# New Features of CoDaPack. An Userfriendly Compositional Data Package

**S. Thió-Henestrosa[1], R. Tolosana-Delgado[1], O. Gómez[1]**
[11]Dept. d´Informàtica i Matemàtica Aplicada. Universitat de Girona, Spain; santiago.thio@udg.es

**Abstract**

The statistical analysis of compositional data is commonly used in geological studies. As is well-known, compositions should be treated using logratios of parts, which are difficult to use correctly in standard statistical packages. In this paper we describe the new features of our freeware package, named CoDaPack, which implements most of the basic statistical methods suitable for compositional data. An example using real data is presented to illustrate the use of the package.

**Kew words:** Compositional data Analysis, Software.

## 1 Introduction

In the eighties, John Aitchison (1986) developed a new methodological approach for the statistical analysis of compositional data. After that, several other authors have published extensions: Martín-Fernández and others (2000), Barceló-Vidal and others (2001), Pawlowsky-Glahn and Egozcue (2001) and Egozcue and others (2003).

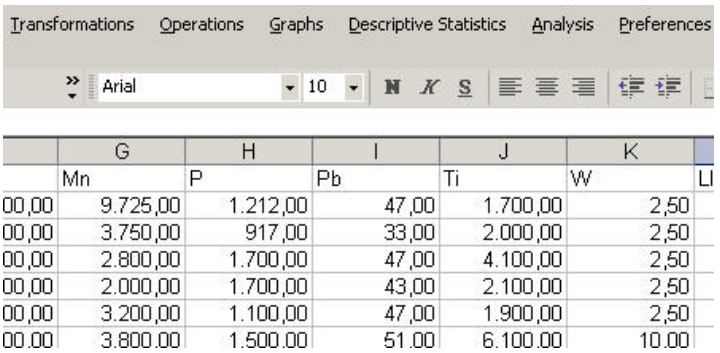
Aitchison implemented this new methodology in Basic routines grouped under the name CODA and later NEWCODA in Matlab (Aitchison, 1997). Also Reyment and Savazzi (1999) presented some routines in C++ and FORTRAN 90 and Reynolds and Billheimer (2002) developed some routines in S+/R.

This methodology is not straightforward to use, neither with that software, nor with standard statistical packages. For this reason it has been developed a new freeware, named CoDaPack (Thió et al., 2003 and 2005), based on CODA routines. At this moment CoDaPack includes some basic statistical methods: transformations, operations within the simplex, zero replacement techniques, ternary diagrams, descriptive statistics, reduction-of-dimension techniques, and some multivariate analysis methods. it is developed in visual basic and Open GL associated with Excel© and it is oriented towards users with a minimum knowledge of computers. It aims to be simple and easy to use.

To use CoDaPack one has to access Excel© and introduce the data in a standard spreadsheet, organized as a matrix where rows correspond to the observations and columns to the parts. The user executes macros that return numerical or graphical results. There are two kinds of numerical results: new variables and descriptive statistics, and both appear on the same sheet. Graphical output appears in independent windows and uses the OpenGL facilities. The latest version of this freeware package can be found on the web site *http://ima.udg.es/~thio/#Compositional Data Package*. Microsoft Excel© under Microsoft Windows© is required to run it.

## 2 CoDaPack Structure

Once installed CoDaPack, the user must open Excel© and introduce the data in a standard spreadsheet. The observations must be in rows and the parts in columns, and the first row of each column can be used to label the variables or it must be left blank (Fig. 1).




| | G | H | I | J | K | |
|---|---|---|---|---|---|---|
| | Mn | P | Pb | Ti | W | Ll |
| 00,00 | 9.725,00 | 1.212,00 | 47,00 | 1.700,00 | 2,50 | |
| 00,00 | 3.750,00 | 917,00 | 33,00 | 2.000,00 | 2,50 | |
| 00,00 | 2.800,00 | 1.700,00 | 47,00 | 4.100,00 | 2,50 | |
| 00,00 | 2.000,00 | 1.700,00 | 43,00 | 2.100,00 | 2,50 | |
| 00,00 | 3.200,00 | 1.100,00 | 47,00 | 1.900,00 | 2,50 | |
| 00.00 | 3.800.00 | 1.500.00 | 51.00 | 6.100.00 | 10.00 | |

**Figure 1.** An example of the organization of data inside the sheet .

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

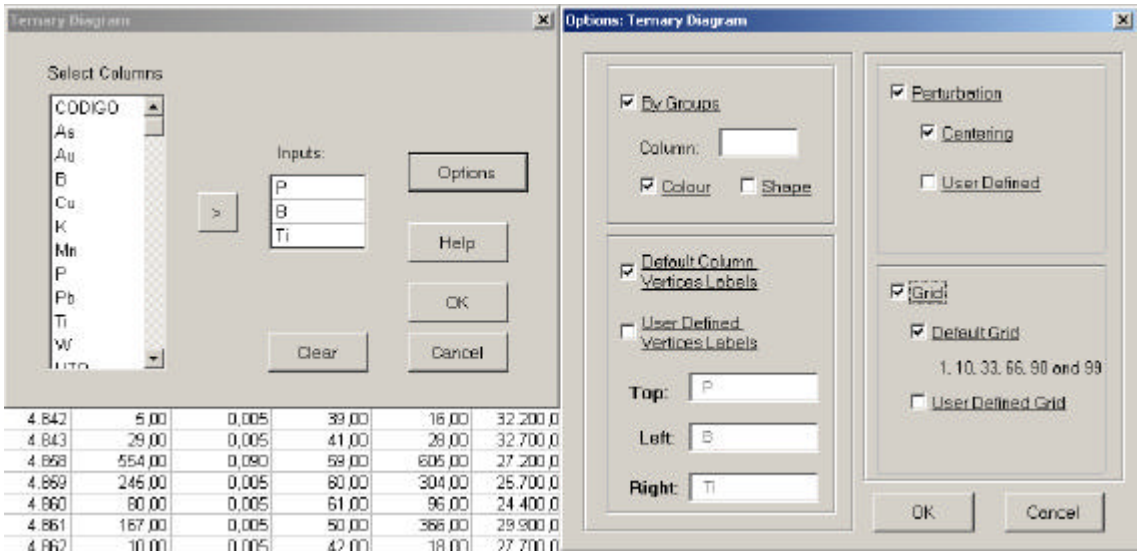



**Figure 2.** Dialogue screen for the ternary diagram menu.

In the present version there are 6 menus, with a total of 23 submenus which, after some dialogue, directly call each macro. The dialogues ask the user to input variables and, further parameters needed, as well as where to put these results. The first menu, *Transformations*, performs several transformations of data from real space to the simplex and viceversa, the same as in older versions of the software (Thió-Henestrosa et al., 2003)

The second menu, *Operations*, performs 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 plots 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*, also without changes, returns some characteristic values for a compositional data set, replacing classical statistics: center, variation matrix, total variance, and atypicality indices.

The fifth menu is a new one, *analysis*. At this time, it performs the logistic normality test (Aitchison, 1986). and, finally, the sixth menu, *preferences*, allows the user to define the constant of the sum constraint.

# 3 New features of CoDaPack

## 3.1 New graphic appareance

One of the flaws of the initial version of CoDaPack was that, in some conditions, it took too much time to draw the graphs. For this reason the functionality and options of these plots are the most striking new features of this package.

In order to solve this drawback several alternatives have been studied and finally the routines have been re-programmed using Open GL facilities, which allows a higher degree of customization and a quicker presentation.

Also the programming of these new graphical routines was done bearing in mind the next step to do in graphical routines: program some of the graphical routines in 3-D.

## 3.2 Macro *Summary* of *Descriptive Statistics* menu

This new macro performs nine descriptive statistics. Three of log-ratios: Variation Array (Aitchison, 1986), CLR Variance and Total Variance; and six compositional descriptive statistics: Centre, Min, Max and quartiles.

1) **Variation Array**: Returns a matrix where the upper diagonal contains the log-ratio variances and the lower diagonal contains the log-ratio means. That is, the *ij*-th component of the upper diagonal is var[ln($\mathbf{X}_i$/ $\mathbf{X}_j$)], and *ij*-th component of the lower diagonal is E[ln($\mathbf{X}_i$/ $\mathbf{X}_j$)], where ($i$,$j$=1,2,…,$D$) and $D$ is the number of parts.

2) **CLR Variance**: Returns the sum of log-ratio variances that involves each part. The sum of all CLR Variances is the Total Variance. So CLR-Variance$_i$= $\dfrac{\sum_{\mathbf{j=1,i\neq j}}^{\mathbf{D}} \mathrm{var}\left[\ln\left(\mathbf{X_i}/\mathbf{X_j}\right)\right]}{2\mathbf{D}}$.

3) **Total variance**: returns the sum of all CLR Variances.

4) **Center**: Returns centre of the data set, that is, $\hat{x} = \mathbf{C}[\mathbf{g}_1,\mathbf{g}_2,\ldots,\mathbf{g}_\mathbf{D}]$, where $\mathbf{g_i} = \left(\prod_{\mathbf{k=1}}^{\mathbf{N}} \mathbf{x_{ki}}\right)^{1/\mathbf{N}}$ symbolizes the geometric mean of part $\mathbf{X}_i$ in data set $\mathbf{X}$.

5) **Minimum and Maximum**: For each part of the data set $\mathbf{X}$ returns the minimum and the maximum of the closed data set $\mathbf{C}(\mathrm{X})$.

6) **Quartiles**: For each part of the data set $\mathbf{X}$ returns Q1, the median and Q3 of the closed data set $\mathbf{C}(\mathrm{X})$.

The user has to select the columns to analyze and where to put the results, and there are two further options on this routine (Fig. 3):

1) To perform the statistics for each group defined by a column,

2) The user can choose which statistics needs (at least one must be chosen).

Figure 4 shows an output of the summary routine with all statistics calculated:

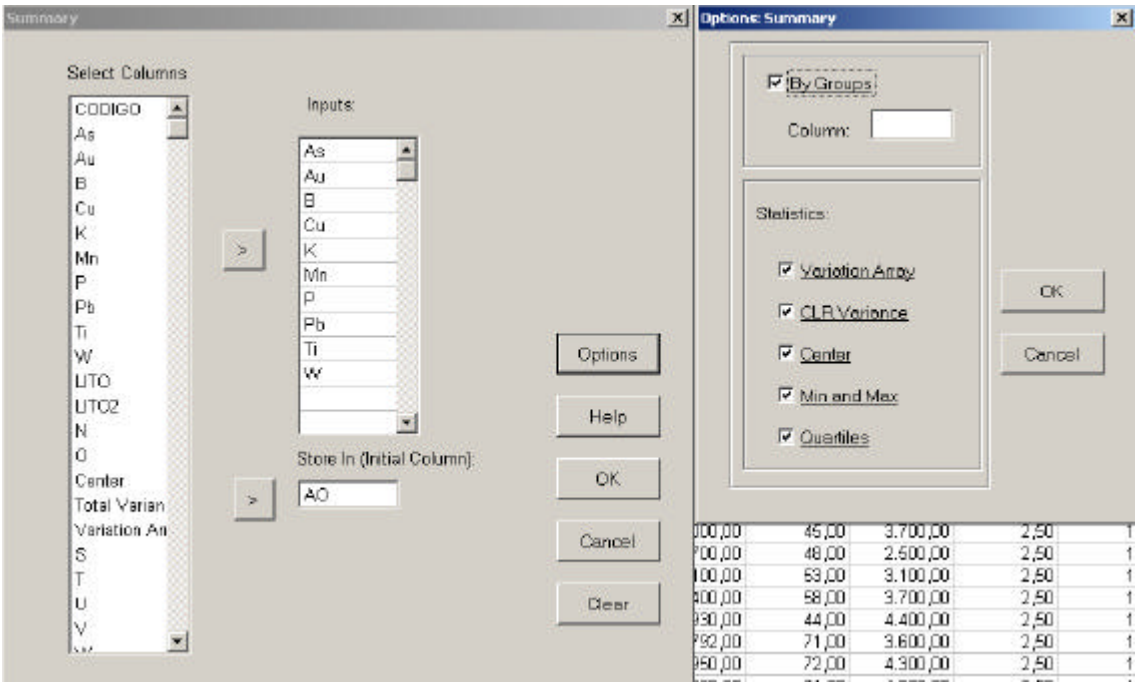



**Figure 3.** From *Descriptive Statistics* menu: Main form and *Options* form of *Summary* routine.

| AO | AP | AQ | AR | AS | AT | AU | AV | AW | AX | AY | AZ | BA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Variation Array | | | | | | | | | | | | |
| | As | Au | B | Cu | K | Mn | P | Pb | Ti | W | | |
| As | | 1,3015 | 1,3867 | 0,8878 | 2,3668 | 1,4652 | 1,4091 | 2,0846 | 2,2600 | 2,8719 | Variances | |
| Au | 7,6803 | | 0,4964 | 1,8522 | 0,7651 | 2,6295 | 0,7261 | 0,6555 | 0,7081 | 1,0267 | | |
| B | -0,9244 | -8,6048 | | 1,6330 | 0,2779 | 2,1373 | 0,1536 | 0,2267 | 0,4309 | 0,8056 | | |
| Cu | -0,8889 | -8,5493 | 0,0555 | | 2,8442 | 0,9150 | 1,3199 | 2,6303 | 2,8237 | 3,5717 | | |
| K | -7,5224 | -16,2027 | -6,5980 | -6,6536 | | 3,4504 | 0,5926 | 0,0700 | 0,2535 | 0,3433 | | |
| Mn | -4,2049 | -11,8852 | -3,2804 | -3,3359 | 3,3175 | | 1,5988 | 3,1927 | 3,5085 | 3,9732 | | |
| P | -4,0246 | -11,7049 | -3,1001 | -3,1556 | 3,4978 | 0,1803 | | 0,5094 | 0,7195 | 0,8967 | | |
| Pb | -1,2669 | -8,9662 | -0,3614 | -0,4169 | 6,2365 | 2,9190 | 2,7387 | | 0,2046 | 0,3391 | | |
| Ti | -5,5320 | -13,2124 | -4,6076 | -4,6631 | 1,9904 | -1,3272 | -1,5075 | -4,2461 | | 0,5005 | | |
| W | 1,2945 | -6,3858 | 2,2189 | 2,1634 | 8,8183 | 5,4994 | 5,3191 | 2,5804 | 6,8285 | | | |
| | Means | | | | | | | | | | Tot var | 6,4647 |
| | | | | | | | | | | | | |
| | As | Au | B | Cu | K | Mn | P | Pb | Ti | W | | |
| CLR Variance | 0,8032 | 0,5081 | 0,3674 | 0,9239 | 0,5492 | 1,1435 | 0,3963 | 0,4962 | 0,5705 | 0,7084 | | |
| | | | | | | | | | | | | |
| Compositional Descriptive Statistics | | | | | | | | | | | | |
| | As | Au | B | Cu | K | Mn | P | Pb | Ti | W | | |
| Center | 0,0004 | 0,0000002 | 0,0011 | 0,0011 | 0,8275 | 0,0300 | 0,0250 | 0,0016 | 0,1131 | 0,0001 | | |
| Min | 0,0001 | 0,0000001 | 0,0004 | 0,0001 | 0,3667 | 0,0012 | 0,0041 | 0,0006 | 0,0226 | 0,0001 | | |
| Max | 0,0139 | 0,0000034 | 0,0029 | 0,0306 | 0,9455 | 0,4463 | 0,0712 | 0,0027 | 0,2460 | 0,0004 | | |
| Q25 | 0,0001 | 0,0000001 | 0,0008 | 0,0002 | 0,7192 | 0,0086 | 0,0145 | 0,0013 | 0,0744 | 0,0001 | | |
| Median | 0,0002 | 0,0000001 | 0,0010 | 0,0005 | 0,8006 | 0,0258 | 0,0245 | 0,0014 | 0,1114 | 0,0001 | | |
| Q75 | 0,0011 | 0,0000002 | 0,0014 | 0,0035 | 0,8619 | 0,1355 | 0,0384 | 0,0017 | 0,1479 | 0,0002 | | |

**Figure 4.** Output of *Summary* routine from *Descriptive Statistics* menu.

### 3.3 Macro *Variation Array* of *Descriptive Statistics* menu

The initial version of CoDaPack calculates the Variation Matrix as defined in Aitchison (1986), that is, the *ij*-th component of the matrix is var[ln($\mathbf{X}_i$/ $\mathbf{X}_j$)] that is a symmetric matrix.

As the Variation Array includes on the upper diagonal the Variation Matrix, the new version of CoDaPack includes the Variation Array instead of Variation Matrix.

So, with this feature the user obtains the Variation array of the selected columns. It returns a matrix where upper diagonal contains the log-ratio variances and the lower diagonal contains the log-ratio means of the data set as described on *summary*.

The user has to select the columns to calculate the variation array and where to put the results.

### 3.4 Macro *Logistic Normality Test* of *Analysis* menu

This feature performs an ALR transformation of the input data and then calculates tests for:

1) All marginal, univariate distributions (with a total of *D*-1 tests)

2) All bivariate angle distributions (with a total of 1/2(*D*-1)(*D*-2) tests)

3) The (*D*-1)-dimensional radius distribution.

For each kind of test Anderson-Darling, Cramer-von Misses and Watson tests are performed.

| BU | BV | BW | BX |
|---|---|---|---|
| TEST OF LOGISTIC NORMALITY OF PARTS (As,Au,B,Cu) | | | |
| | | | |
| MARGINAL UNIVARIATE DISTRIBUTIONS | | | |
| | Anderson-Darling | Cramer-von Mises | Watson |
| marginal 1 | 2,461087567 | 0,459379605 | 0,40414015 |
| marginal 2 | 3,520074995 | 0,567546563 | 0,51276117 |
| marginal 3 | 4,573281394 | 0,810371625 | 0,74532038 |
| | | | |
| BIVARIATE ANGLE DISTRIBUTIONS | | | |
| (i,j) | Anderson-Darling | Cramer-von Mises | Watson |
| (1,2) | 2,151849375 | 0,417769607 | 0,34005398 |
| (1,3) | 3,217655896 | 0,632802237 | 0,39906187 |
| (2,3) | 5,169548979 | 0,978485245 | 0,45062465 |
| | | | |
| RADIUS TESTS | | | |
| | Anderson-Darling | Cramer-von Mises | Watson |
| | 4,37063908 | 0,805494626 | 0,56650617 |

**Figure 5.** Output of *Logistic Normality Test* routine from *Analysis* menu.

### 3.5 New installation an uninstallation macros

Another drawback on the first version of CoDaPack was the installation of menus that had to be done one by one. For this reason the new version includes two new routines to install and uninstall the menus automatically.

These routines have been kindly provided by Agnes Schumann of the Free University of Berlin.

## 4 Example

The analyzed dataset contains 95 samples of soil obtained in a soil geochemical campaign for a mineral prospection in the Iberian Pyrite Belt (SW Spain). Two kinds of rocks are identified in the sub-soil, and registered as a complementary variable (litology); these are: *metavulcanitas* (metacinerites, shales and porphyry acidic tuffs) and *granitoides* (amfibolic granitoid). These two igneous rocks are related to different Cambric rifting events (details can be found in Tolosana-Delgado et al., 2004). All calculations and figures presented are obtained with CoDaPack.

As a first approach for an explanatory analysis it is possible to calculate with CoDaPack, following Aitchison methodology (1986), the center and total variance (Fig. 4). The main part of this data set corresponds to K which has the highest value on the center (0.83) followed by Ti (0.11) while the rest of the values are very small, specially Au ($2.1 \times 10^{-7}$).

The variation array (Fig. 4) shows that Mn is the part with largest relative variability in the clr-transformed data set, followed by Cu and As.

It is possible to see a two dimensional graphical display of both parts and observations of the clr-transformed data set. The compositional biplot (Aitchison and Greenacre, 2002) displayed on Figure 6 explains an 82% of total variability of the data set. It verifies that largest variability appears on clr(Mn), followed by clr(Cu), clr(As) and clr(W). As the link between two vertices is roughly proportional to the logratio variance var[ln($\mathbf{X}_i$/ $\mathbf{X}_j$)], and the corresponding vertices of Pb, K and W are very close, the variation between those parts is expected to be very low.

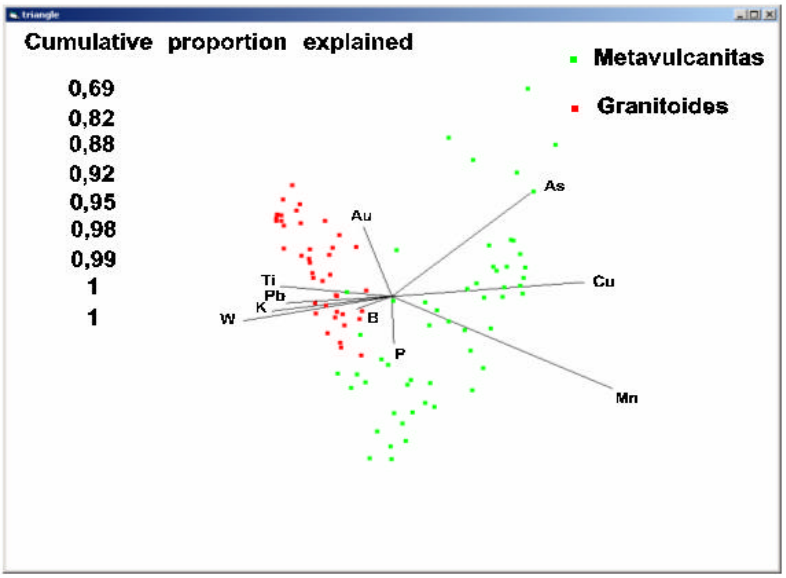



**Figure 6.** Compositional biplot of the data set.

Also it is possible to see that both groups, *metavulcanitas* and *granitoides*, have differences on compositions. *Metavulcanitas* have larger values on Mn, Cu and As while *granitoides* on W, K, Ti and Pb.

Finally when three or more vertices appear approximately on a straight line in a compositional biplot, a one-dimensional variability between the involved parts is suggested. For example, in Figure 7 there are the ternary diagram of As, Pb and W of the raw data set and the centred data set. This centering operation was first mathematically defined by Martín-Fernández and others (1999), although it is a classical trick in geology. Applying this operation, the data set is translated, using the perturbation operation, into the barycentre of the simplex. The operation of centering was performed with the corresponding submenu of *Operations.*

The seen linear pattern can be modelled using compositional principal component analysis (Aitchison 1986) (Fig. 8) of the centered data set.

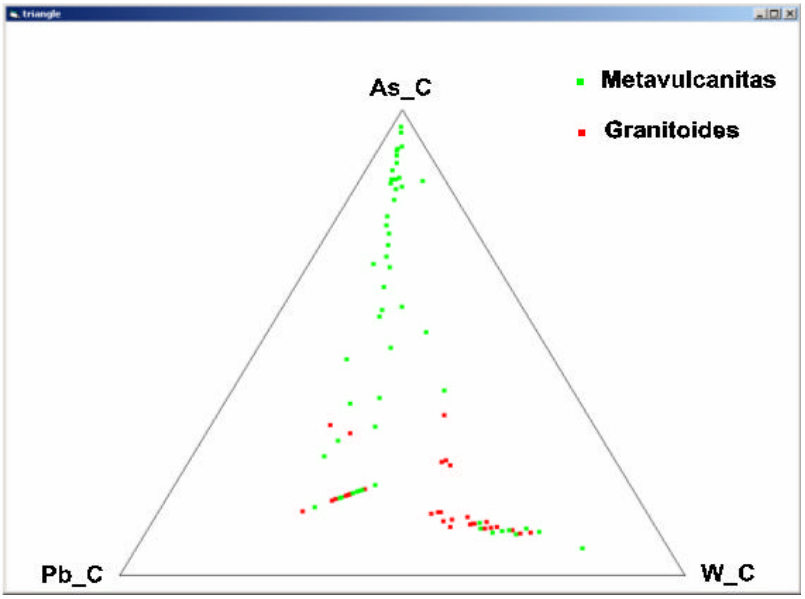



**Figure 7.** [As. Pb, W] subcomposition of the data set:
(Left) not centred; (Right) centred. "_C" in (Right) indicates parts have been centred.

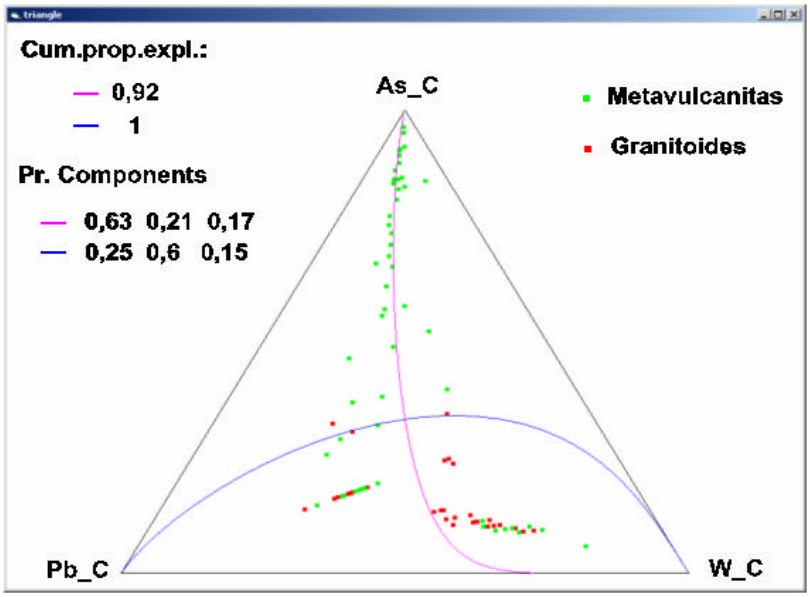



**Figure 8.** [As. Pb, W] subcomposition of the centered data set:
Curves show the backtransformed first and second linear compositional PCA axes,
which become curved in this projection. "_C" indicates parts have been centred.

## 5 Conclusion

CoDaPack is a software that is still growing. In the future news routines will be programmed by the group of Girona or with collaboration with other groups. At this moment the work around CoDaPack goes in three main directions:

**1) New graphical output**: In a few months we would start developing new routines in 3-D such as Tethrahedron diagrams, Biplots and $R^3$ plots. All this 3-D plots should offer the option to be rotated and zoomed.

2) **New statistic capabilities**: The group of Girona are always working, sometimes in collaboration with other groups, on the programming of new statistic routines. At this time we are working on Discrimination Analysis, Balances Analysis and Cluster Analysis. The last one is led by a group of the Universidad de la Plata in Argentina.

3) **Software maintenance**: In this field we are working on several directions in order to improve CoDaPack: a) repairing the bugs, b) providing the software of an internal coherence in order to facilitate the growing of CoDaPack, c) updating the users manual and d) changing CoDaPack appearance: in a future we will create an interactive help and we are considering to introduce a new window to write in textual outputs.

## Acknowledgements

This research has received financial support from the Dirección General de Investigación of the Spanish MCyT through the project BFM2003-05640/MATE and from the Departament d'Universitats, Recerca i Societat de la Informació of the Generalitat de Catalunya through the project 2003XT 00079. The data set was kindly provided by A. Canales, from PRESUR, the Spanish prospecting enterprise.

## References

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

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., and Greenacre, M. (2002). Biplots of compositional data. *Applied Statistics 51*, 375-392.

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-, 20 pp. 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), 279-300.

Martín-Fernández, J.A., Bren, M., Barceló-Vidal, C. and Pawlowsky-Glahn, V, (1999). A measure of difference for compositional data based on measures of divergence. In S. J. Lippard, A. Naess and R. Sinding-Larsen (Eds.). *Proceedings of IAMG'99, The Fifth Annual Conference of the International Association for Mathematical Geology: Trondheim, Norway,* v. 1, p. 211-216.

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.

Reyment, R.A. and Savazzi, E. (1999). *Aspects of Multivariate Statistical Analysis in Geology*. Elsevier.

Reynolds, J.H. and Billheimer, D. (2002). *Basic Compositional Data Analysis Functions for S+/R*. In http://www.biostat.wustl.edu/archives/html/s-news/2003-12/msg00139.html. (on September, 9, 2005).

Thió-Henestrosa, S., Barceló-Vidal, C., Martín-Fernández, J.A., and Pawlowsky-Glahn, V. (2003). CoDaPack. An Excel and Visual Basic Based Sotfware of Compositional Data Analysis. Current Version and Discussion for Upcoming Versions. In S. Thió-Henestrosa and J.A. Martín-Fernández, (Eds.). *Proceedings of CODAWORK'03, The First Compositional Data Analysis Workshop: Girona, Spain*, 8p. (CD, electronic publication).

Thió-Henestrosa, S. and Martín-Fernández, J.A. (2005). Dealing with Compositional Data: The Freware CoDaPack. *Mathematical Geology 37*(7), 777-797.

Tolosana-Delgado, R., Canals, A., Pawlowsky-Glahn, V. (2004). Characterizing a Cu anomaly using univariate kriging techniques. In González, Segura, Colombo, (Eds.). *Geo-Temas: Actas del VI Congreso Geológico de España*. Vol 6(1) pp 117-120.