The data sets from Aitchison’s book in the ”compositions” package

Matevž Bren[1,2]

[1] Institute of Mathematic, Phisics and Mechanics, Slovenia

[2]University of Maribor, FOV, Slovenia

ifcs 2006

Ljubljana, 25 – 29 July 2006

Outline

Aitchison's book

Groundwork on Compositional Data Analysis (CDA) is the book of John Aitchison from 1986 *The statistical Analysis of Compositional Data.*

From the book we quote:

> *The properties of many substances or objects, such as gasoline, metal alloys and cakes, depend on the particular mixture, or composition, of their ingredients. The purpose of the experiments with different mixtures is to obtain some understanding of the nature and extend of the dependence of the properties on the composition. In the analysis of such experiments the composition is confined to the role of a covariate.*

There are forty data sets coming from various disciplines –
Geology, Social science, Economy, Medicine, Agriculture, Pedagogy, Anatomy, Ecology, Sports. . .
– that serve as examples in the book.

Aitchison’s Researchers daily activities data

gives activity patterns of a statistician for 20 days – proportion of a day in activity teaching, consultation, administration, research, other wakeful activities and sleep.

Proportion of a day in activity

teac – teaching
cons – consultation
admi – administration
rese – research
wake – other wakeful activities
slee – sleep

We consider only the activity proportions, not the absolute values – *compositional data, mixtures.*

Researchers daily activities

day	**teac**	**cons**	**admi**	**rese**	**wake**	**slee**
1	0.162	0.041	0.138	0.123	0.254	0.282
2	0.200	0.039	0.073	0.076	0.346	0.266
3	0.201	0.082	0.115	0.146	0.194	0.261
4	0.134	0.077	0.107	0.146	0.214	0.321
5	0.224	0.080	0.091	0.162	0.195	0.248
6	0.144	0.063	0.103	0.123	0.316	0.252
7	0.125	0.054	0.137	0.102	0.312	0.270
8	0.127	0.077	0.110	0.101	0.341	0.244
9	0.139	0.052	0.128	0.111	0.266	0.304
10	0.108	0.052	0.082	0.075	0.413	0.270
11	0.187	0.091	0.113	0.116	0.264	0.228
12	0.184	0.070	0.066	0.151	0.305	0.216
13	0.155	0.086	0.101	0.119	0.225	0.315
14	0.181	0.097	0.081	0.164	0.271	0.206
15	0.224	0.096	0.101	0.142	0.203	0.234
16	0.198	0.067	0.139	0.154	0.162	0.281
17	0.214	0.073	0.102	0.130	0.201	0.281
18	0.132	0.037	0.148	0.099	0.307	0.277
19	0.167	0.073	0.127	0.122	0.266	0.245
20	0.166	0.064	0.101	0.145	0.242	0.282

Compositional data sample space

Compositions (compounds, mixtures, alloy ...) can be represented with vectors of the portions of individual components. The portions are nonnegative and they have a constant sum.

One of suitable sample spaces for compositional data

$$\mathbf{w} = (w_1, \ldots, w_D), \quad w_k \geq 0,\ k = 1, \ldots, D,$$

$$w_1 + \cdots + w_D = \text{const.}$$

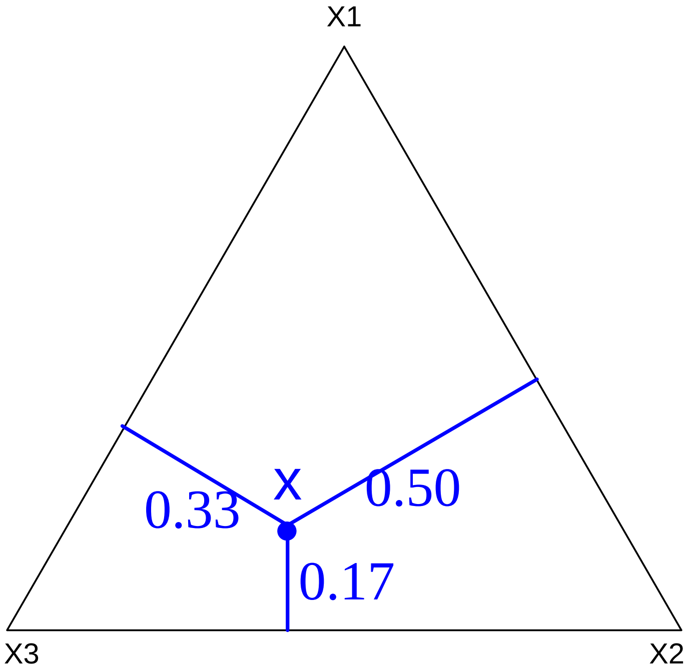
is the d - dimensional *unit simplex* $(d := D - 1)$

$$\mathcal{S}^d := \{\mathbf{x} = (x_1, \ldots, x_D);\ x_k > 0,\ k = 1, \ldots, D \wedge x_1 + \cdots + x_D = 1\}$$

Any vector of positive components $\mathbf{w} \in \mathbb{R}_+^D$ can be projected onto the simplex by the *closure operation*

$$\mathcal{C}(\mathbf{w}) = \left(\frac{w_1}{\sum w_k}, \ldots, \frac{w_D}{\sum w_k} \right) \in \mathcal{S}^d.$$

Ternary diagrams



Graphical representation of three part compositions $\mathbf{x} = (0.17, 0.33, 0.50)$.

are a convenient way of displaying the 3-part compositions.

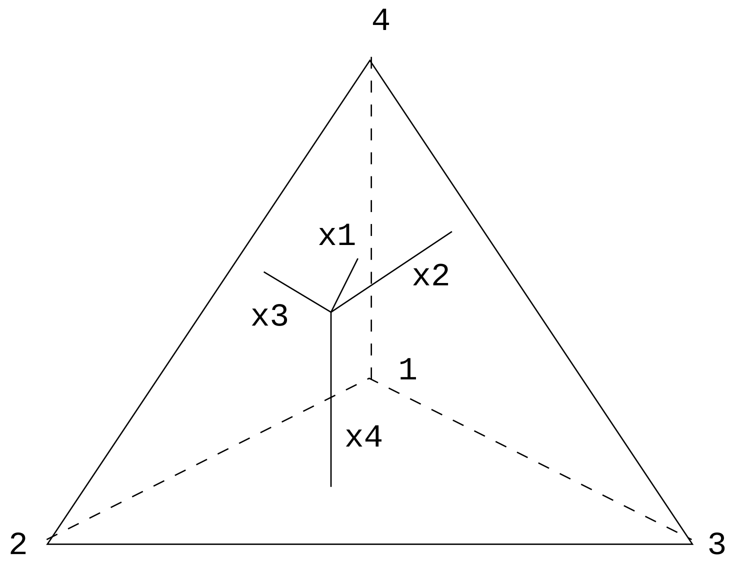
The triangle Δ X1X2X3 is equilateral and has the altitude equal to 1.

For any point x in the triangle the perpendiculars x_1, x_2 and x_3 to the sides opposite X1, X2, X3 satisfy

$$\begin{aligned} x_i &\geq 0, \qquad i = 1, 2, 3 \\ x_1 + x_2 + x_3 &= 1. \end{aligned}$$

One-to-one mapping from the 3-part compositions to the points in the triangle.

Tetrahedral displays



Graphical representation of four part compositions.

a convenient way of graphical representation of a 4-part composition.

The tetrahedron 1234 is regular and has the altitude equal to 1.

For any point x in the tetrahedron the perpendiculars x_1, x_2, x_3 and x_4 to the face opposite the vertex 1, 2, 3 and 4, respectively, satisfy

$$\begin{aligned} x_i &\geq 0, \quad i = 1, 2, 3, 4 \\ x_1 + x_2 + x_3 + x_4 &= 1. \end{aligned}$$

CDA Software tools

- CoDa by John Aitchison, 1986, written in Quick Basic available with the Aitchisons book. Upgraded by John Bacon-Shone.

- CoDaPack freeware SW by Santiago Thió and Martín-Fernández, 2001, in Excel available at `http://ima.udg.es/Recerca/EIO/inici_cat.html`

- atemps in R
  - by Joel Raynolds and Dean Billheimer at `http://www.biostat.wustl.edu/archives/html/s-news/2003-12/msg0013`
  - **`MixeR`** by Batagelj and Bren, 2003, available at `http://vlado.fmf.uni-lj.si/pub/MixeR`
  - compositions package by K. Gerald van den Boogaart and Raimon Tolosana Delgado, June 2005, available at `http://cran.r-project.org/src/contrib/Descriptions/compositions.h`

R a free statistical language and environment

R (`http://www.r-project.org/`) is a free language and environment for statistical computing and graphics. The environment in which many classical and modern statistical techniques have been implemented, but many are supplied as packages. There are 8 *standard* packages and many more are available through the `cran` family of Internet sites `http://cran.r-project.org`

The term *environment* is intended to characterize it as a fully planned and coherent system, rather than an incremental accretion of very specific and inflexible tools, as is frequently the case with other data analysis software.

R is an integrated suite of software facilities for data manipulation, calculation and graphical display.

The R Package 'Compositions'

The package provides functions for the consistent analysis of compositional data (e.g. portions or substances) and positive numbers (e.g. concentrations) in the way proposed by Atchison. In the package are implemented

Graphical Presentations Ternary diagrams, Area plots, Boxplots,

Descriptive Statistics

Transformations Subcompositions, Marginal comp., Grouping. . .

Multivariate Methods Principle Component Analysis, Cluster Analysis, Discrimination Analysis and Linear Models.

The package supports *four different multivariate scales* represented by four classes. The classes differ on the assumption whether or not the total amount is meaningful for the problem and whether the geometry of the differences is a relative (log-scale) distance or a absolute (Euclidean) distance.

. . . The R Package ’Compositions’

”rcomp” The total amount is meaningless or the individual amounts are part of a whole (in equal units) and the data should be analyzed in real (non relative) geometry.

”acomp” The total amount is meaningless or the individual amounts are part of a whole (in equal units) and the data should be analyzed in a relative geometry.

”rplus” The total amount is meaningful and data is analyzed in the real (non relative) geometry.

”aplus” The total amount is meaningful and the data should be analyzed in relative geometry.

The package is based on the concept that the type of analysis is given by the user (e.g. to plot) and the type of the data (e.g. acomp).

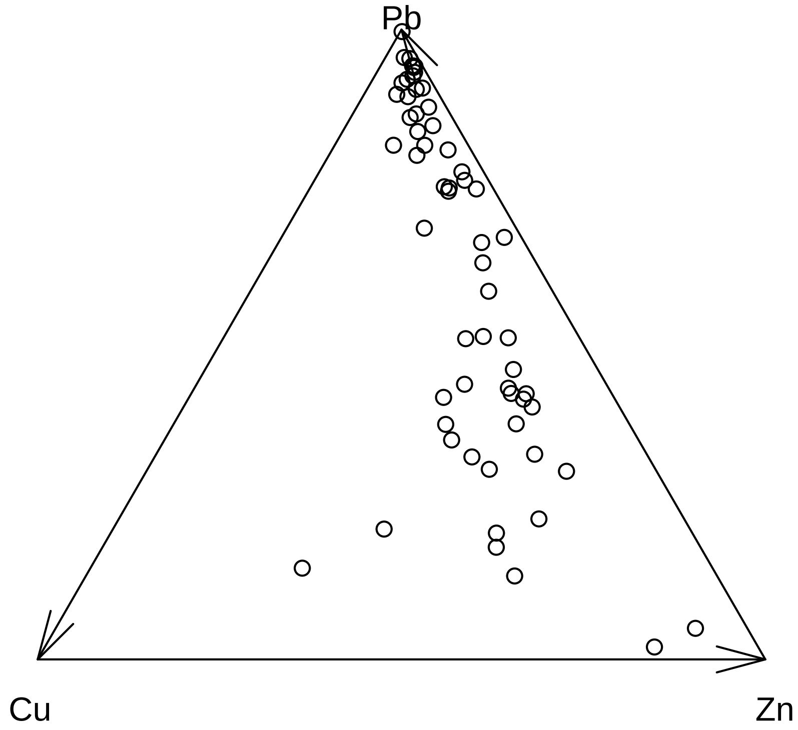
We will illustrate this with an exemplary dataset from the package:

. . . The R Package ’Compositions’

```
> library(compositions)
Welcome to compositions, a package for compositional
data analysis.
> data(SimulatedAmounts)
> dat <- rcomp(sa.lognormals)
> dat
          Cu          Zn         Pb
 [1,] 0.097971136 0.391326782 0.51070208
 [2,] 0.015828238 0.051778890 0.93239287
 [3,] 0.023646054 0.229295970 0.74705798
          ...         ...         ...
[59,] 0.087881617 0.327464869 0.58465351
[60,] 0.078617049 0.120922730 0.80046022
attr(,"class")
[1] "rcomp"
```

. . . The R Package ’Compositions’

```
> plot(dat)
```



Plot of the 3-part compositions in class ”rcomp”.

. . . The R Package ’Compositions’

```
> dat <- acomp(sa.lognormals)
> dat
> plot(dat)
```

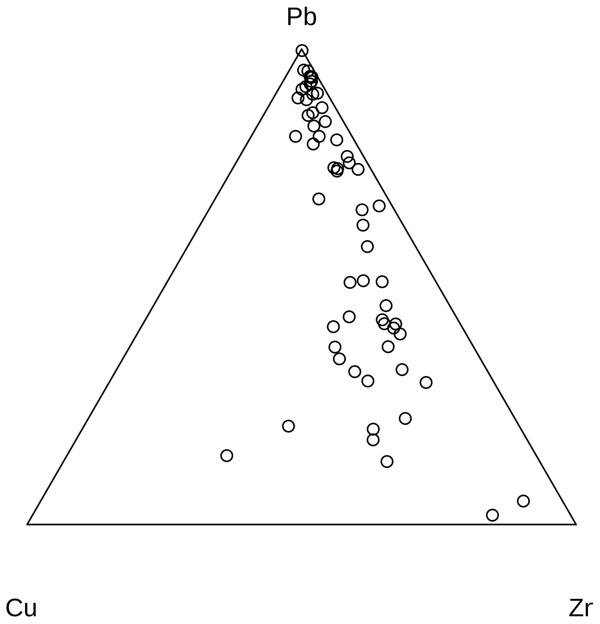
```
         Cu          Zn          Pb
 [1,] 0.097971136 0.391326782 0.51070208
 [2,] 0.015828238 0.051778890 0.93239287
 [3,] 0.023646054 0.229295970 0.74705798

         ...         ...         ...
         ...         ...         ...

[59,] 0.087881617 0.327464869 0.58465351
[60,] 0.078617049 0.120922730 0.80046022

attr(,"class")

[1] "acomp"
```



Plot of the 3-part compositions in class ”acomp”.

. . . The R Package 'Compositions'

```
> dat <- rplus(sa.lognormals)
> dat
> plot(dat)
```

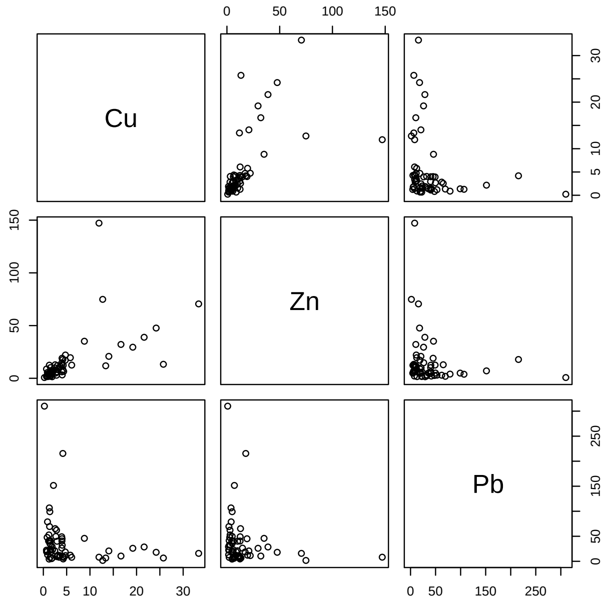
```
          Cu          Zn          Pb
 [1,] 8.8043262 35.1671810 45.895025
 [2,] 0.8115227 2.6547329 47.804310
 [3,] 1.2836130 12.4472047 40.553628

         . . .       . . .       . . .
         . . .       . . .       . . .

[59,] 3.9619526 14.7630454 26.357839
[60,] 3.9854998 6.1301909 40.579417

attr(,"class")

[1] "rplus"
```



Plot of the 3-part data in class "rplus".

...The R Package 'Compositions'

```
> dat <- aplus(sa.lognormals)
> dat
> plot(dat)
```

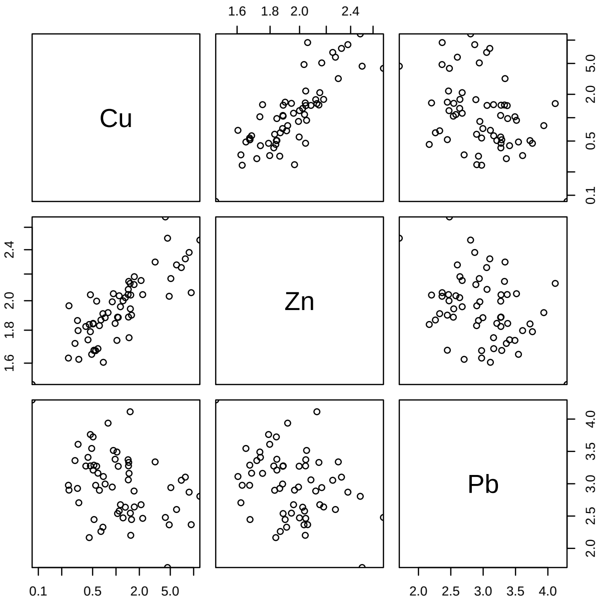
```
             Cu         Zn        Pb
 [1,] 8.8043262 35.1671810 45.895025
 [2,] 0.8115227  2.6547329 47.804310
 [3,] 1.2836130 12.4472047 40.553628

         ...         ...        ...
         ...         ...        ...

[59,] 3.9619526 14.7630454 26.357839
[60,] 3.9854998  6.1301909 40.579417

attr(,"class")

[1] "aplus"
```



Plot of the 3-part amounts in class "aplus".

…The R Package ’Compositions’

Depending on the type of the data `acomp, aplus, rcomp, rplus` a different plot function called `plot.ClassName` is invoked and plots the data in a fashion most suitable for the given data type.

This principle is used all over the package.

In the package we will include the data sets from Aitchison’s book. There are forty data sets that serve as examples in the book and will be available with the ”compositions” package under the GNU Public Library Licence Version 2.

We thank the author for generously giving us the data files.

The data sets

No.	DATA	TOPIC	TITLE
1	Hongite	Geology	Compositions of 25 specimens of hongite
2	Kongite	Geology	Compositions of 25 specimens of kongite
3	Boxite	Geology	Comp. and depth of 25 specimens of boxite
4	Coxite	Geology	Comp., depths and porosities of 25 sp. of coxite
5	ArcticLake	Geology	Arctic lake sediment samples
6	SkyeAFM	Geology	AFM comp. of aphyric Skye lavas
7	Supervisor	Work	Proportions of supervisor's statements
8	HouseholdExp	SocEconomy	Household Expenditures
9	Metabolites	Medicine	Steroid metabolite in adults and children
10	Activity10	Work	Activity patterns of a statistician
11	WhiteCells	Medicine	White-cell comp. of blood samples by two methods
12	Yatquad	Agriculture	Yatquad fruit evaluation
13	Firework	Technic	Firework mixtures
14	ClamEast	ClamEcology	Color-size comp. of clam colonies East Bay
15	ClamWest	ClamEcology	Color-size comp. of clam colonies West Bay
16	SerumProtein	Medicine	Serum Protein compositions of blood samples
17	DiagnosticProb	Pedagogy	Diagnostic probabilities
18	Glacial	Geology	Comp. of glacial tills
19	PogoJump	Sports	Honk Kong Pogo-Jumps Championship
20	Sediments	Geology	Proportions in sediments specimens

. . . The data sets

No.	SUM CONSTRAINT	DIMENSION	ZERO VAL.	ERROR
1	100, rounding errors	25x5	None	
2	100	25x5	None	
3	100, rounding errors	25x5, depth	None	case 6
4	100	25x5, depth, porosities	None	
5	100, rounding errors	39x3, depth	None	
6	100	23x3	None	
7	1, rounding errors	18x4, time, supervisee	None	
8	No	40x4, sex	None	
9	No	67x3, adult or children	None	
10	1	20x6	None	
11	1, rounding errors	30x3, two times	None	
12	1	40x3, two times, un/treated	None	
13	1	81x5, brilliance, vorticity	None	
14	1, rounding errors	20x6	None	
15	1	20x6	None	
16	1, rounding errors	36x4, type	None	case 33
17	1, rounding errors	30x3, type	None	
18	100, rounding errors	92x4, counts	Yes	cases 48, 69, 91
19	No	28x3, finalist	None	
20	1	21x3, type	None	

Activity10.Rd

Each data are documented in the Rd files comprising

name Activity10

docType data

title Activity patterns of a statistician for 20 days

description Proportion of a day in activity teaching, consulting, administrating, research, other wakeful activities and sleep of a statistician for 20 days are given.

usage data(Activity10)

format Each row of the data file contains one record. In the first row there is the title, in the second number of variables, in the third number of cases, then the columns names beginning with Case no., then variables names, and then the cases. A case begins with the case No. and continues with the variables values.

source Aitchison: CODA microcomputer statistical package, 1986, the file name STATDAY.DAT, here included under the GNU Public Library Licence Version 2 or newer.

details The activity of an academic statistician were divided into following six categories

teac	teaching
cons	consultation
admi	administration
rese	research
wake	other wakeful activities
slee	sleep

Data show the proportions of the 24 hours devoted to each activity, recorded on each of 20 days, selected randomly from working days in alternate weeks, so as to avoid any possible carry-over effects, such as short-sleep day being compensated by make-up sleep on a succeeding day.

The six activity may be divided into two categories ’work’ comprising activities 1,2,3,4: and ’leisure’ comprising activities 5 and 6.

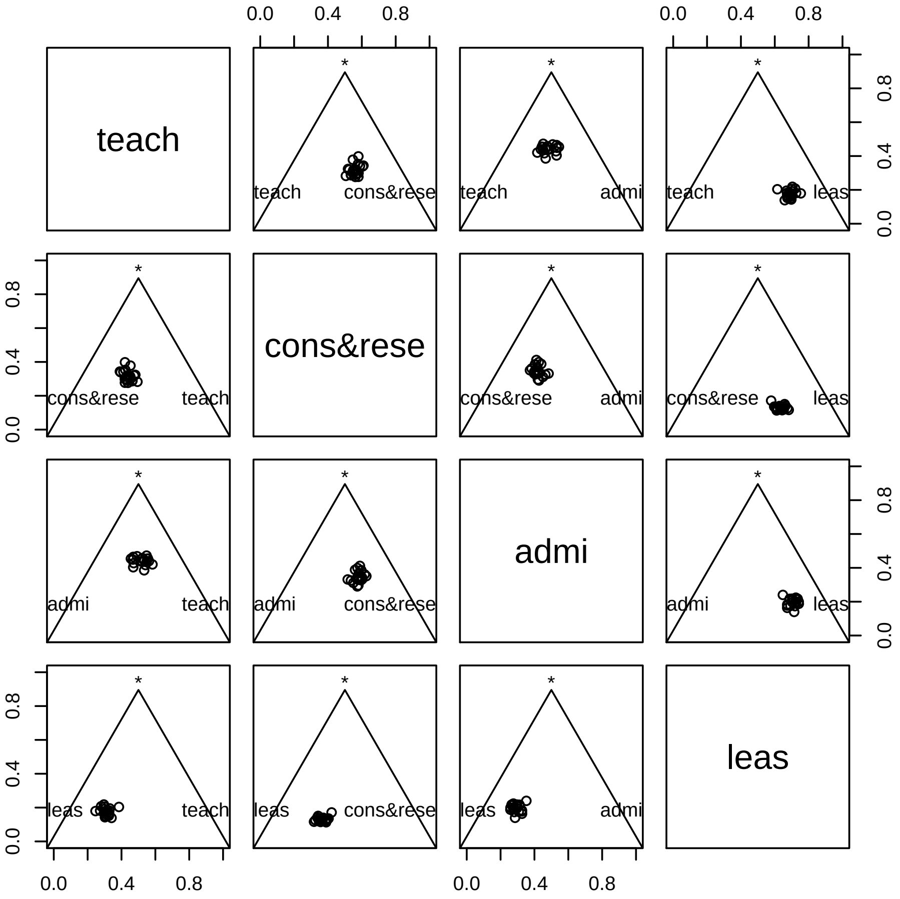
All rows sum to one.

references Aitchison: The Statistical Analysis of Compositional Data, 1986, Data 10, pp15.

keyword datasets

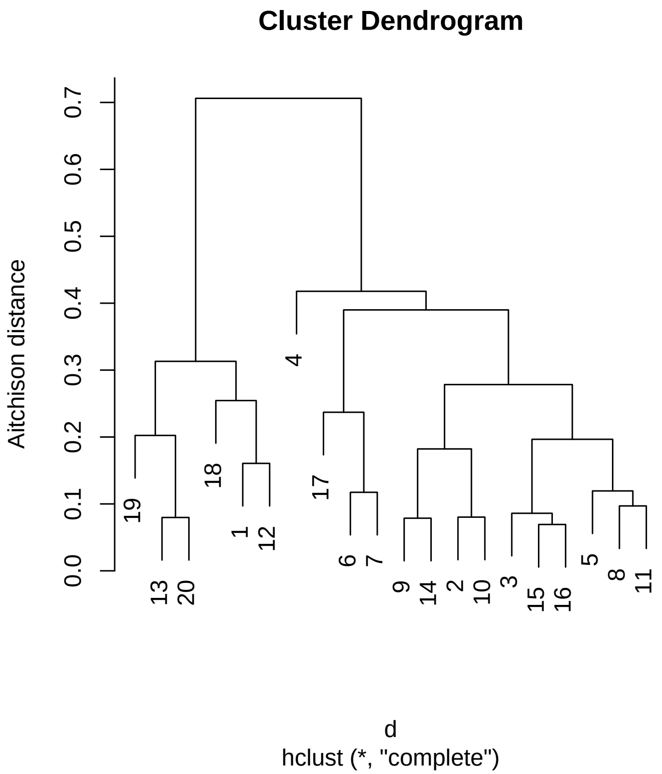
examples

```
> m <- data(Activity10)
> m<-cbind(m[,1],m[,2]+m[,4],m[,3], m[,5]+m[6])
> dimnames(m)[[2]]<-c(’teach’,’cons&rese’,’admi’,
’leas’) # The six activity may be combined into four categories
’teaching’, ’consulting&research’, ’administrating’ and ’leisure’.
> plot(acomp(m))
```

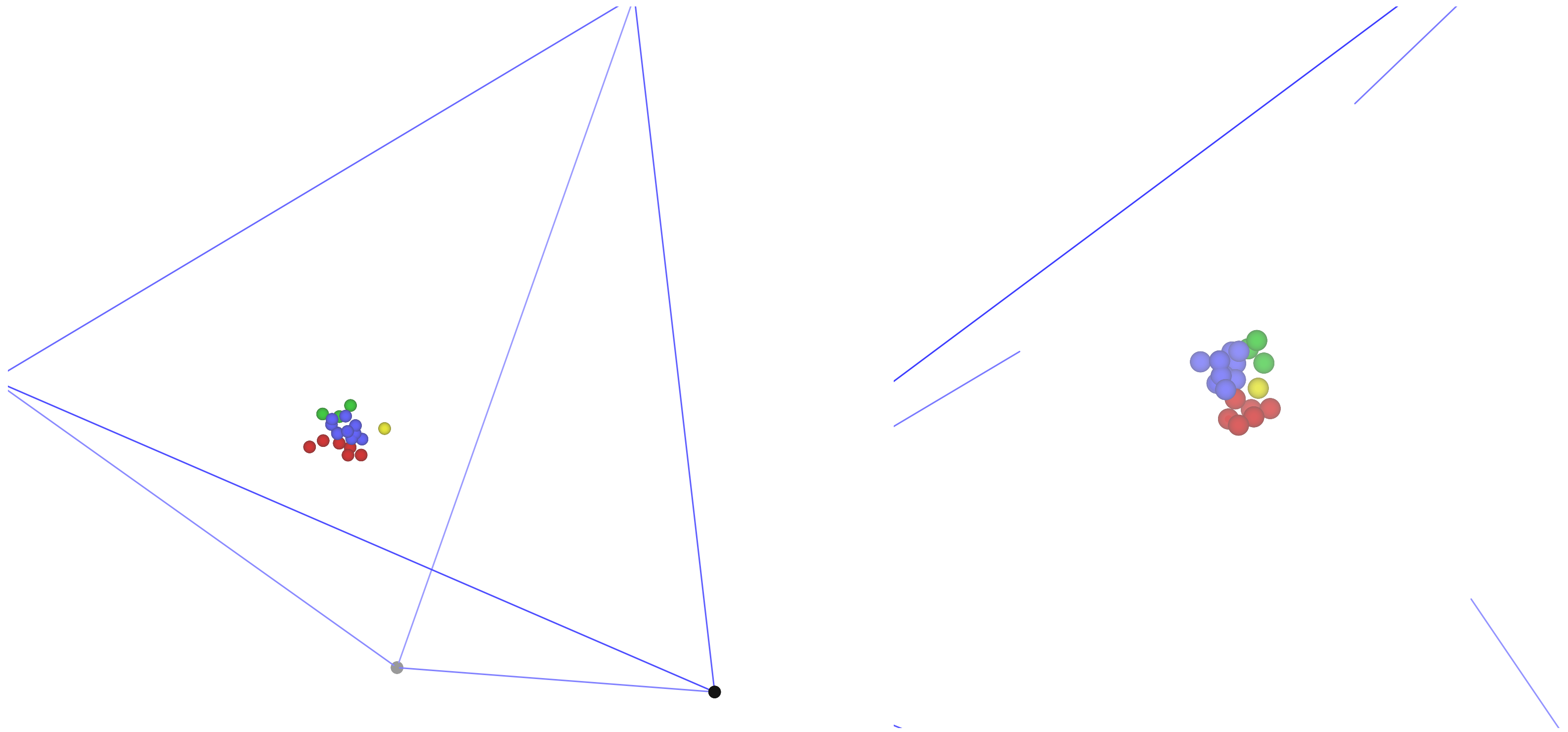


Plot of the 4-part compositions in class "acomp".

```
> d <- dist(acomp(m$mat))   # computes the Aitchison distance
> hc <- hclust(d, method = "complete", members=NULL)
> plot(hc , labels = NULL, hang = 0.1, main = "Cluster
Dendrogram", ylab = "Aitchison distance")
```



```
> mix.Quad2kin(’ac4.kin’, m, clu=cutree(hc,4), scale=0.1)
```



Snapshots of ac4.kin 3D KING view of tetrahedral display of activity data 4-part compositions.

The kin file we display with a 3-D interactive KiNG viewer – a free software available at `http://kinemage.biochem.duke.edu/software/software1.html`

A kinemage is a dynamic, 3-D illustration. We take advantage of that

- by rotating it and twisting it around with the mouse click near the centre of the graphics window and slowly dragging right or left, up or down,
- by clicking on points with the mouse, the label associated with each point will appear in the bottom left of the graphics area,
- also the distance from this point to the last will be displayed,
- With the right button drag we can zoom in and out of the picture.

This animation supports colouring and different sizing of points.

Conclusions

We have demonstrated some ’compositions’ routines and features for visualization of three and four part (sub)compositions with the exemplary data from examples of Aitchison’s book.

The data files, documentation on the data and the table with data files names, titles, topics. . . will be available at ’compositions’ home page

http://cran.r-project.org/src/contrib/Descriptions/compositions.html