COMPOSITIONAL DATA ANALYSIS WITH R

M. Bren[1] **V. Batagelj**[2]
[1]University of Maribor, Slovenia; matevz.bren@fov.uni-mb.si
[2]University of Ljubljana, Slovenia

1 Introduction

We present

- Examples of compositional data:
  - A household budget survey,
  - Slovene communities labour force data.
- The *simplex* – a suitable sample space for compositional data and Aitchison's geometry:
  - the vector space of the simplex
  - the Aitchison distance, the norm and the inner product,
  - transformation to the barycenter.
- R (http://www.r-project.org/), a free language and environment for statistical computing and graphics.
- Conclusions and future work.

2 Examples

Household budget survey
from the Aitchison book *The Statistical Analysis of Compositional Data*:
Sample survey of single persons living alone in a rented accommodation, twenty men and twenty women were randomly selected and asked to record over a period of one month their expenditures on the following four mutually exclusive and exhaustive commodity groups.

H – housing, including fuel and light,
F – foodstuffs, including alcohol and tobacco,
O – other goods, including clothing, footwear. . . ,
S – services, including transport and vehicle.

We consider only the expenditure proportions, not the values – *compositional data.*

Table 1: Household budget survey data

	H	**F**	**O**	**S**
M1	497	591	153	291
M2	839	942	302	365
M3	798	1308	668	584
M4	892	842	287	395
M5	1585	781	2476	1740
M6	755	764	428	438
M7	388	655	153	233
M8	617	879	757	719
M9	248	438	22	65
M10	1641	440	6471	2063
M11	1180	1243	768	813
M12	619	684	99	204
M13	253	422	15	48
M14	661	739	71	188
M15	1981	869	1489	1032
M16	1746	746	2662	1594
M17	1865	915	5184	1767
M18	238	522	29	75
M19	1199	1095	261	344
M20	1524	964	1739	1410

	H	**F**	**O**	**S**
W1	820	114	183	154
W2	184	74	6	20
W3	921	66	1686	455
W4	488	80	103	115
W5	721	83	176	104
W6	614	55	441	193
W7	801	56	357	214
W8	396	59	61	80
W9	864	65	1618	352
W10	845	64	1935	414
W11	404	97	33	47
W12	781	47	1906	452
W13	457	103	136	108
W14	1029	71	244	189
W15	1047	90	653	298
W16	552	91	185	158
W17	718	104	583	304
W18	495	114	65	74
W19	382	77	230	147
W20	1090	59	313	177

Labour force in Slovenia 62 districts by different employment modes on the 30th of Jun 2002.

Employment modes
1. employed in enterprises, companies
2. employees of the self employed
3. contractors
4. peasants and
5. craftsman

Data source
Statistical Data Base of Statistical office of the Republic of Slovenia at the
`http://www.gov.si/zrs/`
and the self constructing tables routine.

Employment modes as proportions – *compositional data.*

Table 2: Data on Labour force in Slovenia 62 distrits by different employmen modes

districts	1.	2.	3.	4.	5.
AJDOVSCINA	0.76006	0.09351	0.07893	0.00365	0.06385
BREZICE	0.66952	0.12339	0.07542	0.00435	0.12733
CELJE	0.84125	0.08562	0.05248	0.00614	0.01451
CERKNICA	0.82552	0.04829	0.05802	0.00284	0.06533
CRNOMELJ	0.80114	0.06576	0.05150	0.00300	0.07860
DOMZALE	0.77614	0.11552	0.07911	0.00836	0.02087
DRAVOGRAD	0.75891	0.09571	0.08771	0.00200	0.05567
GORNJA RADGONA	0.72528	0.10935	0.06035	0.00239	0.10263
GROSUPLJE	0.69324	0.14408	0.09349	0.00397	0.06522
HRASTNIK	0.72123	0.23720	0.03196	0.00442	0.00520
IDRIJA	0.85884	0.05452	0.04619	0.00451	0.03594
ILIRSKA BISTRI	0.71329	0.12338	0.08566	0.00350	0.07418
IZOLA	0.78776	0.11389	0.08641	0.00815	0.00379
JESENICE	0.88318	0.05608	0.05076	0.00586	0.00412
KAMNIK	0.77007	0.09764	0.08976	0.00583	0.03670
KOCEVJE	0.83104	0.08364	0.05542	0.00521	0.02469
KOPER	0.83980	0.06251	0.06417	0.01373	0.01979
KRANJ	0.86506	0.06738	0.04220	0.00532	0.02005
KRSKO	0.73547	0.11438	0.05861	0.00330	0.08824
LASKO	0.75634	0.11689	0.05988	0.00288	0.06402
LENART	0.59817	0.16207	0.06501	0.00447	0.17027
LENDAVA	0.66455	0.10245	0.05712	0.00239	0.17349
LITIJA	0.63791	0.12097	0.09354	0.00582	0.14176
LJ BEZIGRAD	0.92779	0.03694	0.03040	0.00285	0.00202
LJ CENTER	0.92625	0.02122	0.01114	0.04034	0.00105
LJ MOSTE-POLJE	0.87994	0.06415	0.05041	0.00253	0.00298
LJ SISKA	0.87271	0.06204	0.05691	0.00415	0.00419
LJ VIC-RUDNIK	0.83836	0.07752	0.05918	0.00451	0.02043
LJUTOMER	0.63974	0.13506	0.06269	0.00296	0.15954
LOGATEC	0.70203	0.15723	0.09153	0.00440	0.04480
MARIBOR	0.84116	0.07839	0.05577	0.01078	0.01390

districts	1.	2.	3.	4.	5.
MB PESNICA	0.74115	0.09642	0.07205	0.00104	0.08934
MB RUSE	0.78544	0.10331	0.09126	0.00077	0.01923
METLIKA	0.73818	0.12326	0.06395	0.00355	0.07106
MOZIRJE	0.64472	0.15245	0.07531	0.00509	0.12243
MURSKA SOBOTA	0.68694	0.09101	0.04737	0.00454	0.17014
NOVA GORICA	0.81648	0.07525	0.06473	0.00709	0.03646
NOVOMESTO	0.79744	0.09372	0.05018	0.00418	0.05448
ORMOZ	0.68751	0.06113	0.04075	0.00158	0.20903
PIRAN	0.77252	0.11619	0.08517	0.01045	0.01567
POSTOJNA	0.82727	0.08400	0.05878	0.00450	0.02545
PTUJ	0.68960	0.12679	0.06839	0.00457	0.11066
RADLJE	0.73617	0.11813	0.06773	0.00177	0.07620
RADOVLJICA	0.78275	0.10440	0.07675	0.00818	0.02791
RAVNE	0.83196	0.09443	0.04646	0.00386	0.02329
RIBNICA	0.69559	0.13173	0.09794	0.00344	0.07131
SEVNICA	0.60704	0.18470	0.05846	0.00233	0.14747
SEZANA	0.77001	0.11388	0.07487	0.00549	0.03575
SL GRADEC	0.82423	0.07455	0.05051	0.00382	0.04689
SL BISTRICA	0.76222	0.04462	0.10157	0.00370	0.08789
SL KONJICE	0.74258	0.13533	0.05405	0.00395	0.06409
SENTJUR	0.55387	0.18376	0.09517	0.00439	0.16281
SKOFJA LOKA	0.83400	0.06762	0.05166	0.00589	0.04083
SMARJE	0.63434	0.13448	0.07070	0.00307	0.15741
TOLMIN	0.72281	0.11944	0.08662	0.00516	0.06597
TRBOVLJE	0.88728	0.06338	0.04048	0.00419	0.00468
TREBNJE	0.67363	0.17108	0.08062	0.00272	0.07196
TRZIC	0.76345	0.11893	0.09031	0.00394	0.02337
VELENJE	0.88145	0.07137	0.03440	0.00508	0.00769
VRHNIKA	0.71239	0.14703	0.09450	0.00932	0.03676
ZAGORJE	0.63180	0.26759	0.05422	0.00466	0.04174
ZALEC	0.44898	0.44898	0.05376	0.00280	0.04548

3 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 constant sum.

A suitable (one of) sample space 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*

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

The *perturbation operation*

$$\mathbf{x} \circ \mathbf{y} = \mathcal{C}\left(x_1\, y_1, \ldots, x_D\, y_D\right) \quad \text{defined on } \mathcal{S}^d \times \mathcal{S}^d$$

and the *scalar (power) multiplication*

$$\alpha \diamond \mathbf{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha) \quad \text{defined on } \mathbb{R} \times \mathcal{S}^d$$

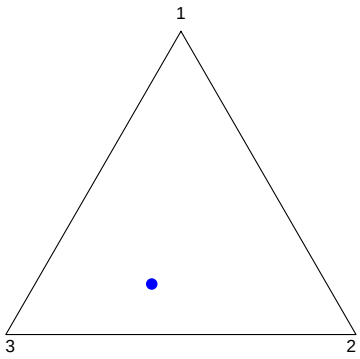

Figure 1: Ternary Diagram – graphical representation of three part compositions $x = (0.17, 0.33, 0.50)$.

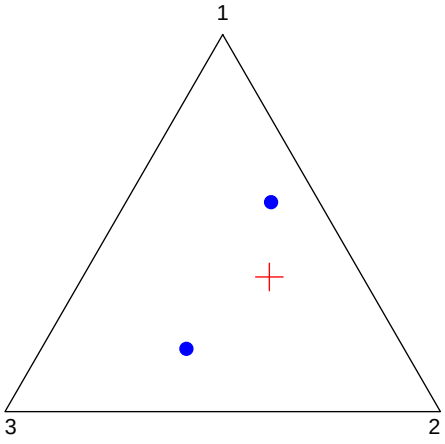

Figure 2: Perturbation operation of compositions $x = (0.17, 0.33, 0.50)$ and $y = (0.56, 0.33, 0.11)$ is $x \circ y = (0.36, 0.43, 0.21)$.

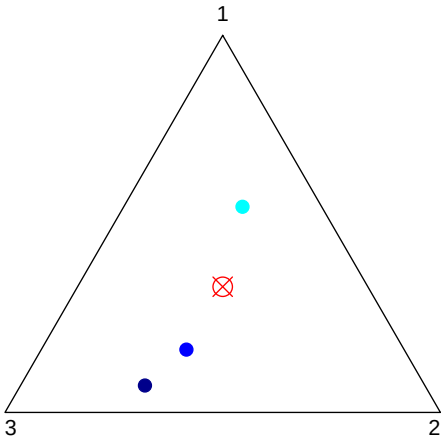

Figure 3: The scalar (power) multiplication of composition x: $2 \diamond x$, $-1 \diamond x$.

induce a vector space structure in to the unit simplex.

$(\mathcal{S}^d, \circ, \diamond)$ is a vector space.

The *neutral element* of this vector space is the *barycenter*

$$\mathbf{e}_D := \left(\frac{1}{D}, \cdots, \frac{1}{D}\right) = \mathcal{C}\,(1,\ \ldots,\ 1)$$

and the *inverse element* of a composition $\mathbf{x} \in \mathcal{S}^d$ is

$$\mathbf{x}' := \mathcal{C}\left(\frac{1}{x_1}, \cdots, \frac{1}{x_D}\right) = -1 \diamond \mathbf{x} =: -\mathbf{x}.$$

Measure of difference for compositional data

$$d\colon \mathcal{S}^d \times \mathcal{S}^d \to \mathbb{R}$$

must verify the *perturbation invariance* condition

$$d(\mathbf{z} \circ \mathbf{x}, \mathbf{z} \circ \mathbf{y}) = d(\mathbf{x}, \mathbf{y}) \quad \text{for all } \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^d.$$

Aitchison proposed the following distance

$$d_{\mathrm{A}}(\mathbf{x}, \mathbf{y}) := \sqrt{\sum_{i=1}^{D} \left(\log \frac{x_i}{g(\mathbf{x})} - \log \frac{y_i}{g(\mathbf{y})}\right)^2}$$

where

$$g(\mathbf{x}) := \sqrt[D]{\prod_{i=1}^{D} x_i}$$

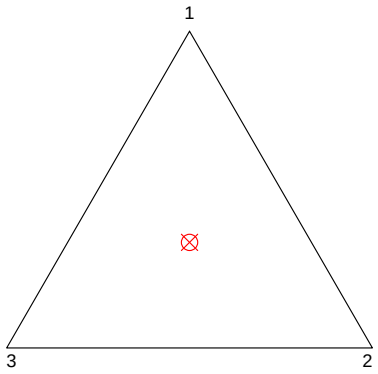

Figure 4: Ternary Diagram with the barycenter.

is the *geometric mean* of the composition $\mathbf{x}$.

The *Aitchison distance* d_{A} is perturbation invariant and therefore we can define a norm by

$$\|\mathbf{x}\|_{\mathrm{A}} := d_{\mathrm{A}}(\mathbf{x}, \mathbf{e}) \qquad \text{for all } \mathbf{x} \in \mathcal{S}^d.$$

The parallelogram identity holds for the $\|\cdot\|_{\mathrm{A}}$ therefore we can define an *inner product:*

$$\langle \cdot, \cdot \rangle_{\mathrm{A}} : \quad \mathcal{S}^d \times \mathcal{S}^d \to \mathbb{R}$$

by

$$\langle \mathbf{x}, \mathbf{y} \rangle_{\mathrm{A}} := \frac{1}{4} \left(\|\mathbf{x} \circ \mathbf{y}\|_{\mathrm{A}} - \|\mathbf{x} \circ (-\mathbf{y})\|_{\mathrm{A}} \right).$$

The $\mathcal{S}^d$ is an unitary space.

The angle: for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}^d$ we have

$$\alpha(\mathbf{x}, \mathbf{y}) := \frac{\langle \mathbf{x}, \mathbf{y} \rangle_{\mathrm{A}}}{\|\mathbf{x}\|_{\mathrm{A}} \, \|\mathbf{y}\|_{\mathrm{A}}}.$$

NOTE
The vector space structure $\left(\mathcal{S}^d, \circ, \diamond\right)$, its algebraic and geometrical concepts: vector, norm, scalar product, and distance play a central role in most of the statistical methods.

Subcomposition

If we are interested only in some of measured properties – only in some part of the composition

$$\mathbf{x} = (x_1, x_2, \ldots, x_D)$$

we just skip the no more observed components and in order to keep the unit sum constraint we divide with the new sum: for the $\mathcal{S} \subset \{1, 2, \ldots, D\}$ and $s := |\mathcal{S}| - 1$ we get the mapping

$$\mathbf{x} \in \mathcal{S}^d \longrightarrow \mathbf{x}_S \in \mathcal{S}^s$$

!

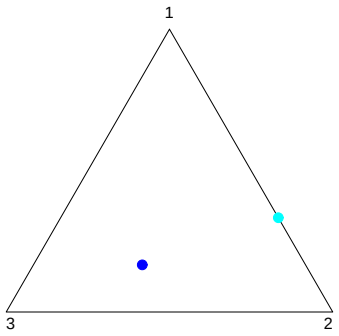

Figure 5: Two part subcomposition.

defined with

$$\mathbf{x}_S := \frac{1}{\sum_{i\in S} x_i}\,(x_1, \cdots, x_s)$$

and we call $\mathbf{x}_S$ the *subcomposition* of the composition $\mathbf{x}$.

And here is the R routine that plots Figure 3

```
> h <- mix.Read("house.dat")
> h4 <- mix.Sub(h,4)
> mix.Ternary(h4)

> mix.Ternary(h)
```

Centered data set

The *geometric mean* of the set of compositions $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\} \subset \mathcal{S}^d$ is defined

$G(X) := \mathcal{C}\,(g_1, \ldots, g_D)$ where $g_k := \left(\prod_{j=1}^{N} x_{jk}\right)^{1/N}$ is the geometric mean of the components $k = 1, \ldots, D$.

Geometric mean is the adequate measure of central tendency for compositional data:

- $G(\mathbf{y} \circ X) = \mathbf{y} \circ G(X)$ for all $\mathbf{y} \in \mathcal{S}^d$,
- $G(\lambda \diamond X) = \lambda \diamond G(X)$ for all $\lambda \in \mathbb{R}$.

In case that the data set $\mathbf{X}$ is near to the corner – this happens when one of the components of the data set is near to 1 it is very difficult to establish if there are differences between the points.

If we perturb the data set $\mathbf{X}$ by the perturbation $-G(\mathbf{X})$ the result data set are centered, i.e. the center of the set $-G(\mathbf{X}) \circ \mathbf{X}$ is the barycenter of the simplex

$$G(-G(\mathbf{X}) \circ \mathbf{X}) = \mathbf{e}.$$

Now we can observe the real pattern of the data (in Aitchison's geometry).

t

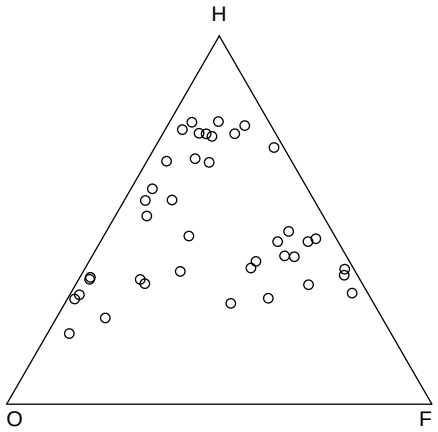

Figure 6: The three part subcomposition of Household data.

NOTE
If we choose a measure of difference which is perturbation invariant and we apply a hierarchic method of classification we obtain the same results for the original data set $\mathbf{X}$ and for the centered data set $-G(\mathbf{X}) \circ \mathbf{X}$.

EXAMPLE: Slovene communities labour force data

This R routine that constructs the three part subcompositions of Slovene communities labour force data in Table 3

```
>   d <- mix.Read("EmploySlo.dat")
>   d124 <- mix.Sub(d,c(5,3))
```

And this R routine plots the Figute 3

```
>     mix.Gmean <- function(m)
>      {
>      if(m$sta>=0){
>       {for (i in 1:ncol(m$mat)) G[i] <- mean.g(m$mat[,i])}
>       mix.Matrix(t(matrix(as.comp(G),byrow = TRUE)),"geometric mean of the data") }
>       }

>      G3 <- rbind(t(matrix(G,byrow = TRUE)),t(matrix(inv.comp(G),byrow = TRUE)),d124$mat)
>      G3 <- mix.Matrix(G3,"Mean and inverze mean and the data")

>    0.878077050 0.11699677 0.00492618
>    0.005354689 0.04018769 0.95445762

>      mix.Ternary(G3,col=c("red","red",rep("blue",nrow(G3$mat)-2)),
>      pch=c(3,4,rep(20,times=nrow(G3$mat)-2)),cex=c(3,3,rep(1,times=nrow(G3$mat)-2)))
```

Table 3: Subcompositions of Slovene communities labour force data

districts	1.	2.	3.
AJDOVSCINA	0.8150777	0.1002788	0.08464343
BREZICE	0.7710433	0.1421004	0.08685638
CELJE	0.8589881	0.08742533	0.05358656
CERKNICA	0.8859127	0.05182276	0.06226458
CRNOMELJ	0.8723214	0.07160279	0.05607578
DOMZALE	0.7995097	0.1189983	0.08149201
DRAVOGRAD	0.8053548	0.1015674	0.0930778
GORNJA RADGONA	0.8103868	0.1221815	0.06743167
GROSUPLJE	0.7447707	0.1547899	0.1004394
HRASTNIK	0.7282283	0.2395016	0.03227012
IDRIJA	0.8950446	0.0568183	0.04813715
ILIRSKA BISTRI	0.7733566	0.1337699	0.09287348
IZOLA	0.7972795	0.1152663	0.0874542
JESENICE	0.892083	0.05664532	0.05127169
KAMNIK	0.8042759	0.1019771	0.09374706
KOCEVJE	0.856654	0.08621792	0.05712813
KOPER	0.8689264	0.06467801	0.06639558
KRANJ	0.8875687	0.06913322	0.04329804
...	...	...	...
...	...	...	...
VRHNIKA	0.7468027	0.1541324	0.09906491
ZAGORJE	0.662535	0.2806074	0.05685763
ZALEC	0.4717564	0.4717564	0.0564872

4 R a free language and environment

R (http://www.r-project.org/) is a free language and environment for statistical computing and graphics. R is similar to the award-winning S system, which was developed at Bell Laboratories by John Chambers et al. It provides a wide variety of statistical and graphical techniques (linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering...).

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

- an effective data handling and storage facility,
- a suite of operators for calculations on arrays, in particular matrices,
- a large, coherent, integrated collection of intermediate tools for data analysis,
- graphical facilities for data analysis and display either on-screen or on hardcopy, and
- a well-developed, simple and effective programming language which includes conditionals, loops, user-defined recursive functions and input and output facilities.

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.

t

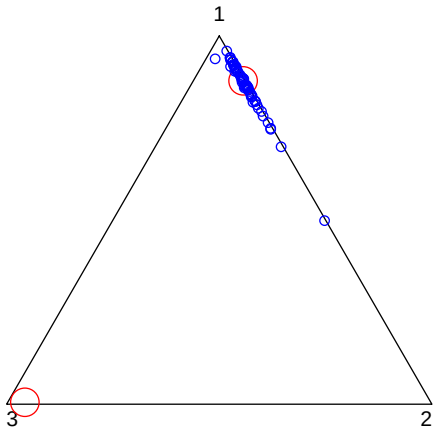

Figure 7: The geometric mean and it's inverse of the three part subcomposition of the Labour data.

The current version of the R library for Compositional Data Analysis is available at

http://vlado.fmf.uni-lj.si/pub/mixture/

```
mix.Check <- function(m, eps=1e-6 )
m$sta <- -2      # matrix contains negative elements
m$sta <- -1      # zero sum row exists
m$sta <- 0       # rows with different row sum(s)
m$sta <- 1       # mixture with constant row sum
m$sta <- 2       # normalized

mix.Read       <- function(file, eps=1e-6)
mix.Normalize  <- function(m)
mix.Random     <- function(nr,nc,s=1)
mix.Ternary    <- function(m, pch = par("pch"), lcex = 1,
               # ternary() is from the on-line answers.
```

5 Conclusions

From the abstract we resume:

What we need is an R library for compositional data analysis comprehending compositional concepts jet not applied originally in R. Programming in R

operations on compositions such as perturbation, power multiplication, subcomposition, distances ...

various logratio transformations of compositions to transform compositions into real vectors that are amenable to standard multivariate statistical analysis,

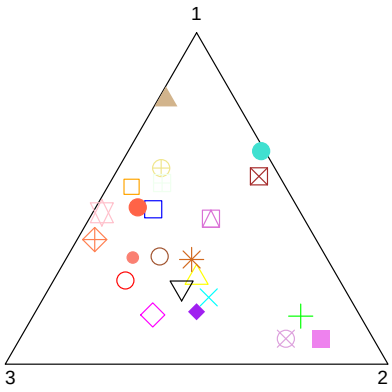

Figure 8: Ternary diagram ploted with R.

compositional concepts such as complete subcompositional independence, the relation of compositions to bases, logcontrast models ... and

graphical presentation of compositions in ternary diagrams and tetrahedrons

will provide an GNU library for compositional data analysis.

And we conclude that we have managed the first and part of the last item. The rest is our goal in future.

6 References

Aitchison, J., 1986, The Statistical Analysis of Compositional Data: Chapman and Hall, New York (USA), 416 p.

Aitchison, J., 1992, On Criteria for Measures of Compositional Difference: Math. Geology, vol. 24, No. 4, p. 365-379.

Aitchison, J., 1997, The one-hour course in compositional data analysis or compositional data analysis is simple, *in* Pawlowsky-Glahn, V., Cimne, eds., Proceedings of IAMG'97 The 1997 Annual Conference of the International Association for Mathematical Geology, Barcelona (E), Part I, pp. 3-35.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V., 1998, Measures of Difference for Compositional Data and HierarchicalClustering Methods, Buccianti, A., Nardi,G. and Potenza, R., eds., Proceedings of IAMG'98. The 1998 Annual Conference of the International Association for Mathematical Geology, Napoli (I), Part 2, p. 526-531.