CoDaPack 2.0: a stand-alone, multi-platform compositional software

M. COMAS[2], and S. THIÓ-HENESTROSA[1]

[1] Universitat de Girona, Spain Santiago.thio@udg.edu
[2] Universitat de Girona, Spain

Abstract

Historically CoDaPack 3D has intended to be a software of Compositional Data with an easy and intuitive way of use. For this reason from the beginning it has been associated to Excel, a software known and used for many people. However, over the years different versions of Excel and Windows have been appeared and CoDaPack has had to be adapted to these new versions due to some incompatibilities.

For this reason, and also because of CoDaPack only works with Excel under windows, the Girona Compositional Data Group has decided to implement a new software with at least the same capabilities and the same profile of users but independent of any other software.

The graphical user interface has three different areas: The variables area, the data area and the results area which has a textual output window and independent graphical output. Also, because the new CoDaPack is being developed under Java code, the final software is going to work in any platform having a Java Virtual Machine: Windows, Linux and other Unix based systems.

1. Introduction

CoDaPack and its follower CoDaPack3D 1.x (Thió-Henestrosa and Martín-Fernández, 2005, 2006) have been the only user friendly software available on Compositional Data Analysis since its first apparition in the beginning of this century. This software ran as menus inside Excel under Windows operating system.

As a Whole package only exists another software called Compositions (Van den Boogaart and Tolosana, 2008) but its use is not intuitive as it is a set of R routines. There are also other packages that perform specific methodology: robCompositions (Templ, 2010) about robust estimation for compositional data, and Compos Analysis (P. G. Smith 2004 and 2010) an Add-In for use with Microsoft Excel but it is not freeware and it is centred on MANOVA analysis. Finally, there is also a lot of software that draws ternary diagrams.

One of the problems of the association of CoDaPack with Excel and Windows were that after every actualization of Excel and Windows versions CoDaPack had have to be adapted as some parts were not supported on new versions. Also, after one specific windows security update CoDaPack stopped working because some components were disabled.

Another of the limitations of CoDaPack was that it only ran under Windows operating system and all users of Unix based systems couldn't work with CoDaPack.

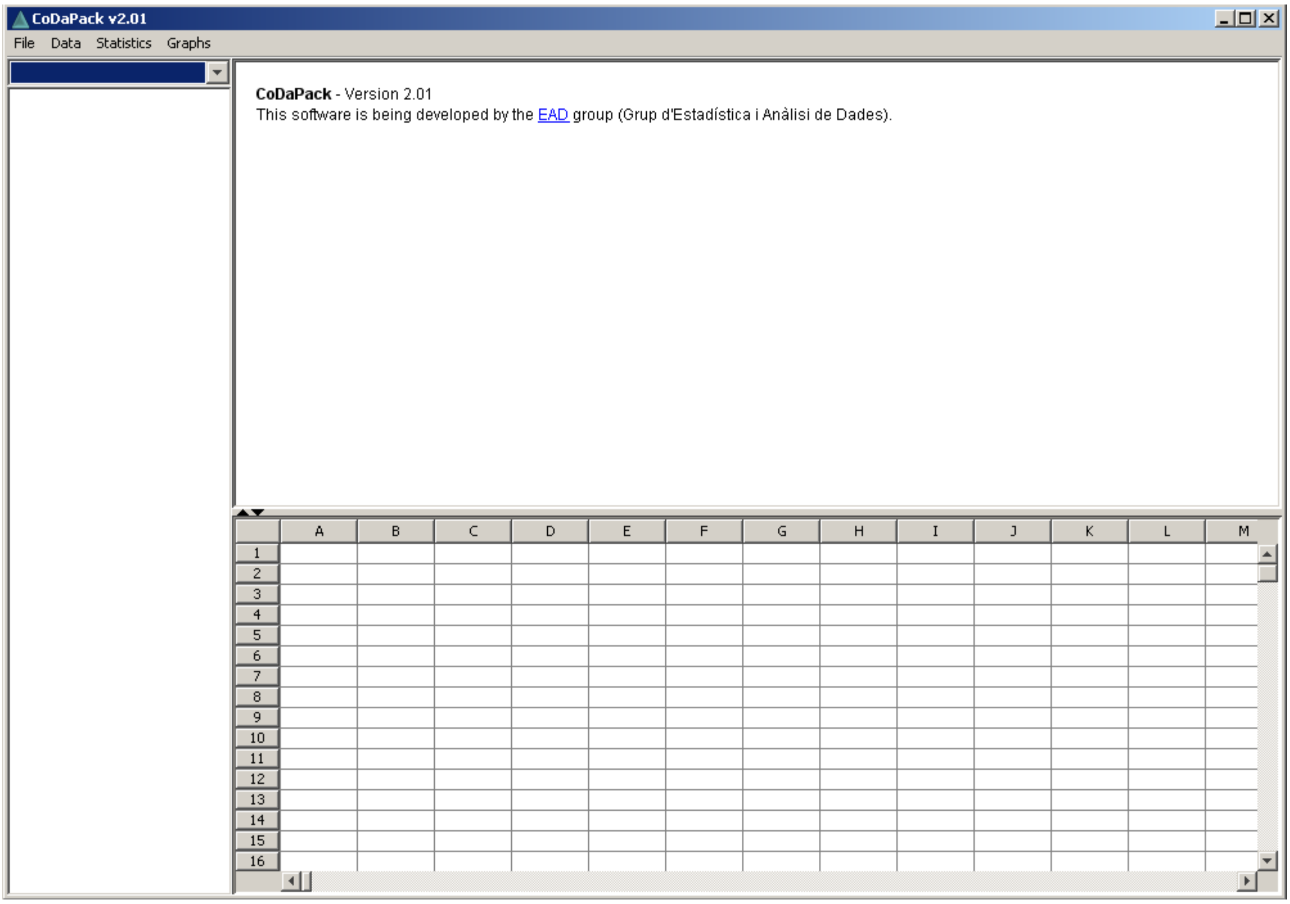
Mainly due to both reasons we decided to reprogram CoDaPack in order not to depend on other software and to build more stable software. This new package, now called CoDaPack 2.0, is being developed under Java code and the final software is going to work in any platform having a Java Virtual Machine: Windows, Linux and Unix based systems including MAC OS. Also it is intended to be easy to use and oriented to a non expert-computer users.

The Package is still in a developing phase and on CoDaWork'11 we present the first capabilities.

2. CoDaPack 2.0 structure

CoDaPack 2.0 main window (Figure 1) has four main areas: Menus, variables, alphanumerical results and data. Also graphical outputs appear in independent windows.

At this moment, the variables area is merely informative as only contains a list of the labels of the variables stored on the data part. In future version, direct access facilities are going to be available from this area. This window is useful if there is a lot of variables as it is easy to see their name and the order where they are placed.



CoDaPack - Version 2.01
This software is being developed by the EAD group (Grup d'Estadística i Anàlisi de Dades).

Figure 1. CoDaPack 2.0 structure: Menus on very top, variables on left side, alphanumerical results on right top and data on left bottom.

The data part contains variables organized in columns. The window has and Excel like appearance but it doesn't work as a spreadsheet as it can't be edited. The data area is headed by a grey row that contains the label of the variables. Also each row is headed by a grey number that indicates the position. Variables could be numerical and alphanumerical but only alphanumerical are treated as categorical data.

If a concrete data is a zero CoDaPack 2.0 distinguishes between non available data and non detected data. Non available data is a missing value or a structural zero and non detected data is a low value under detection limits. In this case it is useful to begin this data with a "<" sign that indicates a value inferior to this limit.

When CoDaPack 2.0 founds an observation with non detected or non available data, it is skipped and not used.

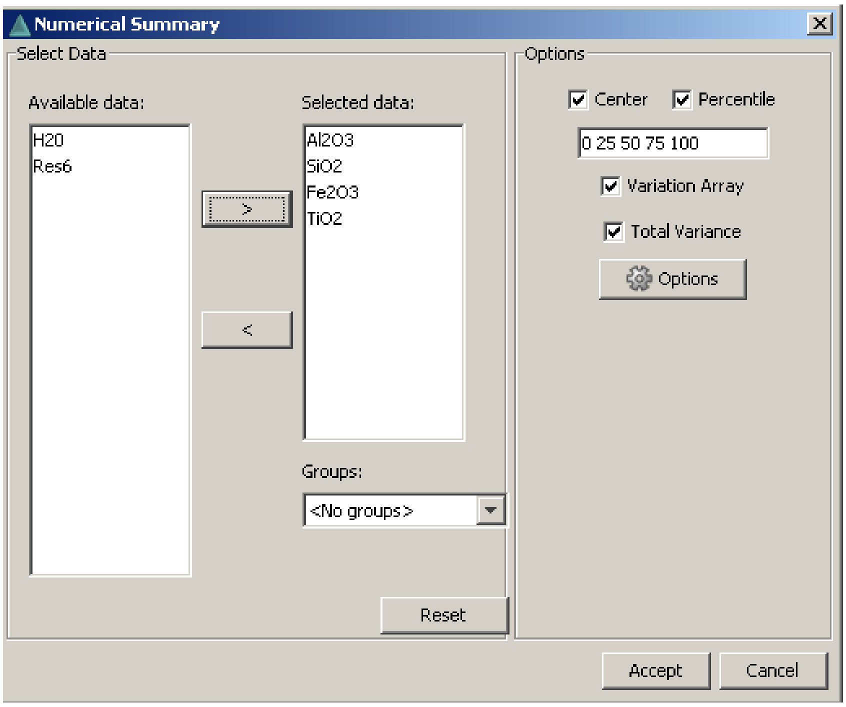
Finally, alphanumeric results appear on the right top of the window. On this part CoDaPack 2.0 traces all commands performed by the user and write the alphanumerical results.

3. CoDaPack 2.0 menus

The package has four main menus: File, Data, Statistics and Graphs.

All procedures of CoDaPack 2.0 open a form with similar structure: on top left there is the *Select Data* area, on top right the *Options* area and on bottom right *Accept* and *Cancel* buttons (Figure 2).

On *Select Data* area user should select the variables needed to perform the procedure. On this area there are five or six elements depending on the procedure. On left side there is a box with the available data, on right side a box the selected data and between these boxes two arrows to transport selected variables from one box to the other. Also it is possible to select or unselect a variable by double click on it. On the bottom of this area there is a *Reset* button that empties the selected data box. Finally, the sixth element, *Groups*, available only on procedures of Statistics and Graphs menus is used to select a variable that contains a partition of the observation. By default this option is <No groups>, if a variable is selected the results will be done for every group of the partition. Depending on the procedure it is only possible to select a limited number of variables. For example Ternary Diagrams only allows selecting three or four variables.



Numerical Summary

Select Data

Available data:
H20
Res6

Selected data:
Al2O3
SiO2
Fe2O3
TiO2

Groups: <No groups>

Options

☑ Center ☑ Percentile

0 25 50 75 100

☑ Variation Array

☑ Total Variance

Figure 2. Numerical summary Form.

The *Options* area is only used on procedures that need extra information to be executed.

3.1 File menu

File Menu (Figure 3) manages with data files. CoDaPack 2.0 stores the data in workspaces. CoDaPack 2.0 workspaces are ASCII files that contain mainly the data and its labels.

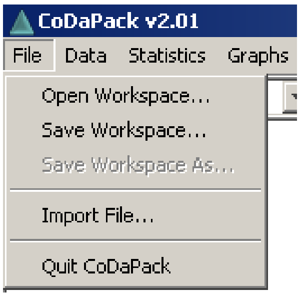
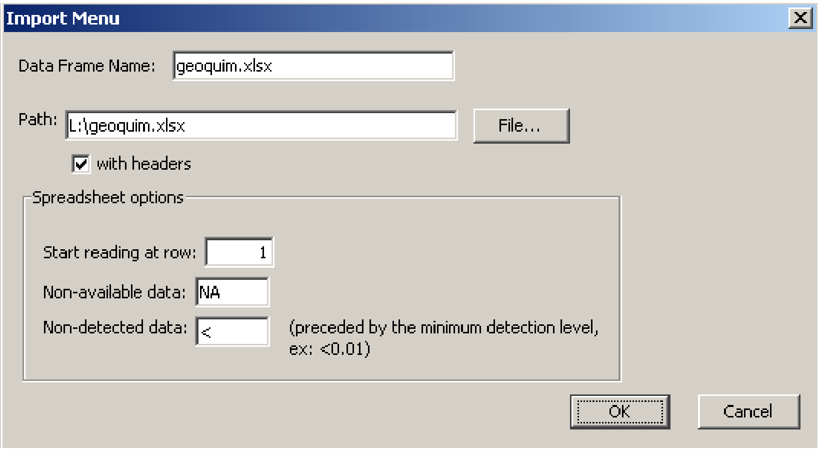


Figure 3. File Menu and its submenus.

Also, it is possible to import data from external applications. At this time it is only possible with Excel file. When a data set is imported from Excel a form appears (Figure 4) asking for the name file and where it can be found. Also this menu requires the first row of the spreadsheet to be read and if the first row of the data contains the labels and how to identify non available data and not detected data.



Import Menu ☒

Data Frame Name: geoquim.xlsx

Path: L:\geoquim.xlsx File...

☑ with headers

Spreadsheet options

Start reading at row: 1

Non-available data: NA

Non-detected data: < (preceded by the minimum detection level, ex: <0.01)

OK Cancel

Figure 4. Import menu form.

If the Excel file has more than one sheet then CoDaPack 2.0 asks for which of the sheets the user wants to use.

3.2 Data menu

Data Menu (Figure 5) merges two menus of ancient CoDaPack: Transformations and Operations and also adds some other capabilities.

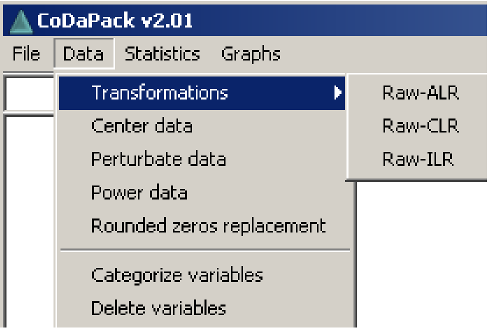
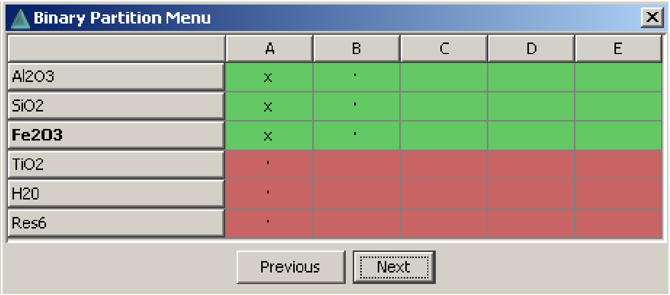


Figure 5. Data menu and submenus.

3.2.1 Transformation

The first option of this menu, transformations, concerns to the three transformations between real space to the simplex and their inverses. To perform the ILR transformation a sequential binary partition is needed to be introduced by the user by means of a form (Figure 6).

On this form green indicates which parts could be selected at this step of the partition while red indicates which parts not.



Binary Partition Menu ☒

	A	B	C	D	E
Al2O3	x	·			
SiO2	x	·			
Fe2O3	x	·			
TiO2	·				
H2O	·				
Res6	·				

Previous Next

Figure 6. Form to enter the partition.

3.2.2 Operations inside simplex: Center, Perturbation, Power and Rounded zero Replacement

Data menu also has operations inside the simplex, that is, operations which result are new elements in the simplex that appear as new variables at the data area of CoDaPack 2.0.

Center data centers a set of parts selected. Also, optionally the center of the parts selected is written on the results area.

Perturbate data performs a perturbation operation between selected parts and a perturbation vector given on options area of the form.

Power data performs a power transformation of selected data with a constant placed by the user on options area.

Rounded Zero Replacement substitutes the non detected data by its detection limit value following Martín-Fernández et al. (2003) methodology. Non available data are not substituted

3.2.3 Categorize and delete variables

The last two procedures of Data menu concern to variables management. It is possible to categorize a variable just converting it as string. Also it is possible to delete a set of variables.

3.3 Statistics menu

At this time Statistics Menu (Figure 7) only produces a Numerical summary of selected data.

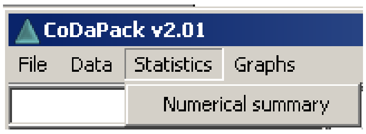
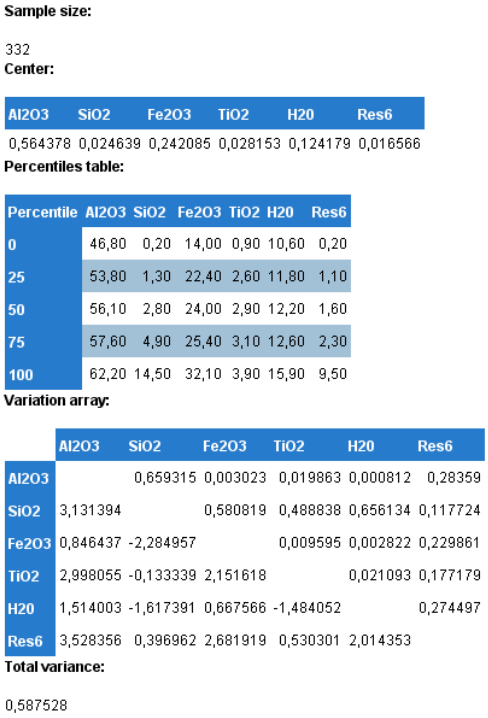


Figure 7. Statistics Menu.



Sample size:

332

Center:

Al2O3	SiO2	Fe2O3	TiO2	H20	Res6
0,564378	0,024639	0,242085	0,028153	0,124179	0,016566

Percentiles table:

Percentile	Al2O3	SiO2	Fe2O3	TiO2	H20	Res6
0	46,80	0,20	14,00	0,90	10,60	0,20
25	53,80	1,30	22,40	2,60	11,80	1,10
50	56,10	2,80	24,00	2,90	12,20	1,60
75	57,60	4,90	25,40	3,10	12,60	2,30
100	62,20	14,50	32,10	3,90	15,90	9,50

Variation array:

	Al2O3	SiO2	Fe2O3	TiO2	H20	Res6
Al2O3		0,659315	0,003023	0,019863	0,000812	0,28359
SiO2	3,131394		0,580819	0,488838	0,656134	0,117724
Fe2O3	0,846437	-2,284957		0,009595	0,002822	0,229861
TiO2	2,998055	-0,133339	2,151618		0,021093	0,177179
H20	1,514003	-1,617391	0,667566	-1,484052		0,274497
Res6	3,528356	0,396962	2,681919	0,530301	2,014353	

Total variance:

0,587528

Figure 8. Output for Numerical summary.

Given a set of selected parts this procedure returns alphanumerical results. If selected by the users the output consists on (Figure 9) sample size, percentiles 0, 25, 50, 75 and 100, variation array and Total variance.

3.4 Graphs menu

There are seven types of graphs (Figure 9). One of the differences of the 2.0 version of CoDaPack3D is that all 2D and 3D graphics uses the same menu and depending on the number of variables they appear as 2D or 3D, except of Dendrograms of balances that are only two dimensional.

Once a graph is drawn it is possible to zoom it by means of the mouse or a button. Also with the mouse it is possible to move the graph and, on 3D graphs, to rotate it.

Also it is possible to save snapshots of the graph with Jpeg, Ps, Png and Bitmap formats.

All graphs allow drawing observations by means of a group codification. Only categorical variables could be used to define groups.

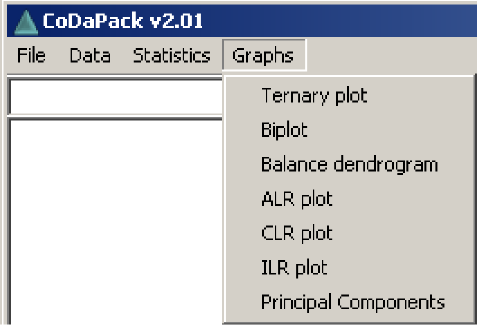


Figure 8. Graphs Menu.

3.4.1 Ternary plots

Ternary diagram (Figure 10) is displayed by default as not centred and without grid (only 2D) but both options could be activated or deactivated interactively while graph is open. Also it is possible to change the position of axes.

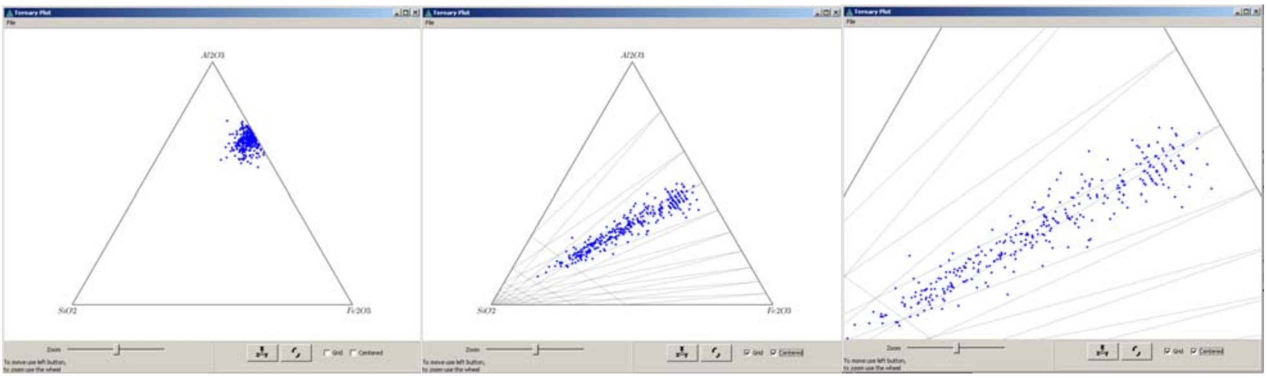


Figure 10. Ternary diagram. a) Default display, b) With centring and grid, and c) After some zoom.

If user selects four parts a tetrahedron is displayed. Figure 11 shows three snapshots of a 3D ternary diagram.

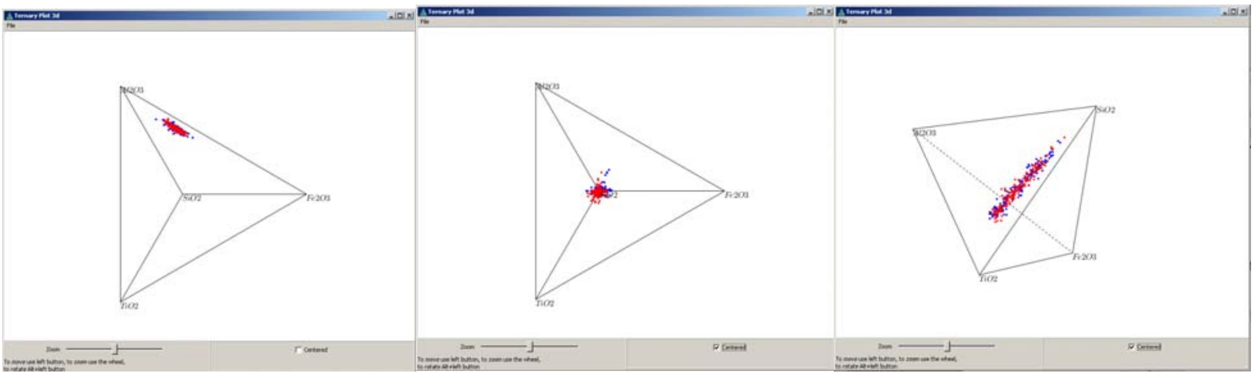

Ternary Plot 3d
File
Al2O3
SiO2
Fe2O3
TiO2
Zoom
Centered
To move use left button, to zoom use the wheel,
to rotate Alt+left button
Ternary Plot 3d
File
Al2O3
Fe2O3
TiO2
Zoom
☑ Centered
To move use left button, to zoom use the wheel,
to rotate Alt+left button
Ternary Plot 3d
File
SiO2
Al2O3
Fe2O3
TiO2
Zoom
☑ Centered
To move use left button, to zoom use the wheel,
to rotate Alt+left button

Figure 11. 3D Ternary diagram. a) Default display, b) With centring, and c) After some rotation.

3.4.2 Biplots

By default CoDaPack 2.0 display 3 dimensional biplots unless user only selected three variables. Once the biplot is drawn the user could choose which biplot wants, Covariance, Symmetric scaling and Form (Figure 12).

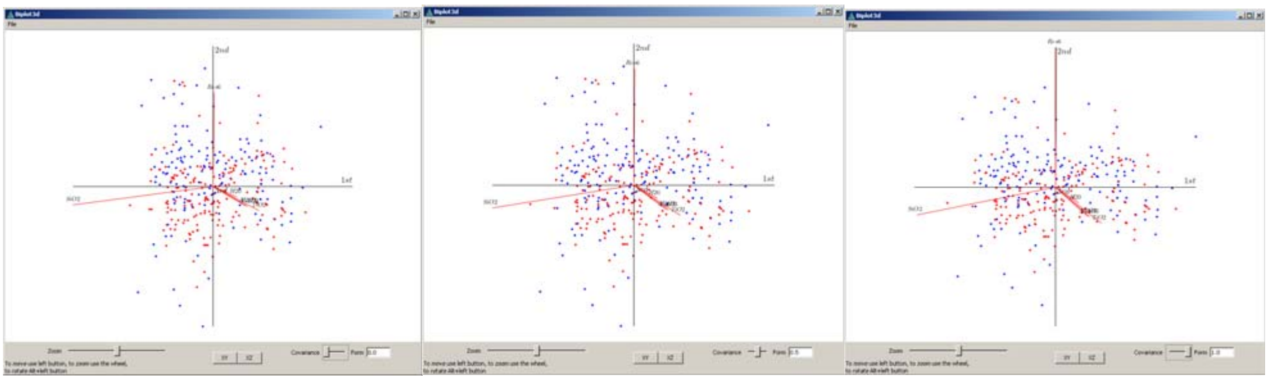

Biplot 3d
File
2nd
1st
SiO2
Zoom
XY
XZ
Covariance
Form 0.0
To move use left button, to zoom use the wheel,
to rotate Alt+left button
Biplot 3d
File
2nd
1st
SiO2
Zoom
XY
XZ
Covariance
Form 0.5
To move use left button, to zoom use the wheel,
to rotate Alt+left button
Biplot 3d
File
2nd
1st
SiO2
Zoom
XY
XZ
Covariance
Form 1.0
To move use left button, to zoom use the wheel,
to rotate Alt+left button

Figure 12. Biplot a) Covariance, b) Symmetric scaling, and c) Form.

Also it is possible to show the display of alternatively first and second or first and third axis just clicking on a button and it is possible to move, rotate and zoom the display (Figure 13).

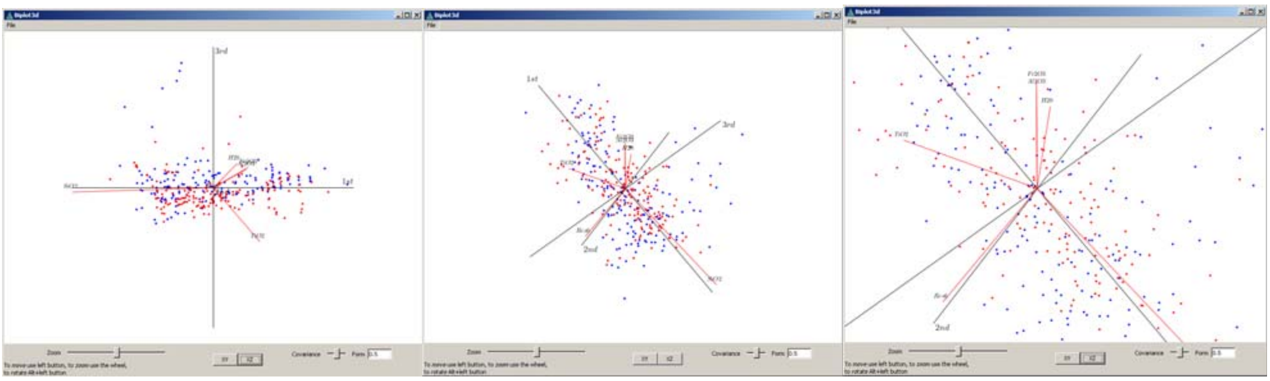
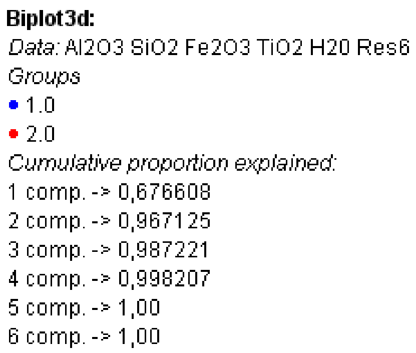

Biplot 3d
File
3rd
1st
SiO2
TiO2
Zoom
XY
XZ
Covariance
Form 0.5
To move use left button, to zoom use the wheel,
to rotate Alt+left button
Biplot 3d
File
1st
3rd
2nd
SiO2
Zoom
XY
XZ
Covariance
Form 0.5
To move use left button, to zoom use the wheel,
to rotate Alt+left button
Biplot 3d
File
Fe2O3
Al2O3
TiO2
2nd
Zoom
XY
XZ
Covariance
Form 0.5
To move use left button, to zoom use the wheel,
to rotate Alt+left button

Figure 13. Biplot a) display first and third axis, b) after some rotation, and c) after some zoom.

At this time numerical output consists on just the cumulative proportion explained (Figure 14).



Biplot3d:

Data: Al2O3 SiO2 Fe2O3 TiO2 H20 Res6

Groups

● 1.0

● 2.0

Cumulative proportion explained:

1 comp. -> 0,676608

2 comp. -> 0,967125

3 comp. -> 0,987221

4 comp. -> 0,998207

5 comp. -> 1,00

6 comp. -> 1,00

Figure 14. Numerical output for Biplot.

3.4.3 Balance dendrograms

Balance dendrogram routine asks for the user to define a sequential binary partition following the same steps as ILR transformation.

Figure 15 shows the display of a dendrogram obtained from a data set with two groups. As a difference of CoDaPack earlier versions now, for each balance, a boxplot of each group is displayed and the balances are connected on the median of the whole population (Figure 15).

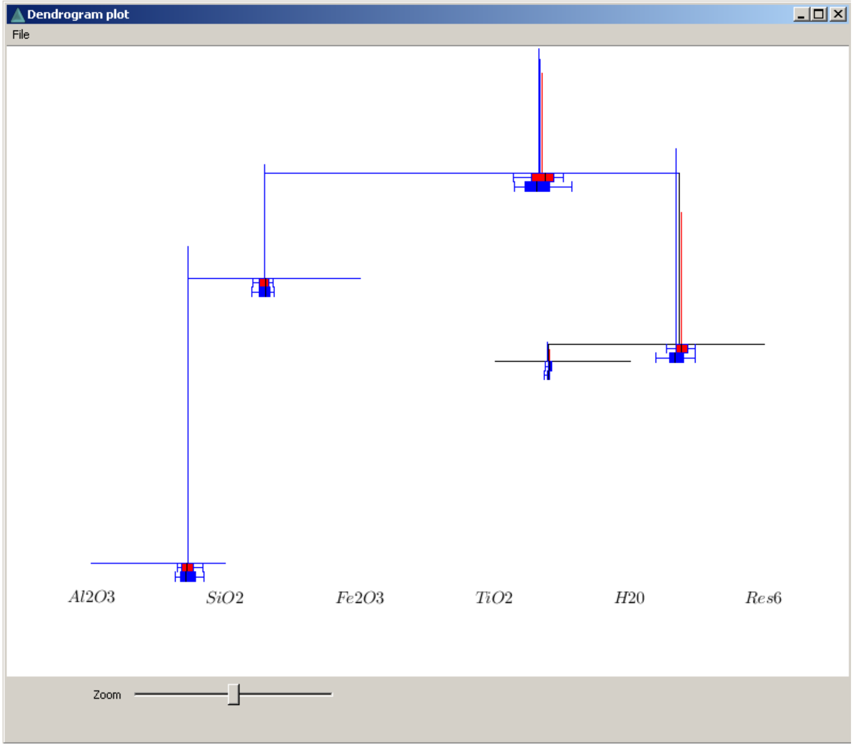


Figure 15. Graphical output of Balance dendrogram routine.

As numerical output this routine puts on data area the ILR transformation made with the data.

3.4.4 ALR, CLR and ILR Plots

CoDaPack 2.0 also allows plotting ALR, CLR and ILR transformations. Depending on the variables used the graphs will be two or three dimensional.

On 3D graphs is it possible to se a view of XY and XZ by means of a button. Also it is possible to rotate and zoom (Figure 16).

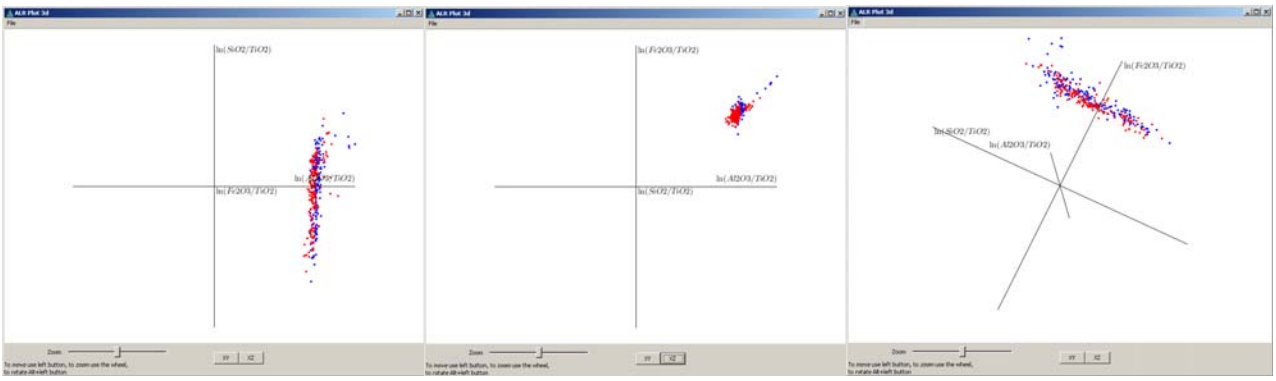


Figure 16. ALR Plot a) XY view, b) XZ view, and c) after some rotation.

3.4.5 Principal components

Principal components routine has the same capabilities that ternary diagram routine adding the drawn of the principal components (Figure 17)

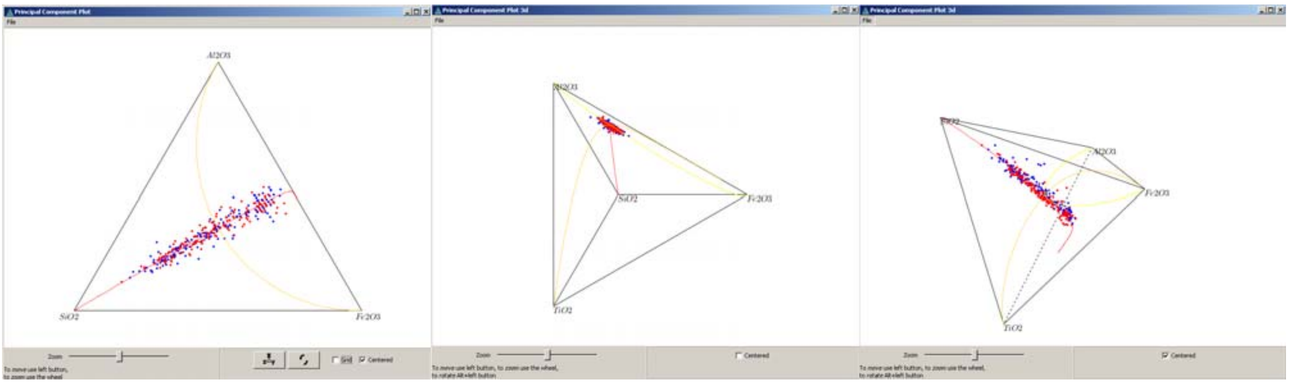
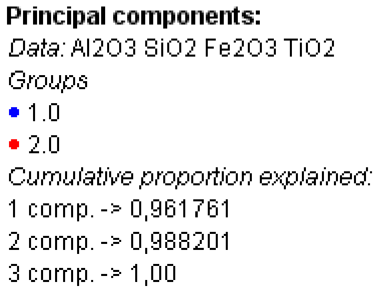


Figure 17. Principal components a) Ternary, b) Tetrahedron, and c) Tetrahedron with centred data set ans some rotation.

Also, as a numerical output this routine returns the cumulative proportion explained of these principal components (Figure 18).



```
Principal components:
Data: Al2O3 SiO2 Fe2O3 TiO2
Groups
● 1.0
● 2.0
Cumulative proportion explained:
1 comp. -> 0,961761
2 comp. -> 0,988201
3 comp. -> 1,00
```

Figure 18. Numerical output for Principal components.

4 Conclusion

CoDaPack 2.0 is just at its beginnings. It will be presented for the first time at CoDaWork'11. But there is still a lot of work to do adding more routines and adding more capabilities on the existent.

Respect to versions 1.x CoDaPack 2.0 takes a lot of advantages, the more important is that now is independent of other packages like Excel. Also it is available on multiple platforms as it is available on Unix based operating systems like MAC OS.

Also it is faster executing and graphical output has been improved as they are interactive.

References

Martín-Fernández, J.A., C. Barceló-Vidal and V. Pawlowsky-Glahn, (2003). Dealing with zeros and missing values in compositional data sets. *Mathematical Geology* 35, 3, 253-278.

Smith, P. G., (2004). Automated log-ratio analysis of compositional data: software suited to analysis of habitat preference from radiotracking data. *Bat Research News* 45: 16.

Smith, P. G., (2010). *Compos Analysis version 6.3 user's guide.* Version 6.3. Smith Ecology Ltd., Ty Major, Forest Coal Pit, Abergavenny, NP7 7LH, UK. i + 22 pp. http://www.smithecology.com/software.htm

M. Templ, Hron K., Filzmoser P., (2010) *robCompositions: Robust Estimation for Compositional Data. Manual and package, version 1.3.3*. http://cran.r-project.org/package=robCompositions.

Thió-Henestrosa, S. and J.A. Martín-Fernández, (2005). Dealing with compositional data: the freeware CoDaPack. M*athematical Geology* 37 (7), 773–793.

Thió-Henestrosa, S., and J.A. Martín-Fernández (2006). Detailed guide to CoDaPack: a freeware compositional software. In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (Eds.), *Compositional Data Analysis in the Geosciences: From Theory to Practice*, vol. 264. Geological Society, London, pp. 101–118 (Special Publications).

Van den Boogaart, K. and R. Tolosana-Delgado (2008). "compositions": A unified R package to analyze compositional data. *Computers and Geosciences* 34 (2009), 320–338.