

CoDaPack. AN EXCEL AND VISUAL BASIC BASED SOFTWARE OF COMPOSITIONAL DATA ANALYSIS. CURRENT VERSION AND DISCUSSION FOR UPCOMING VERSIONS

S. Thió-Henestrosa¹, C. Barceló-Vidal², J. A. Martín-Fernández², and V. Pawlowsky-Glahn²

¹Universitat de Girona, Spain; santiago.thio@udg.es

²Universitat de Girona, Spain

1 Introduction

In the eighties, John Aitchison (1986) developed a new methodological approach for the statistical analysis of compositional data. This new methodology was implemented in Basic routines grouped under the name CODA and later NEWCODA in Matlab (Aitchison, 1997). After that, several other authors have published extensions to this methodology: Martín-Fernández and others (2000), Barceló-Vidal and others (2001), Pawlowsky-Glahn and Egozcue (2001, 2002) and Egozcue and others (2003).

This methodology is not straightforward to use, neither with the original CODA or NEWCODA, nor with standard statistical packages. For this reason the Girona Compositional Data Analysis Group has developed a new freeware, named CoDaPack, based on CODA routines, which includes at this moment some basic statistical methods suitable for compositional data. It is developed in VisualBasic associated to Excel and it is oriented towards users with minimum knowledge on computers, with the aim to be simple and easy to use. Using menus, one can execute macros to return the numerical results on the same sheet and graphical outputs that appear in independent windows inside Excel.

In the present version there exist 5 menus with a total of 23 macros. The first menu, Transformations, performs several transformations of data from real space to the simplex and viceversa, that is, 1) Unconstrain/Basis, 2) Raw-ALR (additive log-ratio transformation, alr, and its inverse transformation, the generalised additive logistic transformation, agl), 3) Raw-CLR (centred log-ratio transformation, clr, and its inverse) and 4) Raw-ILR (the isometric log-ratio transformation, ilr, and its inverse transformation).

The second menu, Operations, performs the following operations inside the simplex 1) Perturbation, 2) Power transformation, 3) Centering, 4) Standardisation, 5) Amalgamation, 6) Subcomposition/Closure and 7) Rounded Zero Replacement.

The third menu, Graphs, performs two dimensional graphs like ternary diagrams, plots of alr or clr transformed data sets, biplots, principal components plot, additive logistic normal predictive regions and confidence regions, the three last ones in the ternary diagram. In all of these graphs the user can customize the appearance of the graph and, in some cases, the user can mark the observations in the graph according to a previous classification.

The fourth menu, Descriptive Statistics, returns characteristic values for a data set, like 1) Center, 2) Variation matrix, and 3) Total variance.

And, finally, the fifth menu, Preferences, allows the user to customize the application.

The web site

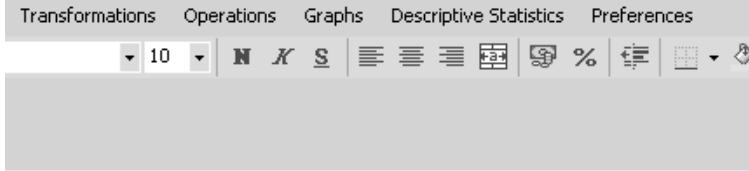
[http://ima.udg.es/~thio/#Compositional Data Package](http://ima.udg.es/~thio/#Compositional%20Data%20Package)

contains this freeware package and to install it the user only needs to have Excel installed on his computer.

To illustrate the use of the program an example with real data of male and female physical activity is presented. Finally this paper includes a discussion section with open questions of which are additional features expected in a new version.

2 CoDaPack structure

Once installed, to use CoDaPack, one has to access Excel and introduce the data in a standard spreadsheet. The observations must be in rows and the variables in columns, and the first row of each column can be used to label the variables or it has to rest blank (Fig. 1).



The screenshot shows the main menu of the software with options: Transformations, Operations, Graphs, Descriptive Statistics, and Preferences. Below the menu is a toolbar with various icons for data manipulation and analysis.

G	H	I	J	K
Activity	Strong Interest	Undecided	Little Interest	Sex
Archery	163	121	104	1
Badminton	31	153	204	1
Basketball	267	53	68	1
Bicycling	199	109	79	1
Bowling	167	119	102	1
Floor Hockey	58	172	158	1
Flag Football	146	118	124	1

Figure 1: An example of the spreadsheet with the main menus and the data.

Using menus, one can execute macros that return numerical results to the same sheet and graphical outputs that appear in independent windows inside Excel. Each macro asks the user which are the data and where to put the results (if there are numerical results). Some of the macros, specially those with graphical output, have an option button to modify the default values (Fig. 2).

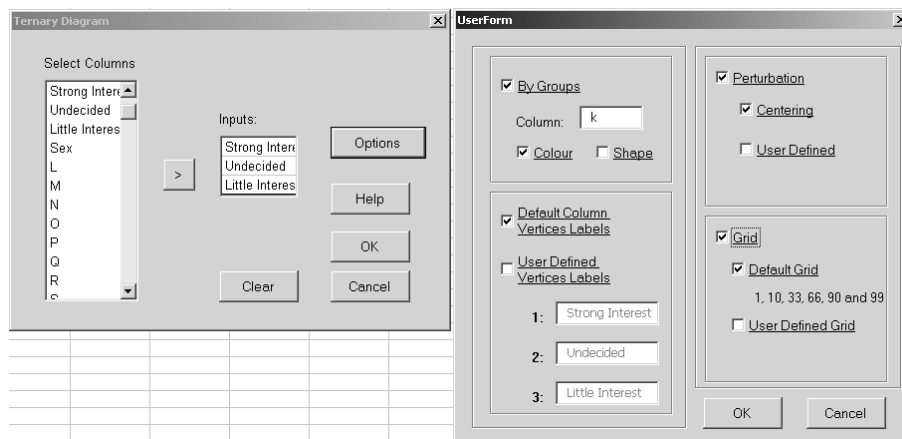


Figure 2: Dialog screen for ternary diagram menu.

3 Features of CoDaPack - Description of menus and macros

This section describes all macros of CoDaPack in the same order as they appear on the five menus. For the general notation, we follow Aitchison and others (2002), who represent a generic D -part composition as a row vector $\mathbf{x} = [x_1, \dots, x_D]$, where the x_i ($i = 1, \dots, D$), and a data set consisting of N D -part compositions, $\mathbf{x}_1, \dots, \mathbf{x}_N$, as an $N \times D$ matrix $\mathbf{X} = [\mathbf{x}_1; \dots; \mathbf{x}_N]$.

Transformations: This menu performs several transformations of data from real space to the simplex and viceversa. Consider a set of compositional observations \mathbf{X} and a vector \mathbf{p} associated to their size as defined in Aitchison (1986).

1) Unconstrain/Basis: this routine returns the data set unconstrained, that is, for each observation \mathbf{x}_i , it returns $\mathbf{y}_i = [x_{i1}p_i, \dots, x_{iD}p_i]$, where p_i is the size of each observation.

2) Raw-ALR: this routine performs the additive log-ratio transformation (alr) and its inverse transformation, that is, the generalised additive logistic transformation (agl). Division in the alr transformation is performed with the last component according to the sequence selected by the user.

$$\mathbf{y} = \text{alr}(\mathbf{x}) = \left[\ln \frac{x_1}{x_D}, \dots, \ln \frac{x_{D-1}}{x_D} \right], \text{ where } \mathbf{y} \in \mathfrak{R}^{D-1},$$

and

$$\mathbf{x} = \text{agl}(\mathbf{y}) = \left[\frac{\exp(y_1)}{\sum_{i=1}^{D-1} (\exp(y_i) + 1)}, \dots, \frac{\exp(y_{D-1})}{\sum_{i=1}^{D-1} (\exp(y_i) + 1)}, 1 - x_1 - \dots - x_{D-1} \right].$$

3) Raw-CLR: this routine performs the centred log-ratio transformation (clr) and its inverse (clr^{-1}):

$$\mathbf{y} = \text{clr}(\mathbf{x}) = \ln \frac{\mathbf{x}}{g_D(\mathbf{x})},$$

where $\mathbf{y} \in \mathfrak{R}^D$, and $g_D(\mathbf{x})$ is the geometric mean $\left(\prod_{k=1}^D x_k \right)^{1/D}$ of \mathbf{x} , and

$$\mathbf{x} = \text{clr}^{-1}(\mathbf{y}) = \left[\frac{\exp(y_1)}{\sum_{i=1}^D \exp(y_i)}, \dots, \frac{\exp(y_D)}{\sum_{i=1}^D \exp(y_i)} \right].$$

4) Raw-ILR: this routine performs the isometric log-ratio transformation (ilr), according to the sequence selected by the user, as well as its inverse transformation (ilr^{-1}):

$$\mathbf{y} = \text{ilr}(\mathbf{x}) = (y_1, \dots, y_{D-1}) \in \mathfrak{R}^{D-1},$$

where

$$y_k = \frac{1}{\sqrt{k(k+1)}} \ln \left(\frac{\prod_{j=1}^k x_j}{(x_{k+1})^k} \right) \quad (k = 1, \dots, D-1),$$

and

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{y}) = \left[\left(1 + \frac{\sum_{i=0, i \neq 1}^D f(i)}{f(0)} \right)^{-1}, \dots, \left(1 + \frac{\sum_{i=0, i \neq D}^D f(i)}{f(D-1)} \right)^{-1} \right],$$

where

$$f(j) = \left(\frac{1}{f(j-1)} \exp(\sqrt{j(j+1)}y_j) \right)^{-1/j} \text{ and } f(0) = 1.$$

Operations: This menu performs the following operations inside the simplex

1) Perturbation: returns a D -composition $\mathbf{y} = \mathbf{p} \oplus \mathbf{x} = \mathcal{C}(p_1x_1, \dots, p_Dx_D)$, where \mathcal{C} is the closure operation $\mathcal{C}(x_1, \dots, x_D) = (\frac{x_1}{\sum_{i=1}^D x_i}, \dots, \frac{x_D}{\sum_{i=1}^D x_i})$, and \mathbf{p} is a fixed D -composition.

2) Power transformation: for $a \in \mathfrak{R}$, the power transformation return $a \otimes \mathbf{x} = \mathcal{C}(x_1^a, \dots, x_D^a)$.

3) Centering: this routine centres the data set, that is, it returns the data set \mathbf{Y} formed by the D -compositions $\mathbf{y} = g_n(\mathbf{X})^{-1} \oplus \mathbf{x}$, where $g_n(\mathbf{X}) = \left[(\prod_{i=1}^n x_{i1})^{1/n}, \dots, (\prod_{i=1}^n x_{iD})^{1/n} \right]$ is the vector of geometric means of the data set \mathbf{X} . Thus, the centre of the set \mathbf{Y} is \mathbf{e} , the baricentre of the simplex.

4) Standardisation: this routine returns a sample of D -compositions \mathbf{y} , centred at \mathbf{e} and with unit total variance.

5) Amalgamation: the result of the amalgamation of some of the components of a D -composition selected by the user is the sum of those components.

6) Subcomposition/Closure: this routine closes the data, that is, returns $\mathbf{y} = \mathcal{C}(\mathbf{x})$. If we select S variables ($S < D$) a subcomposition with S -parts is obtained.

7) Rounded Zero Replacement: it consists in the substitution of an observation \mathbf{x} , with zeros in some parts, by an observation \mathbf{y} using the expression:

$$y_k = \begin{cases} \delta_k, & \text{if } x_k = 0 \\ x_k (1 - \sum_{x_i=0} \delta_i), & \text{if } x_k > 0 \end{cases} ,$$

where δ_k is the replacement value for the k -th component defined by the user.

Graphs: This menu performs two dimensional graphs in separate windows. In all of this graphs the user can customize the appearance of the graph and, in some cases, the user can mark the observations in the graph according to a previous classification.

1) Ternary Diagram: Displays the Ternary Diagram of 3 selected columns.

There are four options to modify the appearance of the graph: 1) to differentiate, by color or by shape, each point depending on a previous classification, 2) to label the vertices of the triangle (the default labels are the column names), 3) to perturb the data with the inverse of the centre (centering) or with a given vector, and 4) to display a grid of values. The default values of the grid are 1, 10, 33, 66, 90 and 99 but the user can define other values in a column.

2) ALR Plot: Displays a plot according to the additive log-ratio transformation of 3 selected columns.

There are two options to modify the appearance of the graph: 1) to differentiate, by color or by shape, each point depending on a previous classification, and 2) to label the axis (the default labels are $\log(x_1/x_3)$ and $\log(x_2/x_3)$).

3) CLR Plot: Displays a plot according to the centred log-ratio transformation of 3 selected columns.

There are two options to modify the appearance of the graph: 1) to differentiate, by color or by shape, each point depending on a previous classification, and 2) to label the axis.

4) Biplot: Performs a biplot of selected columns.

There are six options to modify the appearance of the graph: 1) axes name: the user can indicate a column with the labels of the axes, 2) to differentiate, by color or by shape, each point depending on a previous classification, 3) the user can choose the factor plane indicating which components to display, 4) to label the observations (the default is no label), 5) to display or not the observations (the default is yes), and 6) to display with a different mark the observations that are outliers (the default is not).

5) Principal Components: Performs a Principal Components Analysis of 3 selected columns and displays the result in a ternary diagram.

There are two options to modify the appearance of the graph: 1) to differentiate, by color or by shape, each point depending on a previous classification, and 2) to label the vertices of the triangle (the default labels are the column names).

The display includes the cumulative proportion explained and the Principal Components.

6) ALN Predictive Region: Performs the Additive Logistic Normal Predictive Region of the selected columns and displays the result in a ternary diagram.

There are two options to modify the appearance of the graph: 1) to label the vertices of the triangle (the default labels are the column names), and 2) to choose the default predictive levels (the default levels are 0.90, 0.95 and 0.99).

7) ALN Confidence Region: Performs the Additive Logistic Normal Confidence Region of the selected columns and displays the result in a ternary diagram.

There are three options to modify the appearance of the graph: 1) to perform an ALN confidence region for each group defined by a column. 2) to label the vertices of the triangle (the default labels are the column names) and 3) to define the confidence level (the default is 0.95).

Descriptive Statistics: This menu returns characteristic values for a data set, like

1) Center: returns the center of the data, that is, the compositional geometric mean of the data set \mathbf{X} .

2) Variation matrix: returns a matrix with the variance of the logarithms of the quotients of all the parts. That is, the ij -th component of the variation matrix is $\text{var}(\ln(x_i/x_j))$, where $i \neq j$.

3) Total variance: returns the sum of all the elements in the variation matrix divided by $2D$, that is $\frac{\sum_{i=1}^{D-1} \sum_{j=i+1}^D \text{var}(\ln(x_i/x_j))}{D}$.

Preferences: This menu allows the user to indicate:

1) Screen size: it is used to indicate the size of the screen in pixels, in order to perform the graphs with the right size according to the screen. The graphical outputs are customised for a default screen size of 1152×864 pixels, but if the user has a different configuration, the size of the graphs can be adapted.

2) Sum constraint: to indicate the sum constraint used. The default value is 1.

4 Example

To illustrate with an example some of the features of CoDaPack we use a data set (Greenwood and others, 2001) consisting of 3-part samples of interest in 23 physical activities. The available data set is divided in two groups, 388 males and 363 females, and there is counted for each of 23 physical activities how many people of each group declare "strong interest", "little interest" or "undecided".

First of all, we can visualise (Fig. 3) the data in a ternary diagram differentiating the sex by different colors, and also displaying a grid of values. The default values of the grid are 1, 10, 33, 66, 90 and 99 but the user can define other values in a column. Can also visualise with different colours the activities in order to see the differences between sex (Fig. 4).

In order to see if the mean of two groups are different we can plot confidence regions under normality assumptions. Figure 5 shows that both 95% confidence regions for the respective centers

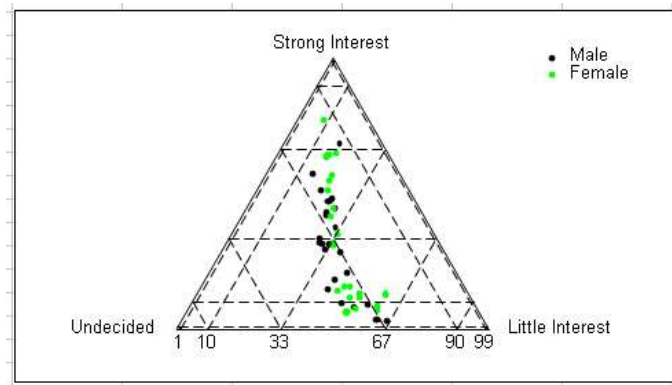


Figure 3: Ternary diagram with a grid.

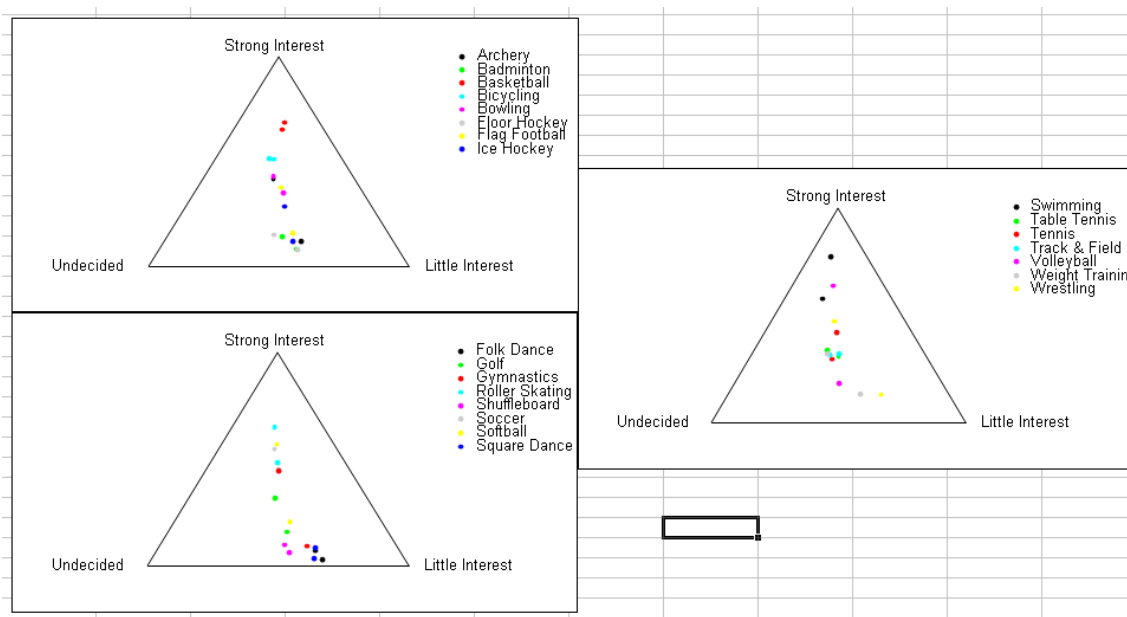


Figure 4: Ternary diagram by activity.

are disjoint, a way to assess visually that the two groups can be considered as different.

It is also possible to perform a principal component analysis in order to describe the data. Figure 6 shows that the data can be well fitted by the first component, as 98 percent of the inertia of the data is explained. Also, this first axis opposes the activities depending on its Strong Interest and its Little Interest.

5 Discussion

This package is in its first version and only the basic methodology has been implemented.

Now we are planning to program a new version. Because there is still a lot of work to be done, cooperation of users will be essential to develop a really useful tool. Therefore, suggestions about the philosophy as well as about new features and options in the actual functions are welcome.

A first point to discuss is the convenience of using Excel as the basis of CoDaPack or to create a

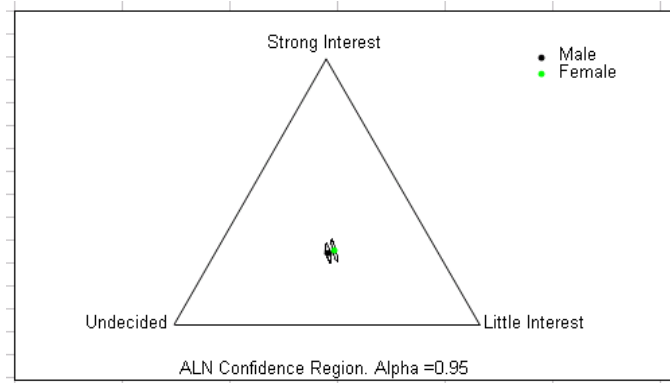


Figure 5: ALN confidence region of the two groups.

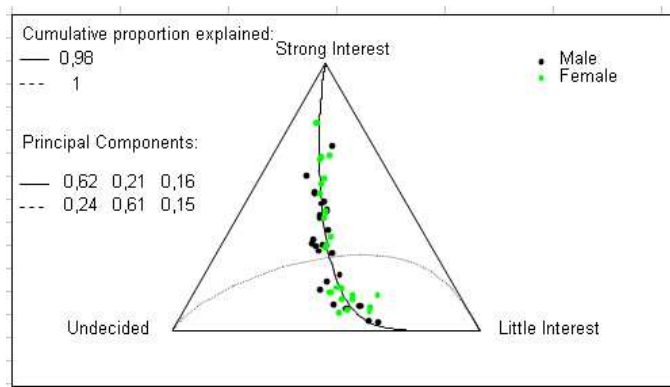


Figure 6: Principal component analysis.

new input/output platform without any dependence on commercial software. At this time we are thinking of a platform similar to other statistical packages with different windows for input data, results and graphs.

Our purpose is to add new useful features for application oriented users such as discriminant analysis, regression, cluster analysis and other methodologies proposed by them. But we would like to know the interest of including features of theoretical oriented users such as drawing arbitrary parallel "straight" lines inside the ternary diagram.

A special mention deserves the graphical output. We plan to include 3-D graphs with the additional option to perform zooms, to rotate the graphs to see them from different perspectives, and to export the graphs in vector format to other packages. Another feature to include in graphical output is the option to create graphs sequentially adding at different steps new parts to the same graph.

6 References

Aitchison, J., 1986, *The Statistical Analysis of Compositional Data: Monographs on Statistics and Applied Probability*. Chapman & Hall Ltd., London (UK). 416 p.

Aitchison, J., 1997, *NEWCODA: a software package for compositional data analysis*. Available from Social Science Research Centre, University of Hong Kong, Pokfulam Road, Hong Kong.

Aitchison, J., Barceló-Vidal, C., Egozcue, J. J., and Pawlowsky-Glahn, V., 2002, *A concise guide*

to the algebraic-geometric structure of the simplex, the sample space for compositional data analysis: *in* Proceedings of IAMG'02 — The annual conference of the International Association for Mathematical Geology, Berlin, Germany, September 15-20, 2002. p. 387-392.

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V., 2001, Mathematical foundations of compositional data analysis: *in* G. Ross (Ed.), Proceedings of IAMG'01 — The sixth annual conference of the International Association for Mathematical Geology, Volume CD-, pp. 20 p. electronic publication.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu Figueras, G. and Barceló-Vidal, C., 2003, Isometric Logratio Transformations for Compositional Data Analysis: *Mathematical Geology* 35(3), p. 279-300.

Greenwood, M., Stillwell, J. and Byars, A., 2001, Activity preferences of middle school physical education students: *The Physical Educator* 58(1), p. 26-29.

Martín-Fernández, J. A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2000, Zero replacement in compositional data sets, *in* Advances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98), p. 49-56. Berlin (Springer-Verlag).

Pawlowsky-Glahn, V. and Egozcue, J. J., 2001, Geometric approach to statistical analysis on the simplex: *Stochastic Environmental Research and Risk Assessment* 15, p. 384-398.

Pawlowsky-Glahn, V. and Egozcue, J. J., 2002, BLU estimators and compositional data: *Mathematical Geology* 34(3), p. 259-274.

Thió-Henestrosa, S., Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V., 2002, CoDaPack. A userfriendly freeware: *in* Proceedings of IAMG'02 — The annual conference of the International Association for Mathematical Geology, Berlin, Germany, September 15-20, 2002, p. 429-434.

Acknowledgements

This research has received financial support from the *Dirección General de Enseñanza Superior e Investigación Científica (DGESIC)* of the Spanish Ministry of Education and Culture through the project BFM2000-0540 and from the *Direcció General de Recerca* of the *Departament d'Universitats, Recerca i Societat de la Informació* of the *Generalitat de Catalunya* through the project 2001XT 00057.