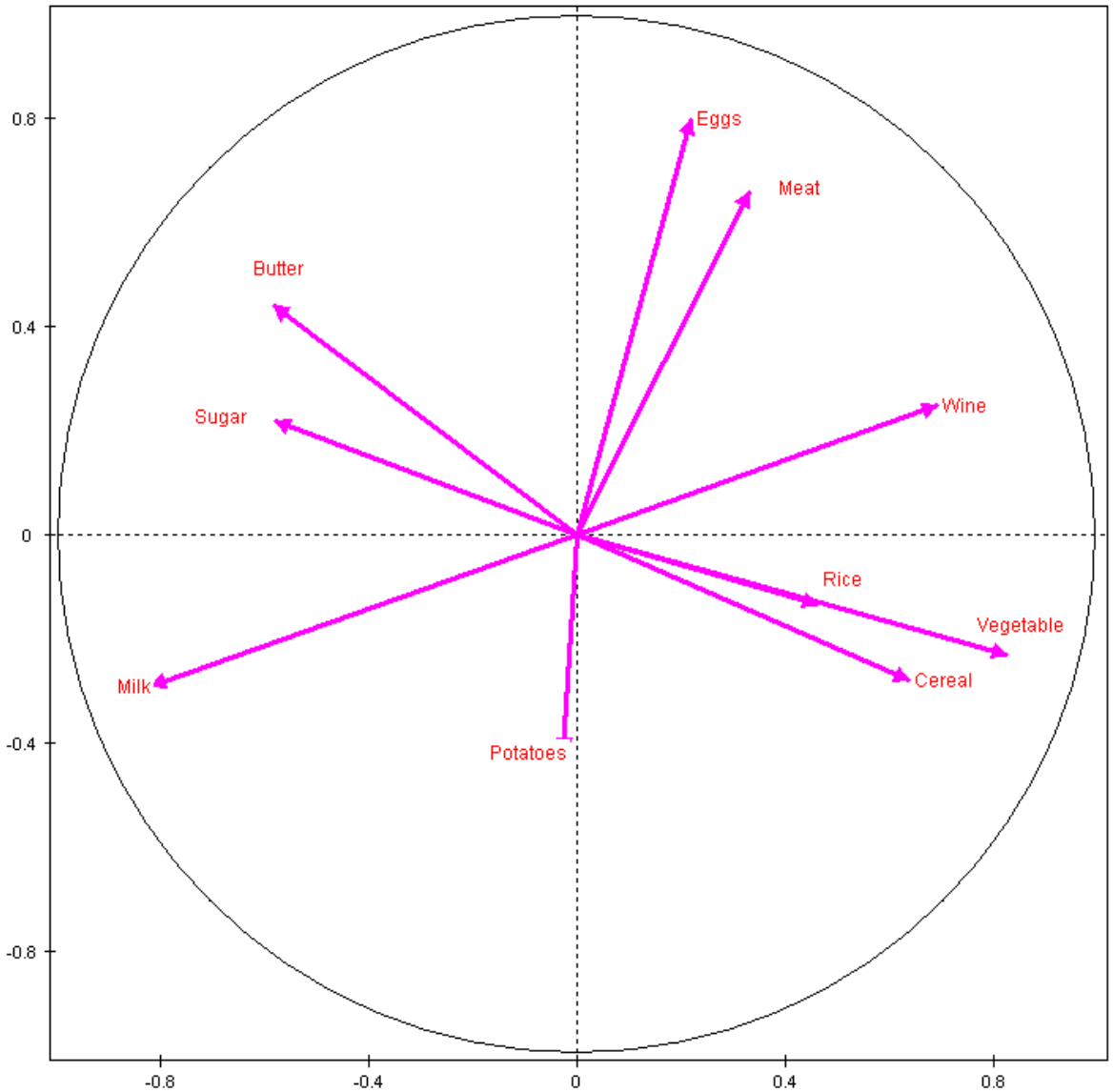
Hypotesis: the new factors are linear combinations of the original variables.
Now few factors are sufficient to explain the most part of the variability in the data.

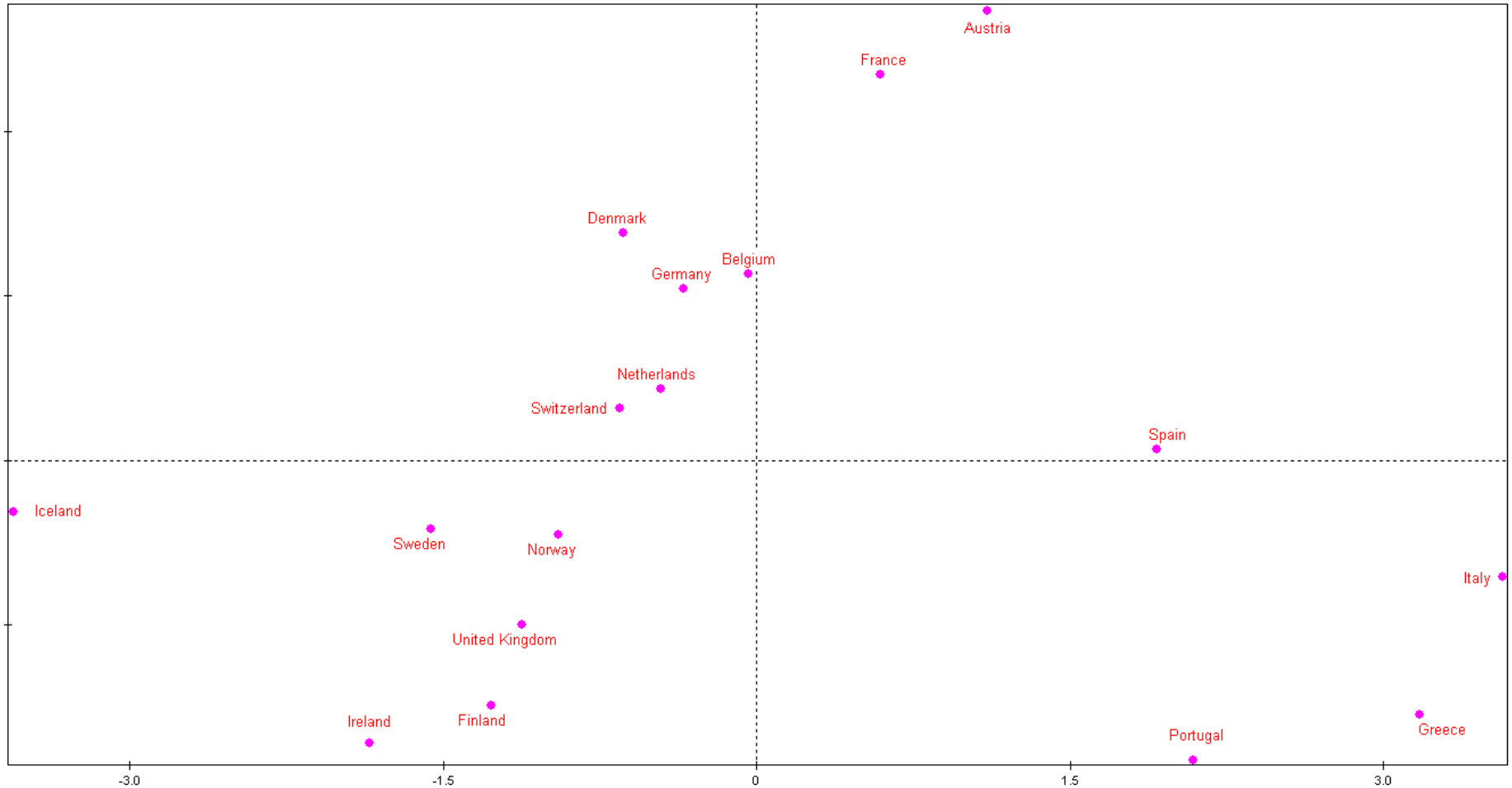
Geometric interpretation



An example

	Cereal	Rice	Potatoes	Sugar	Vegetable	Meat	Milk	Butter	Eggs	Wine
Belgium	73,8	3,9	99	40,3	110,7	103,2	83,4	6,4	14,4	19,5
Denmark	71	2,3	57	40,2	45	105,6	143,4	2,3	16,1	23,1
Germany	71,2	2,5	73,3	32,4	80	93,1	92,3	6,9	13,3	23,3
Greece	104,7	5,2	87,4	29,1	246,9	83,2	64,4	1,1	10,9	30
Spain	73,6	6,6	92,3	28,6	162,1	108,4	125,8	0,5	14,9	42,5
France	80,4	4,2	72,6	34,4	76	106,8	95,3	8,6	15,8	63,5
Ireland	77,9	3,2	171,9	38,1	87,7	90,6	196,2	5,9	9,3	5,8
Italy	120,1	4,9	41	25,6	175,4	89,4	62,1	2,2	10,5	62,8
Netherlands	50,4	8,3	81,8	30,8	118,5	90,2	129	6	13,2	13,1
Austria	63,3	10	60,6	34	79,8	234	111,1	5,2	13,7	43
Portugal	68	15,5	145,5	28,8	112,9	87	100,8	1,5	8,7	58,8
Finland	66,4	6,6	59,7	14,4	63,1	43	201,5	5,3	10,4	5,5
Sweden	65	2,5	121	42,7	45	65	153,4	5,8	10,1	43
United Kingd	82,9	3,7	108,3	36,6	34	73,3	138,5	4,1	10,2	11,6
Iceland	58,2	2,5	47,1	54,5	33,5	67,5	201,3	6,2	8,6	6,4
Norway	80,5	5,4	42,7	44,5	58,3	56,9	160,2	3	11,3	24
Switzerland	66,1	5,2	44,7	42,3	84,4	60,7	115,6	6,3	10,5	41,3


Eggs
Meat
Butter
Wine
Sugar
Rice
Vegetable
Cereal
Milk
Potatoes
0.8
0.4
0
-0.4
-0.8
-0.8
-0.4
0
0.4
0.8


Austria
France
Denmark
Belgium
Germany
Netherlands
Switzerland
Spain
Iceland
Sweden
Norway
Italy
United Kingdom
Ireland
Finland
Portugal
Greece
-3.0
-1.5
0
1.5
3.0

Correspondence Analysis

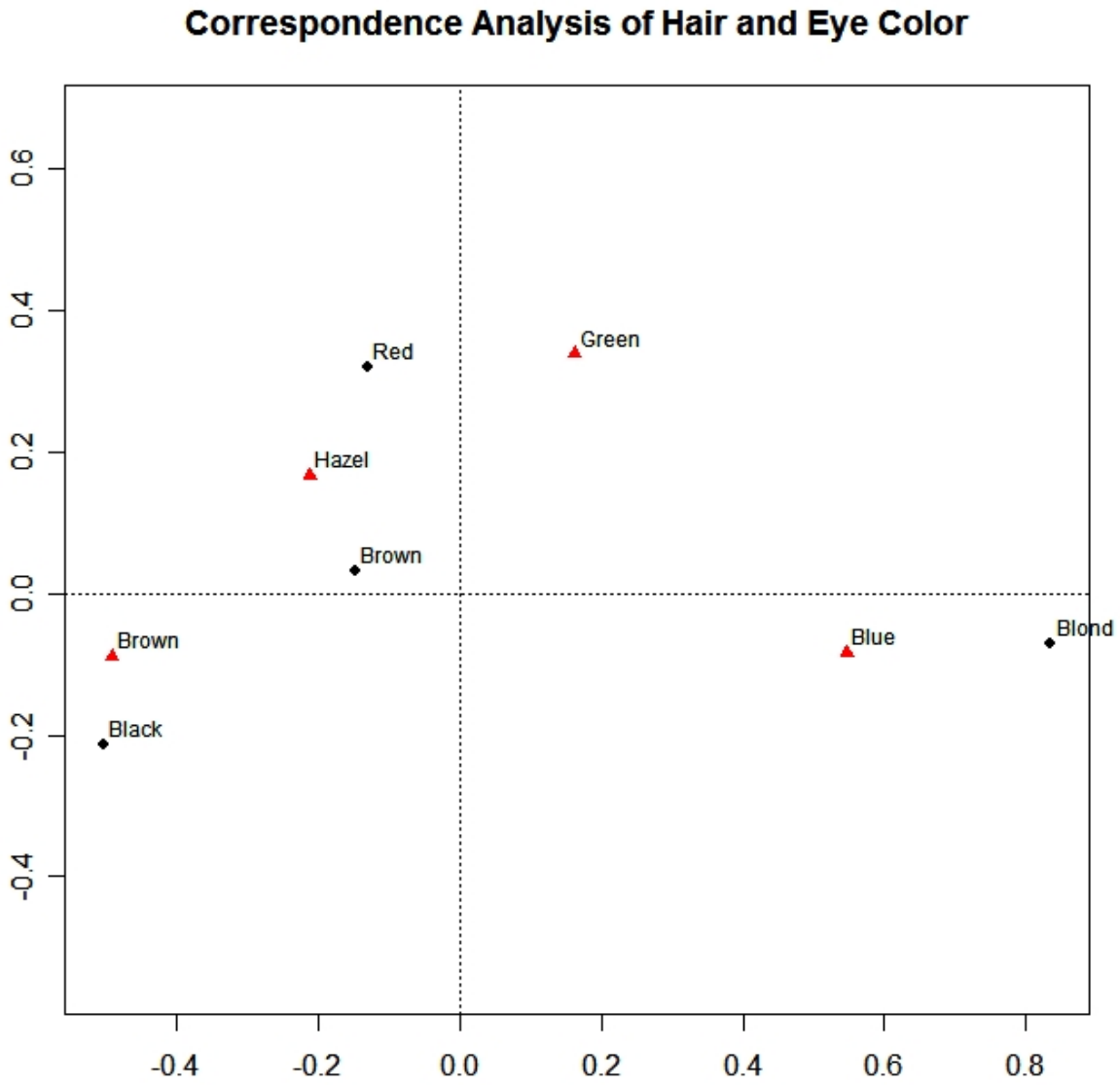
Simple Correspondence analysis is one of the most known tools for qualitative data.

Correspondence Analysis

Simple Correspondence analysis is one of the most known tools for qualitative data.

It studies the relationships between the modalities of two qualitative variables.

Correspondence Analysis



Multiple Correspondence Analysis

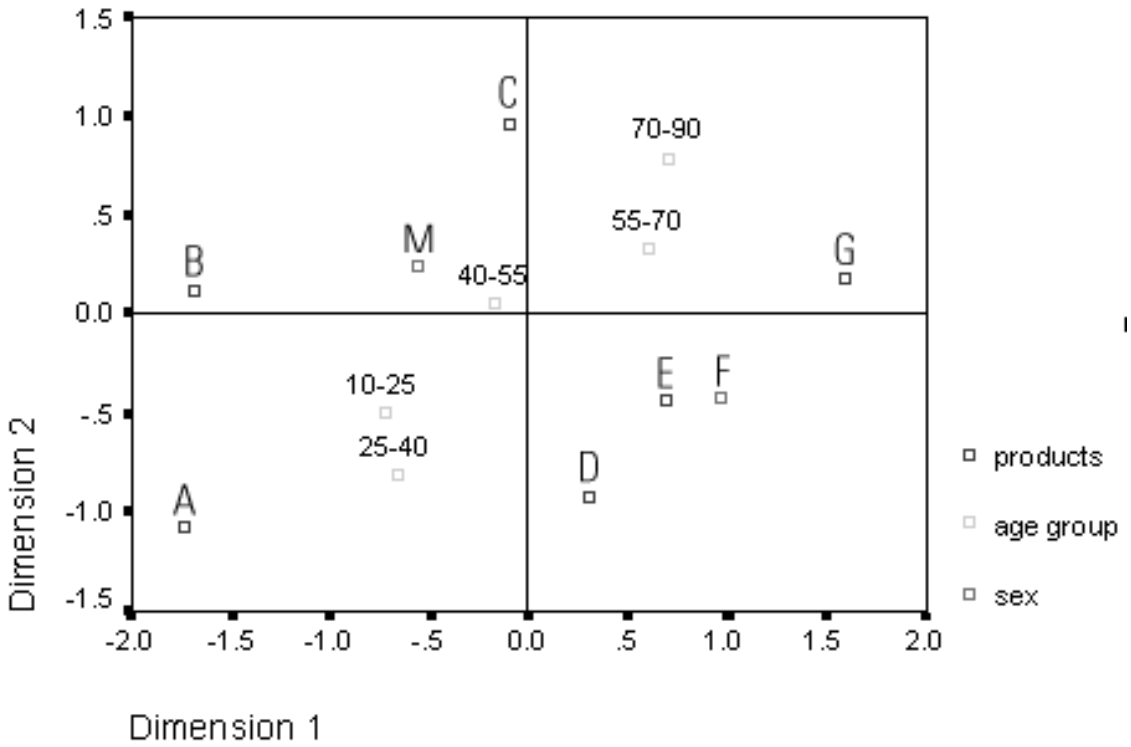
When there are more than 2 qualitative variables the Simple Correspondence Analysis is not possible and the relationships between the characters may be studied with the Multiple Correspondence Analysis.

Multiple Correspondence Analysis

When there are more than 2 qualitative variables the Simple Correspondence Analysis is not possible and the relationships between the characters may be studied with the Multiple Correspondence Analysis.

This technique is used in economic sciences, healt science, marketing analysis.

Multiple Correspondence Analysis



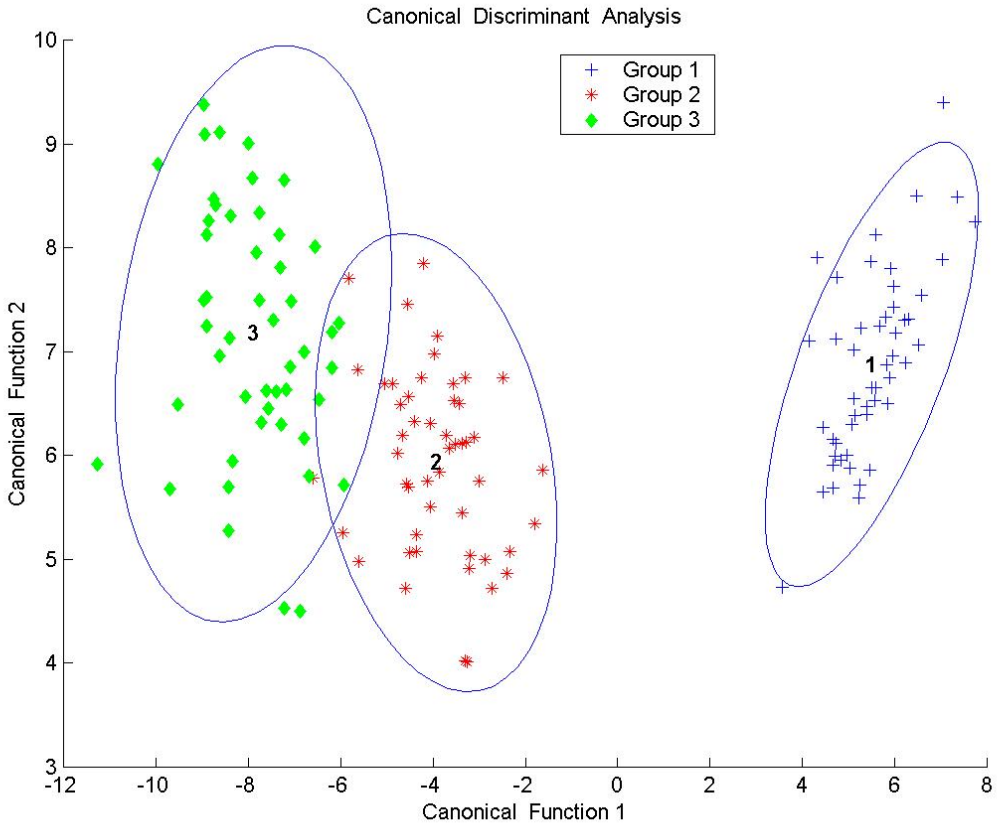
Discriminant analysis

Descriptive aim: verify if the prior classification is confirmed after using the explicative variables.

Discriminant analysis

Descriptive aim: verify if the prior classification is confirmed after using the explicative variables.

Decision aim: classify a new observation in one of the group.

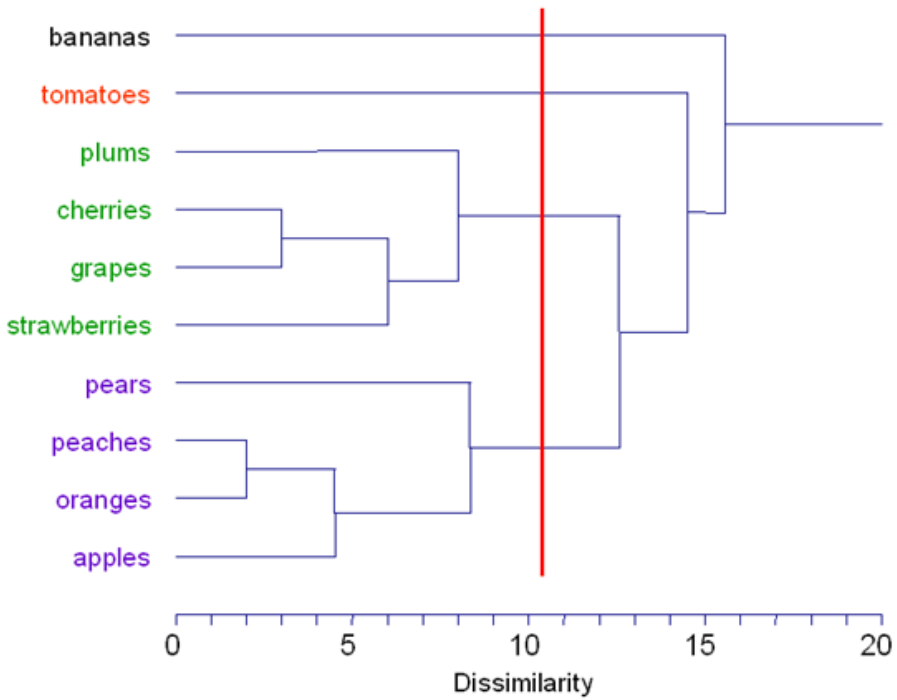


Cluster Analysis

Group of techniques that have the aim to classify observations or individuals in clusters.

The observations in each cluster must be similar and the clusters must be well separated.
Partition and hierarchy.

Cluster Analysis



List of some softwares

Most used softwares for Multidimensional Data Analysis

- Spad
- XI-stat
- Spss
- S-plus
- R
- Pspp

SPAD

Advantages

It can perform many statistical analysis:

- Descriptive Statistics
- Factorial Analysis
- Classification
- Segmentation
- Textual analysis

SPAD

Advantages

It can perform many statistical analysis:

- Descriptive Statistics
- Factorial Analysis
- Classification
- Segmentation
- Textual analysis

It has good graphical tools and it is easy to use.

SPAD

Disadvantages

Data importation is not direct.

SPAD

Disadvantages

Data importation is not direct.

It is expensive:

Price for University (single user)	1050 €
Price for University (15 users)	3000 €
Price for others	21000 €

SPAD

Disadvantages

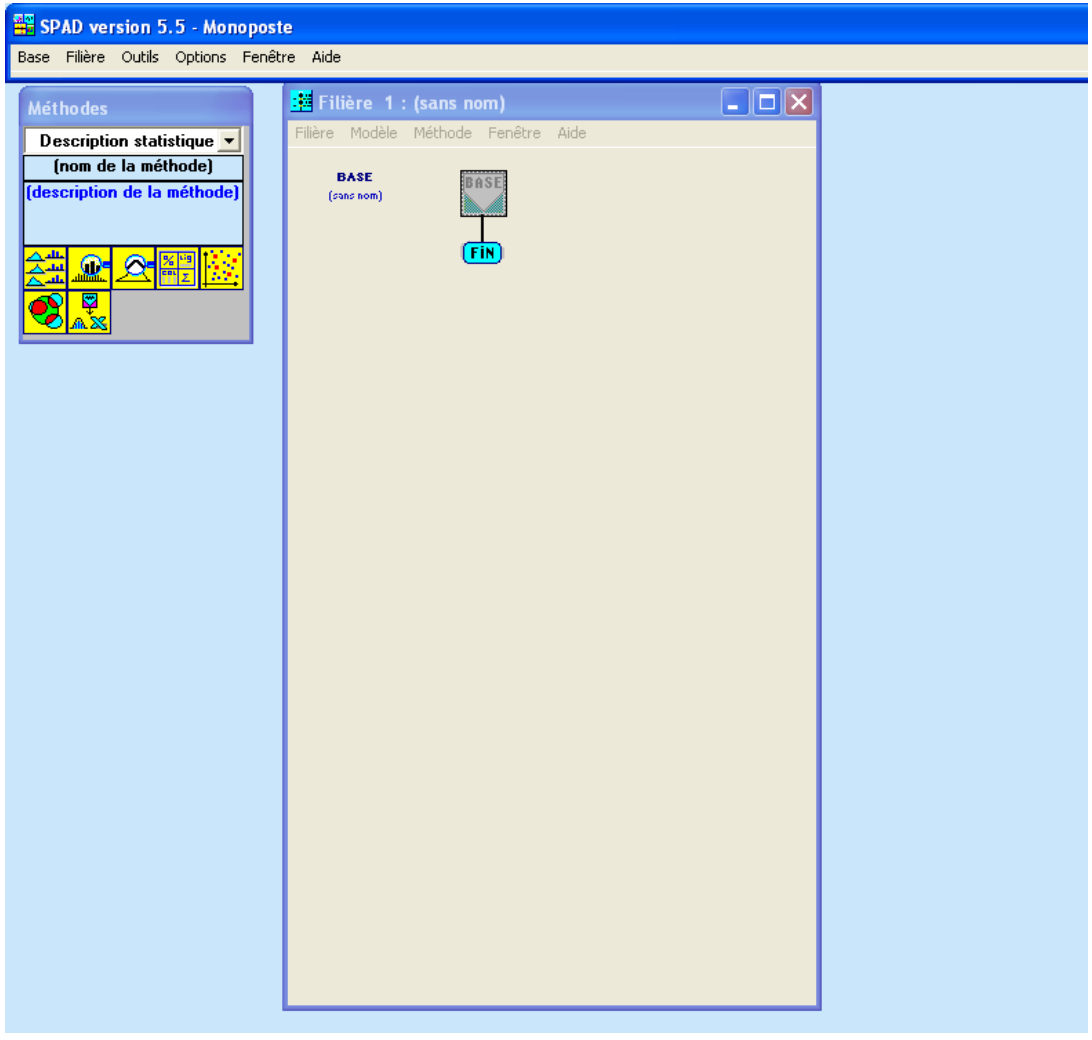
Data importation is not direct.

It is expensive:

Price for University (single user)	1050 €
Price for University (15 users)	3000 €
Price for others	21000 €

Link: http://www.coheris.fr/fr/page/produits/Spad.html

SPAD


SPAD version 5.5 - Monoposte
Base Filière Outils Options Fenêtre Aide
Méthodes
Description statistique
(nom de la méthode)
(description de la méthode)
Filière 1 : (sans nom)
Filière Modèle Méthode Fenêtre Aide
BASE
(sans nom)
BASE
FIN

XLSTAT

Advantages

Very easy to use, excel interface.

XLSTAT

Advantages

Very easy to use, excel interface.

Possibility to download a trial version for 30 days.

XLSTAT

Advantages

Very easy to use, excel interface.

Possibility to download a trial version for 30 days.

It is less expensive than SPAD:

XLSTAT

Advantages

Very easy to use, excel interface.

Possibility to download a trial version for 30 days.

It is less expensive than SPAD:

Price for single user	295 €
Price for 20 users	2495 €

XLSTAT

Advantages

Very easy to use, excel interface.

Possibility to download a trial version for 30 days.

It is less expensive than SPAD:

Price for single user	295 €
Price for 20 users	2495 €

Disadvantages

About Multidimensional Analysis it has less options than SPAD.

XLSTAT

Advantages

Very easy to use, excel interface.

Possibility to download a trial version for 30 days.

It is less expensive than SPAD:

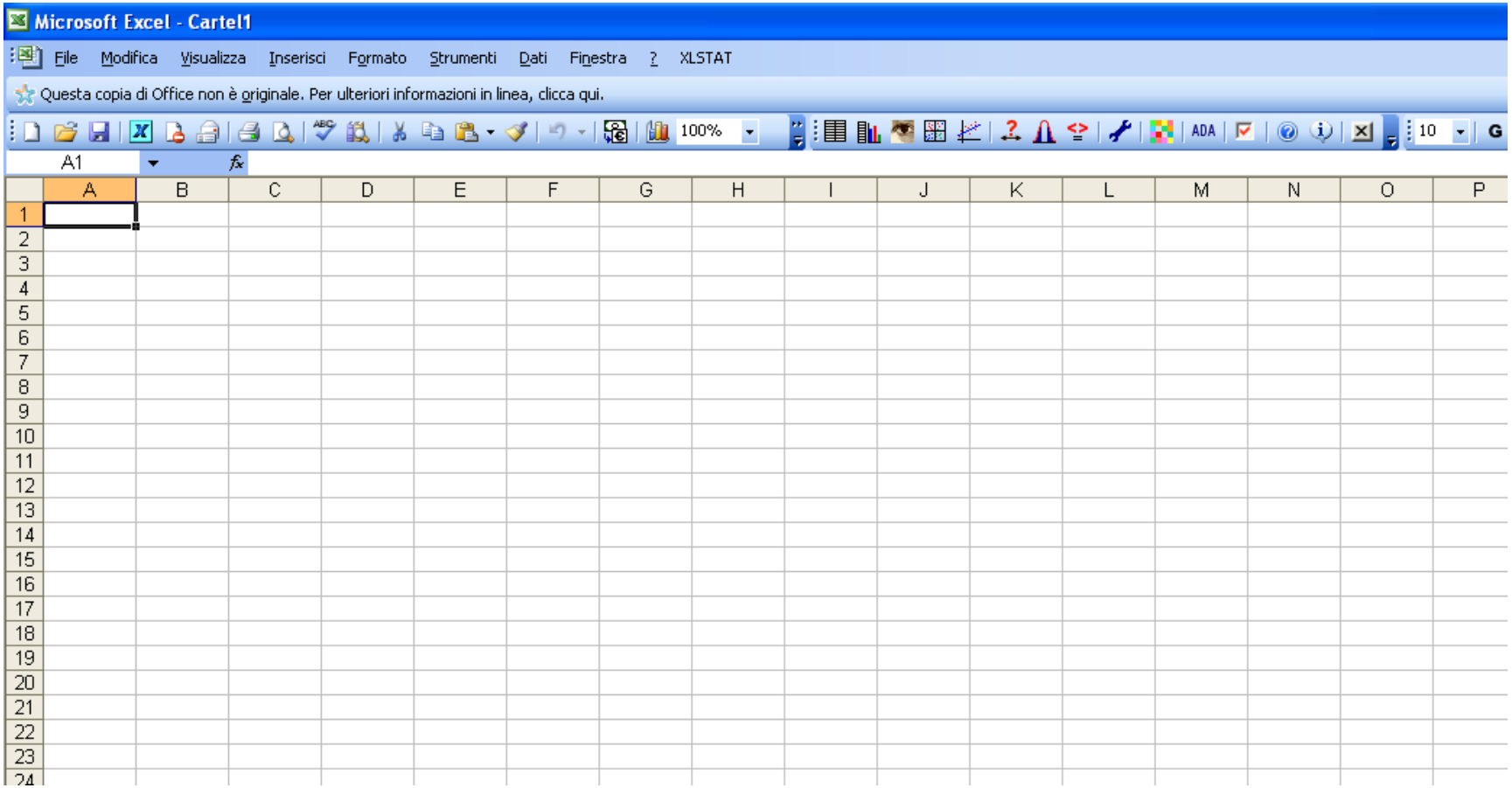
Price for single user	295 €
Price for 20 users	2495 €

Disadvantages

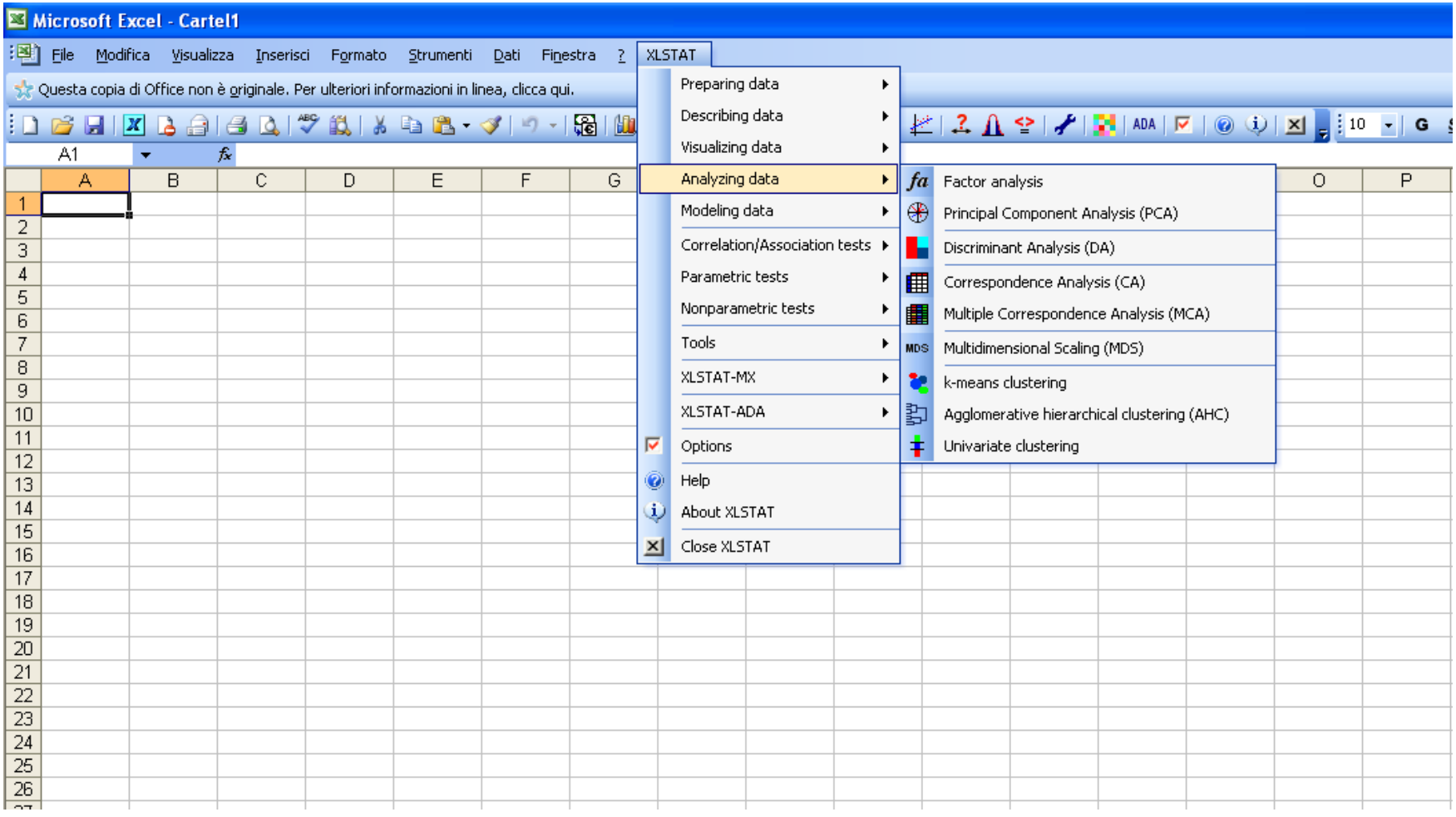
About Multidimensional Analysis it has less options than SPAD.

Link: http://www.xlstat.com/en/

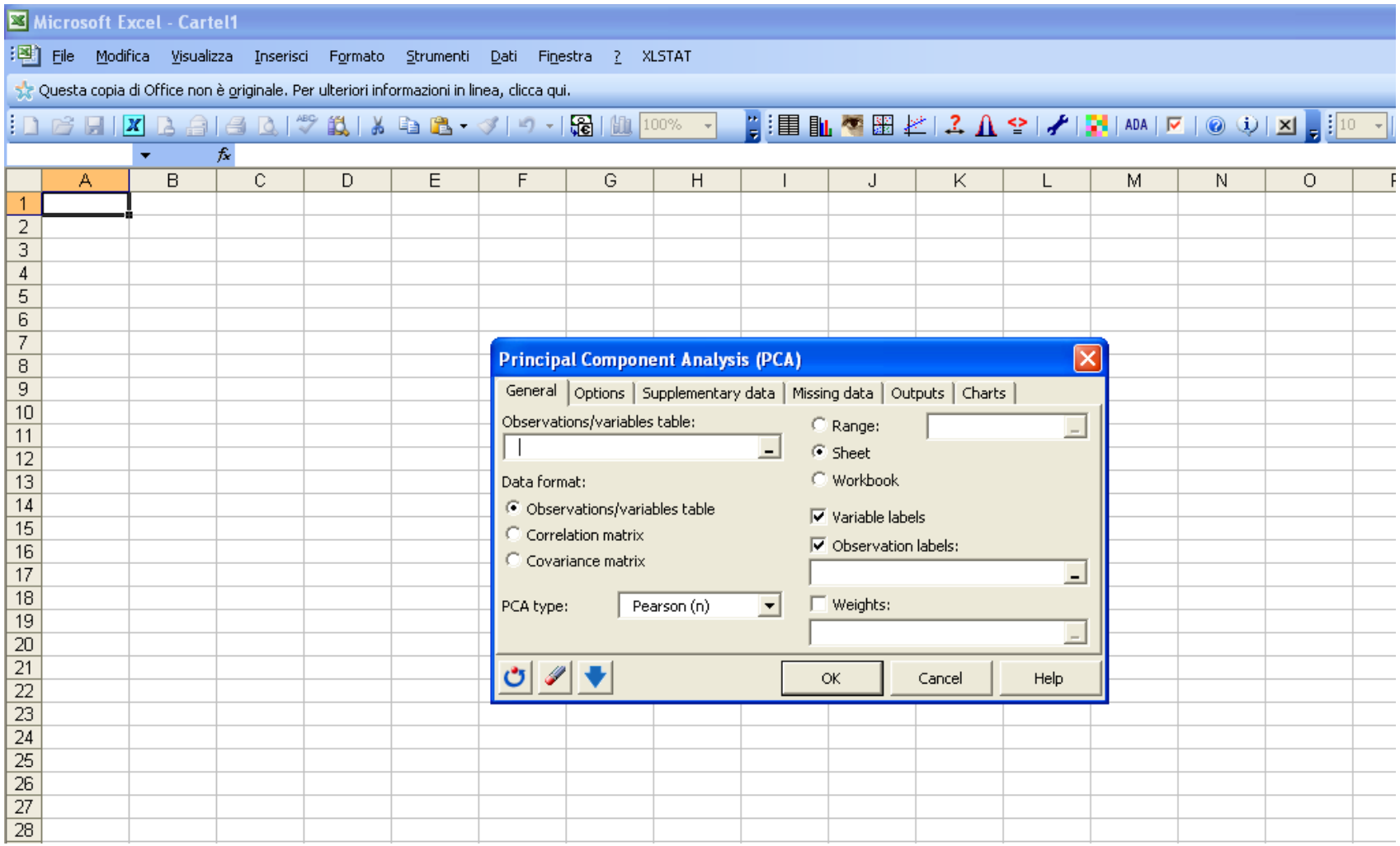
XLSTAT


Microsoft Excel - Cartel1
File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ? XLSTAT
Questa copia di Office non è originale. Per ulteriori informazioni in linea, clicca qui.
100%
ADA
10
A1
A B C D E F G H I J K L M N O P
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24

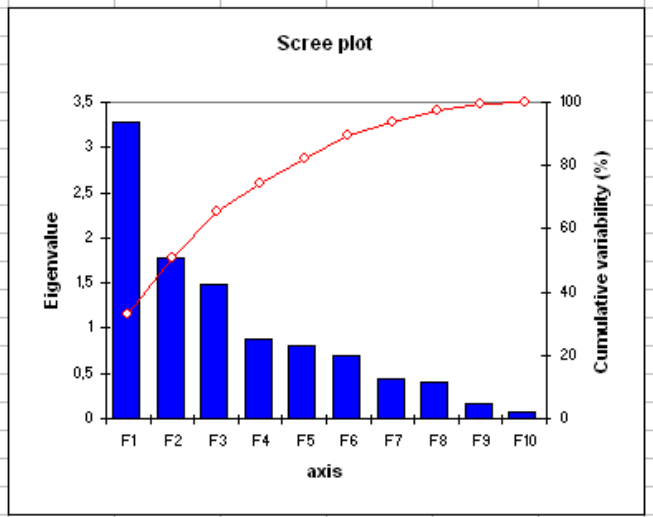
XLSTAT


Microsoft Excel - Cartel1
File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ? XLSTAT
Questa copia di Office non è originale. Per ulteriori informazioni in linea, clicca qui.
A1
Preparing data
Describing data
Visualizing data
Analyzing data
Modeling data
Correlation/Association tests
Parametric tests
Nonparametric tests
Tools
XLSTAT-MX
XLSTAT-ADA
Options
Help
About XLSTAT
Close XLSTAT
Factor analysis
Principal Component Analysis (PCA)
Discriminant Analysis (DA)
Correspondence Analysis (CA)
Multiple Correspondence Analysis (MCA)
Multidimensional Scaling (MDS)
k-means clustering
Agglomerative hierarchical clustering (AHC)
Univariate clustering

XLSTAT


Microsoft Excel - Cartel1
File Modifica Visualizza Inserisci Formato Strumenti Dati Finestra ? XLSTAT
Questa copia di Office non è originale. Per ulteriori informazioni in linea, clicca qui.
Principal Component Analysis (PCA)
General Options Supplementary data Missing data Outputs Charts
Observations/variables table:
Range:
Sheet
Workbook
Data format:
Observations/variables table
Correlation matrix
Covariance matrix
Variable labels
Observation labels:
PCA type:
Pearson (n)
Weights:
OK
Cancel
Help

XLSTAT



Eigenvectors:

	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
Cereal	0,351	-0,208	-0,493	0,145	0,087	-0,017	0,195	-0,471	-0,190	0,515
Rice	0,255	-0,100	0,643	-0,077	-0,357	0,045	0,057	0,175	-0,215	0,543
Potatoes	-0,015	-0,315	0,398	0,496	0,650	-0,027	-0,253	-0,086	-0,041	0,016
Sugar	-0,320	0,164	-0,193	0,575	-0,232	-0,506	-0,002	0,340	0,019	0,279
Vegetable	0,455	-0,173	-0,114	-0,063	0,241	-0,065	0,246	0,520	0,588	0,084
Meat	0,183	0,492	0,328	0,080	0,140	-0,376	0,545	-0,345	0,077	-0,165
Milk	-0,453	-0,217	0,144	-0,255	-0,055	-0,170	-0,044	-0,393	0,611	0,317
Butter	-0,321	0,330	-0,002	0,222	0,161	0,703	0,345	0,087	0,120	0,277
Eggs	0,120	0,598	-0,065	-0,292	0,347	-0,109	-0,520	0,053	0,003	0,365
Wine	0,382	0,186	0,045	0,431	-0,403	0,244	-0,382	-0,266	0,419	-0,131

Factor loadings:

dieta2 PCA

SPSS

Advantages

- Easy to use, similar to excel.
- It is simple to find manual or tutorial on internet that show how to use it.
- It has a large diffusion also if it is not open source.

SPSS

Advantages

- Easy to use, similar to excel.
- It is simple to find manual or tutorial on internet that show how to use it.
- It has a large diffusion also if it is not open source.

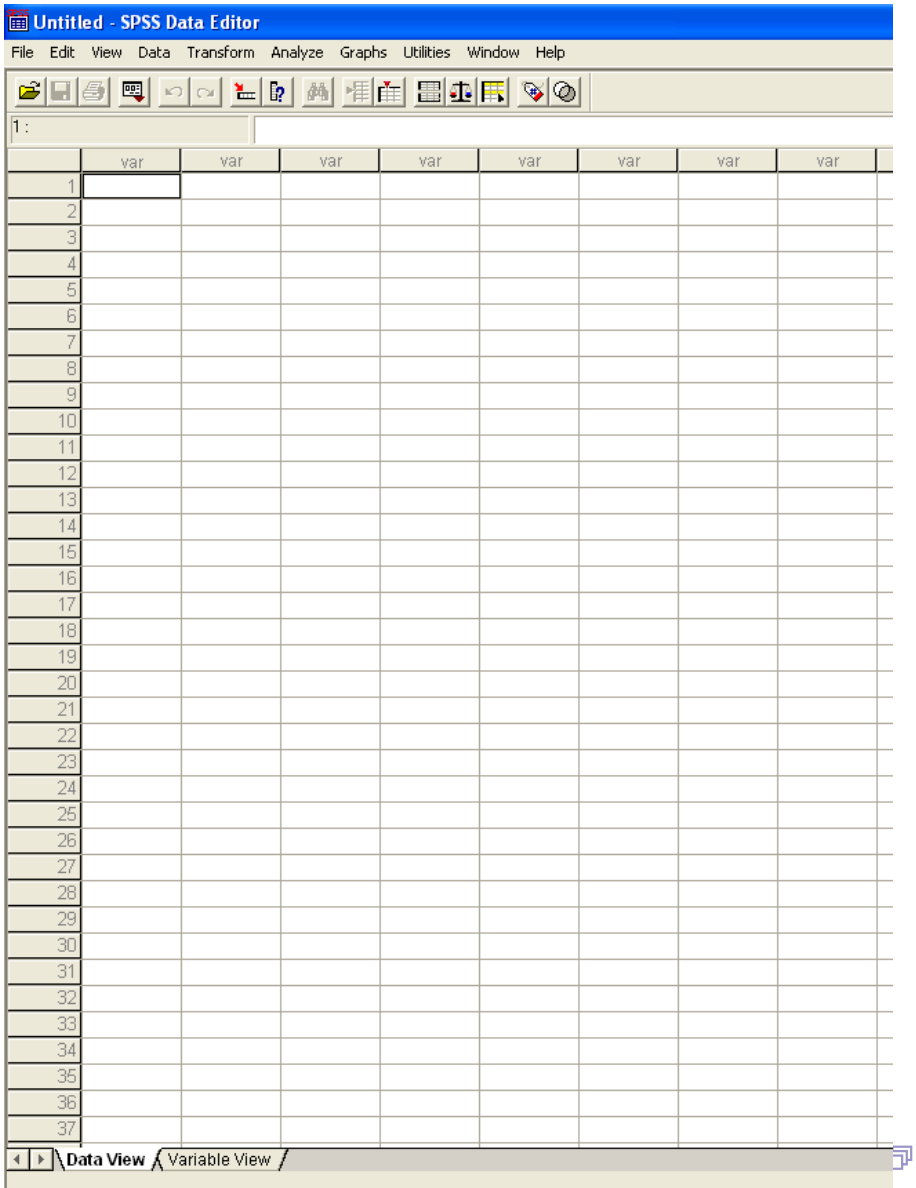
Disadvantages

For some analyses it has less options than other packages.

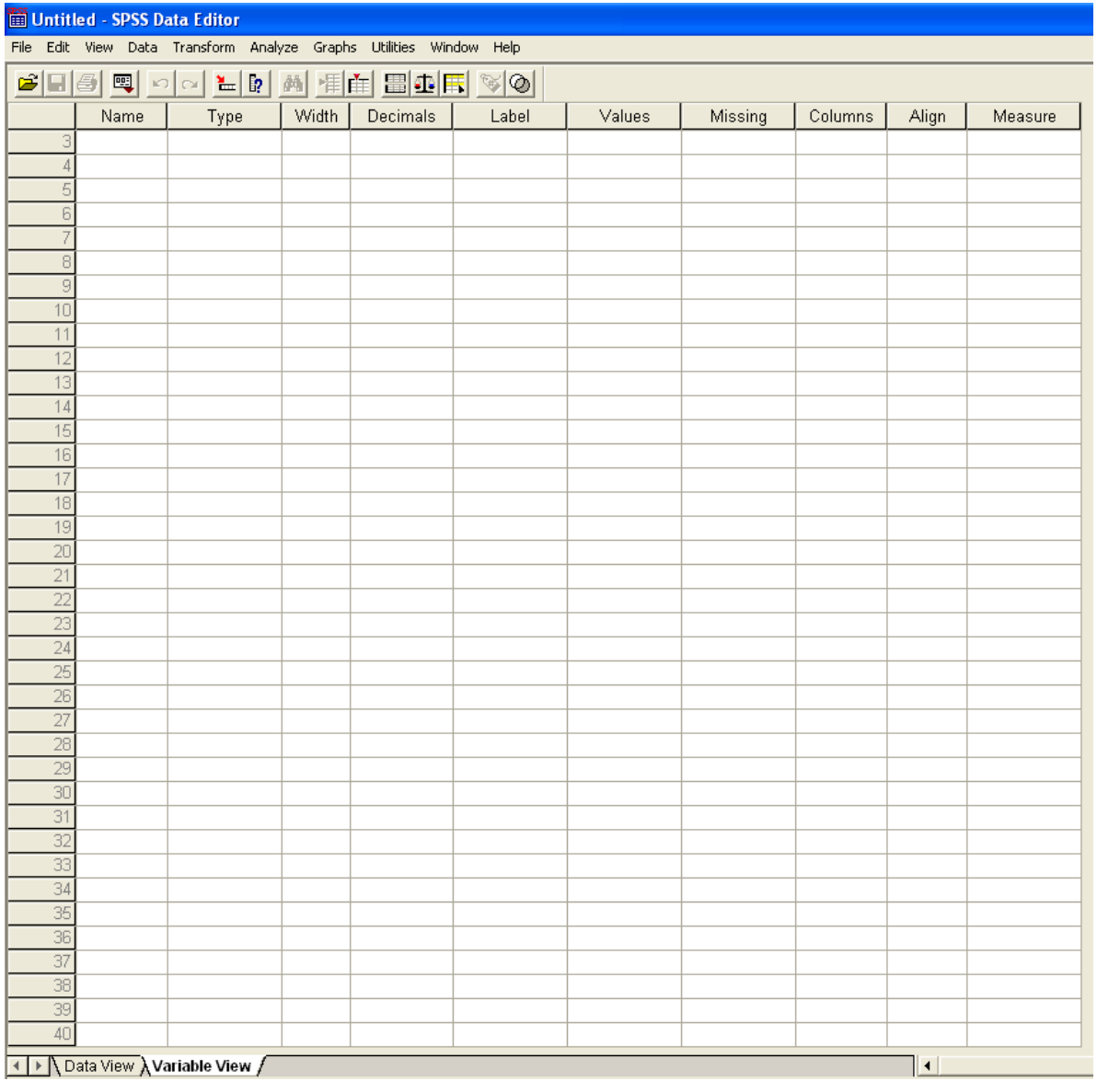
Trial version only for 14 days and limited licence. (about 2000 €)

Link:
http://www-01.ibm.com/software/analytics/spss/products/statistics/stats-standard/

SPSS


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
1 :
var
Data View
Variable View

SPSS


Untitled - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Window Help
Name Type Width Decimals Label Values Missing Columns Align Measure
Data View Variable View

R

Advantages

- Open source.
- Many statistical procedures.
- Availability on internet of many routines developed by users.

R

Advantages

- Open source.
- Many statistical procedures.
- Availability on internet of many routines developed by users.

Disadvantages

Not user-friendly Link: http://cran.r-project.org/

PSPP

Advantages

- It is a free replacement for the proprietary program SPSS.
- The copy of PSPP will not expire and there are no additional packages to purchase.
- It is designed to perform its analyses as fast as possible, regardless of the size of the input data.

PSPP

Advantages

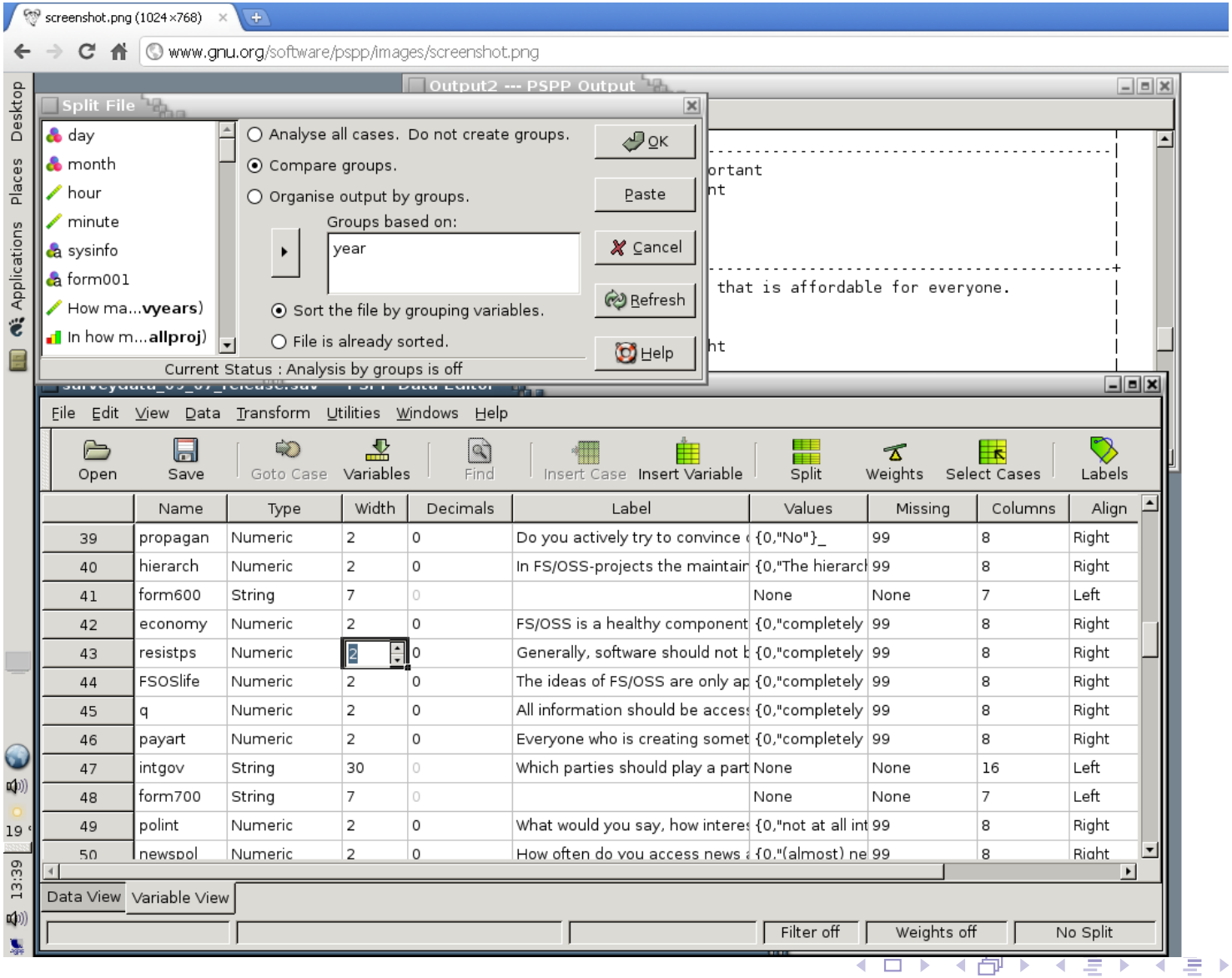
- It is a free replacement for the proprietary program SPSS.
- The copy of PSPP will not expire and there are no additional packages to purchase.
- It is designed to perform its analyses as fast as possible, regardless of the size of the input data.

Disadvantages

Sometimes there are problems to find the right mirror for the installation.

Link: http://www.gnu.org/software/pspp/

PSPP


screenshot.png (1024×768)
www.gnu.org/software/pspp/images/screenshot.png
Output2 --- PSPP Output
Split File
day
month
hour
minute
sysinfo
form001
How ma...vyears)
In how m...allproj)
Analyse all cases. Do not create groups.
Compare groups.
Organise output by groups.
Groups based on:
year
Sort the file by grouping variables.
File is already sorted.
OK
Paste
Cancel
Refresh
Help
Current Status : Analysis by groups is off
that is affordable for everyone.
File Edit View Data Transform Utilities Windows Help
Open Save Goto Case Variables Find Insert Case Insert Variable Split Weights Select Cases Labels
Name Type Width Decimals Label Values Missing Columns Align
39 propagan Numeric 2 0 Do you actively try to convince {0,"No"}_ 99 8 Right
40 hierarch Numeric 2 0 In FS/OSS-projects the maintair {0,"The hierarcl 99 8 Right
41 form600 String 7 0 None None 7 Left
42 economy Numeric 2 0 FS/OSS is a healthy component {0,"completely 99 8 Right
43 resistps Numeric 2 0 Generally, software should not b {0,"completely 99 8 Right
44 FSOSlife Numeric 2 0 The ideas of FS/OSS are only ap {0,"completely 99 8 Right
45 q Numeric 2 0 All information should be acces {0,"completely 99 8 Right
46 payart Numeric 2 0 Everyone who is creating somet {0,"completely 99 8 Right
47 intgov String 30 0 Which parties should play a part None None 16 Left
48 form700 String 7 0 None None 7 Left
49 polint Numeric 2 0 What would you say, how interes {0,"not at all int 99 8 Right
50 newspol Numeric 2 0 How often do you access news a {0,"(almost) ne 99 8 Right
Data View Variable View
Filter off Weights off No Split

Other softwares

Not only for multidimensional analysis...

- Matlab
- Stata
- Eviews
- Gauss

They are not open source and some of them perform only some techniques of multidimensional analysis.

Open source softwares

	Pr	Ano	Log	Prob	Glm	Nopar	Time	PCA	CCA	CA	Disc	Clus
Ade4		*						*	*	*	*	*
Dataplot	*	*				*	*	*	*		*	
Easyreg	*		*	*	*	*	*					
Gretl	*		*	*	*	*	*	*				
Instat +	*	*				*	*					
Macanova	*	*	*		*		*	*				*
Matrixer	*		*	*		*	*					
Microsiris	*	*	*					*				*
Tanagra		*				*		*		*	*	*
Vista		*						*		*		*
Winidams		*					*	*			*	*

Open source softwares

Thank you for your attention!