Compositional Data

Datalab Seminar

Thoralf Mildenberger

Institut für Datenanalyse und Prozessdesign
School of Engineering
Zürcher Hochschule für Angewandte Wissenschaften

06.07.2016

Zürcher Hochschule
für Angewandte Wissenschaften

zh aw School of Engineering
IDP Institut für Datenanalyse und Prozessdesign

Compositional Data

Compositional Data are vectors that add up to a fixed total (usually 1 or 100%).

Components are **proportions**. Amounts are not interesting in themselves but only in relation to each other.

Examples:

- Chemical compositions of substances or objects
- Ingredients in food
- Proportion of money spent on food vs. housing
- relative importance of sectors in different economies

Compositional Data

Example: Mesaurements of chemical compositions consisting of 7 elements plus an "other" category, normalized to one.

```
> load("compo.RData")
> head(compo)
          Al         Si        Ca         Na         C          N         O      Other
1 0.01648873 0.04436536 0.2582444 0.01245552 0.1487544 0.08220641 0.3491103 0.08837485
2 0.01772376 0.02430687 0.3900494 0.01088745 0.1220408 0.06646411 0.2529434 0.11558425
3 0.02372645 0.02779716 0.3085601 0.01209584 0.1325890 0.07443592 0.2585485 0.16224703
4 0.01217391 0.01252174 0.3235942 0.01286957 0.1460870 0.08162319 0.3106087 0.10052174
5 0.01470403 0.01457836 0.3032550 0.01369863 0.1619957 0.08948096 0.3172050 0.08508232
6 0.01470760 0.01470760 0.2396405 0.01272324 0.1651687 0.08591105 0.3865997 0.08054161
> rowSums(compo)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
```

Compositional Data

The operation of normalizing a vector to a fixed sum c (ususally, $c = 1$) is called **closing**.

For $x \in \mathbb{R}_+^D$ we have:

$$\mathrm{clo}(x) = c\left(\frac{x_1}{\sum_{i=1}^D x_i}, \frac{x_2}{\sum_{i=1}^D x_i}, \ldots, \frac{x_D}{\sum_{i=1}^D x_i}\right).$$

Compositional Data

Compositions of D parts are vectors:

$$x = (x_1, x_2, \ldots, x_D) \in \mathbb{R}^D$$

but are subject to **two constraints**:

- $x_i \geq 0$ (or even $x_i > 0$) for all i
- $\sum_{i=1}^{D} x_i = c$.

(live demo)

Compositional Data

Alternatively, one can also define compositions as equivalence classes of vectors, where two vectors $x, y \in \mathbb{R}_+^n$ are in the same equivalence class when there is $\alpha > 0$ such that

$$(y_1, y_2, \ldots, y_n) = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n),$$

i.e. total is not fixed but irrelevant (vectors with different sums but same proportions are equivalent).

In that case, we also have

$$\mathrm{clo}(x) = \mathrm{clo}(y).$$

Compositional Data

Subcompositions: Given a composition, we often want to look only at a subcomposition, i.e. only consider some parts of interest.

We can do so by only using the relevant components and close the resulting vector:

(see live demo)

Most compositions are in fact subcompositions, since we never measure everything (e.g. trace elements).

Actually, this does not matter as we will see.

Compositional Data

Change of units: Change of units is trivial in case we want to change from meters to inches, from grams to ounces etc. as proportions stay the same.

However, sometimes we want to change from mass proportions to volume proportions, from mass to energy content (in food), from proportions of employees in departements to proportions of salaries etc.

Assume we have nutrition data (all measurements originally in g):

```
> nutrition
          Fat        Protein   Carbonates
soy   0.23333333 0.4333333 0.3333333
peas  0.03488372 0.2558140 0.7093023
wheat 0.01190476 0.2738095 0.7142857
corn  0.05555556 0.1111111 0.8333333
beans 0.01176471 0.2823529 0.7058824
```

Compositional Data

Suppose we are more interested in energy intake. We consider the mass proportions for soy:

$$x = (0.2333333, 0.4333333, 0.3333333)$$

If we have a vector giving the energy content in kJ per g for fat, protein, carbonates

$$y = (37, 17, 17)$$

we obtain the energy proportion by componentwise multiplicaton of x and y (and closing the result afterwards):

```
> clo(x*y)
       Fat    Protein Carbonates
 0.3984615  0.3400000  0.2615385
```

Compositional Data

Since we close the result anyway, we might as well close y before componentwise multiplication.

The vector y giving the energy content per gram may thus also be interpreted as a composition.

The operation $\oplus$ defined by

$$x \oplus y = \mathrm{clo}(x_1 y_1, x_2 y_2, \ldots, x_n y_n)$$

is also called **perturbation** of a composition x by a composition y.

Note that

$$x \oplus y = y \oplus x.$$

Compositional Data

The sum and nonegativity constraints imply that...

- all points always lie in some $n-1$ dimensional subspace
- Covariance / correlation matrix is always singular
- Covariance / correlation matrix negatively biased
- Even within the subspace, data are always in a bounded set, so they can e.g. never be normally distributed...

Compositional Data

Correlation between two proportions depend on what else was measured...

(R example)

Can basically give anything!

Compositional Data

Obviously, the constraints make analyzing the data kind of tricky.

Possible quick fix: Kick out one column, breaks linear dependence

But then results depend on which one was removed.

A more principled approach is needed.

Compositional Data: The 4 Axioms

A classical principle in statistics (and other fields of mathematics, physics etc) is **Equivariance**:

A reasonable statistical procedure should be compatible with "natural" transformations that do not change the essential features of the problem.

Example: We measure some temperatures in degrees Celsius. Any reasonable measure of location should give the same answer whether we

1. calculate the measure with the data as given and convert the result to Fahrenheit, or:
2. first convert all the measurements to Fahrenheit and calculate the measure from that.

This is called affine equivariance.

Often the term "invariance" is used although results should indeed change under transformations.

Compositional Data: The 4 Axioms

Four **Axioms** were postulated by Aitchison (to justify his approach?).

Any reasonable procedure for analysis of compositional data should honour the following four principles:

1. Scaling invariance
2. Perturbation invariance
3. Subcompositional coherence
4. Permutation invariance

Compositional Data: The 4 Axioms

Scaling invariance: In compositions, only proportions matter. Totals are irrelevant and may therefore be fixed at a constant value. Any reasonable analysis should respect this and should not give different answers when data vectors are multiplied by a constant.

Also means that it should not depend on the value c to which vectors are closed.

This requirement seems to be rather uncontroversial.

It can be shown that any scaling invariant function of a composition can be expressed as a function of log-ratios of the components.

Compositional Data: The 4 Axioms

Perturbation invariance: It should not matter whether we first perturb our data by a composition and then perform the analysis or whether we perform the analysis with the original data and then perturb the result.

Perturbation means change of units, so the actual requirement is that results should e.g. not depend on whether we perform analysis on mass or volume proportions.

Clearly nice, but it is not clear whether this is really neccessary?

Compositional Data: The 4 Axioms

Subcompositional coherence: Results should not depend on whether we perform an analysis on a subcomposition or on the full composition and form subcompositions from the results.

Another implication is that distance between two compositions must not be smaller than distance only measured on a subcomposition.

A more controversial requirement; classical multivariate approaches used on compositions without transformation generally violate this.

Classical correlation is not subcompositionally coherent.

Compositional Data: The 4 Axioms

Permutation invariance: Results should not depend on the order of components.

Seems obvious, but is violated by some quick fixes like dropping one component to remove linear dependence.

Compositional Data: The 4 Axioms

It can be shown that there is essentially only one type of transformation that is compatible with these requirements.

Essentially, one takes logarithms of the ratios of the components.

Three main types: alr, clr, ilr.

None of these work with zero components.

Compositional Data: The 4 Axioms

alr (Additive Log-Ratio Transform): The oldest version:

$$alr(x) = \left(\log \frac{x_1}{x_D}, \log \frac{x_2}{x_D}, \ldots, \log \frac{x_{D-1}}{x_D}\right).$$

Maps D-dimensional compositions to $\mathbb{R}^{D-1}$.

Fulfills Axioms 1-3 but not 4, as one component plays a special role.

Should not be used anymore.

Compositional Data: The 4 Axioms

clr (Centered Log-Ratio transform):

$$clr(x) = \left(\log \frac{x_1}{\sqrt[D]{x_1 \cdot \dots \cdot x_D}}, \log \frac{x_2}{\sqrt[D]{x_1 \cdot \dots \cdot x_D}}, \dots, \log \frac{x_D}{\sqrt[D]{x_1 \cdot \dots \cdot x_D}}, \right).$$

Maps D-dimensional compositions to $\mathbb{R}^D$.

Fulfills Axioms 1-4.

Main disadvantage: Data lie in $D-1$-dimensional supspace, covariance matrices are singular etc.

Compositional Data: The 4 Axioms

ilr (Isometric Log-Ratio transform): Uses result from clr and projects it to $\mathbb{R}^{D-1}$. Maps D-dimensional compositions to $\mathbb{R}^{D-1}$.

Fulfills Axioms 1-4.

Main disadvantage: Fixes singularity of covariance matrix in clr, but new coordinates are hard to interpret.

Results depend formally on basis chosen, but this does not matter when it is kept constant.

Techniques are available to construct bases that make the transformation more interpretable ("balances").

Compositional Data: General Approach

Priciple of working on the coordinates:

1. Transform compositions to euclidean geometry using clr or (preferably) ilr.
2. Use a classical multivariate technique on the transformed data (e.g. clustering, classification etc.)
3. Transform the results back.

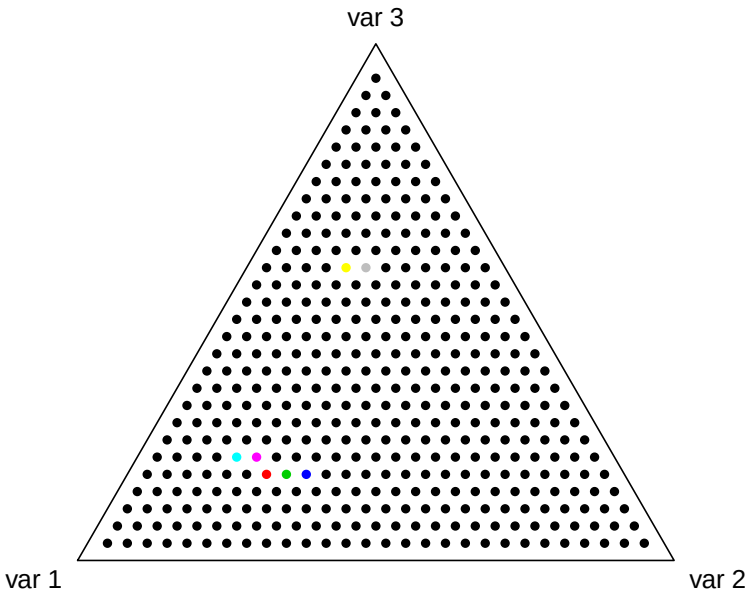
This is compatible with Axioms 1-4 and does also not depend on the basis chosen for ilr as long as the multivariate techniques are affine invariant.

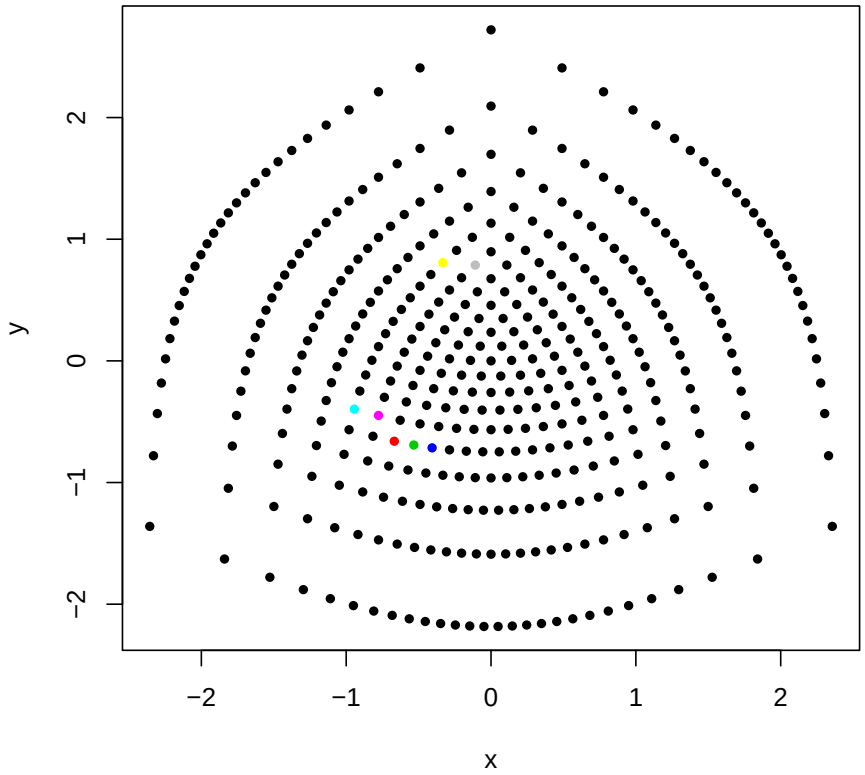
Since we transform back afterwards, interpretability of ilr-coordinates is not directly of interest.

Basically, makes any standard multivariate technique usable for compositional data.

Compositional Data: General Approach

Left: ternary plot, right: ilr transform





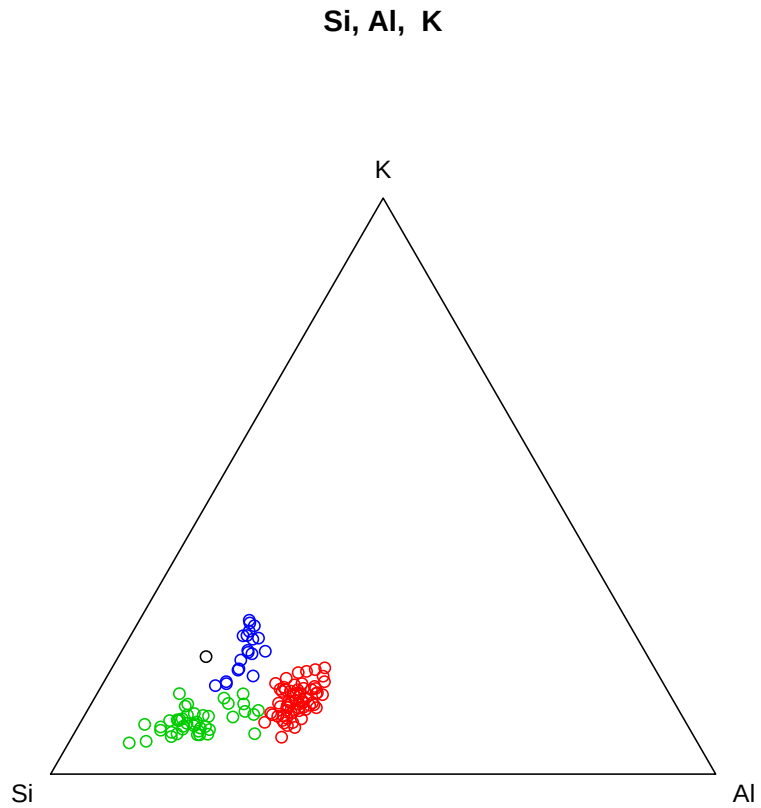
Example: Particle Classifier

We used this approach in the CTI project "Universal Particle Calssifier" for classifying airborne particles based on chemical composition.

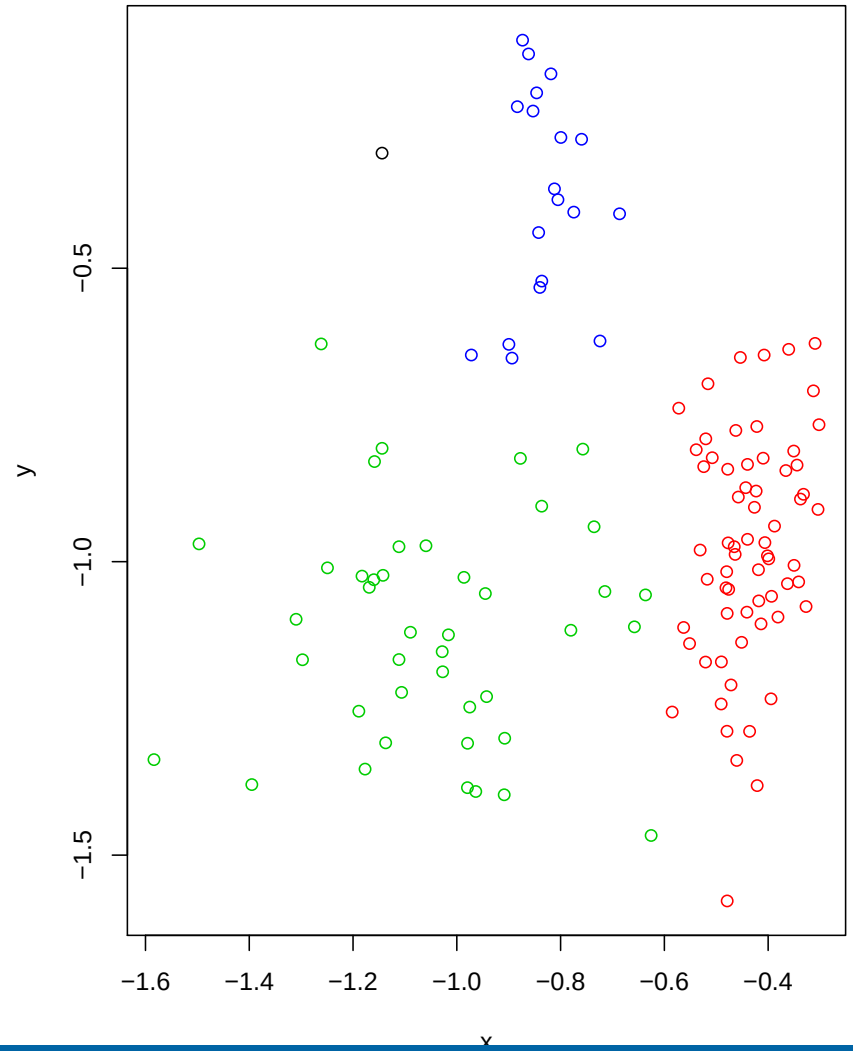
- Use ilr-transform on chemical composition data of particles
- Find ellipsoidal clusters in transformed space or classify new particles accoring to distance to nearest cluster
- transform back and plot in ternary diagrams

Project will be presented in some detail the IDP colloquium in autumn.

Example: Particle Classifier



k=3, outlier correction=FALSE, thresh=0.005



Aitchison geometry

We can take this one step further: The Space

$$A = \left\{ x \in \mathbb{R}^D \middle| \sum_{i=1}^{D} x_i = 1 \text{ and } x_1, \ldots, x_D > 0 \right\}$$

actually is a real vector space in its own right.

Addition and scalar multiplication are given by perturbation and powering, repectively.

These operations correspond to addition and scalar multiplication in the ilr-transformed space.

Aitchison geometry

Perturbation is compositional addition: $(A, \oplus)$ is a commutative group with

$$x \oplus y = \mathrm{clo}\left(x_1 y_1, x_2 y_2, \ldots, x_D y_D\right).$$

Neutral element is $\mathbb{1} = \left(\frac{1}{D}, \ldots, \frac{1}{D}\right)$.

Inverse element of x is given by

$$\ominus x = \mathrm{clo}\left(\left(\frac{1}{x_1}, \ldots, \frac{1}{x_D}\right)\right)$$

with

$$x \oplus (\ominus x) = \mathbb{1}.$$

We write

$$x \oplus (\ominus y) = x \ominus y$$

Aitchison geometry

Powering is scalar multiplication: We define for $a \in \mathbb{R}$:

$$a \odot x := \mathrm{clo}\left(x_1^a, x_2^a, \ldots, x_D^a\right)$$

Interpretation less clear than for permutations, but corresponds to scalar multiplication in ilr-transform space.

Aitchison geometry

Inner product: We can define an inner product by:

$$\langle x, y \rangle_A = \frac{1}{D} \sum_{i>j} \log \frac{x_i}{x_j} \log \frac{y_i}{y_j}$$

Hence, we can define what orthogonality means or what angles are.

Of course, this is also a **kernel** that can be plugged into any kernel method (SVM etc.) to work in ilr-transformed space without actually doing the transformation.

We also get a norm by

$$\|x\| = \sqrt{\langle x, x \rangle_A}$$

and a metric

$$d_A(x, y) = \|x - y\|.$$

Aitchison geometry

Straight lines: What is a straight line in Aitchison geometry?

$$\{x = a \oplus t \odot b \mid t \in \mathbb{R}\}$$

for compositions a and b.

One possible interpretation: We have some composition that changes over time because components grow or decay at dfifferent (but fixed) rates.

E.g. population of biological organism growing at different rates, portfolio with several investments with different returns (which are immediately reinvested in the same type of asset), compound of radioactive substances that decay at different rates etc.

Important: we are still **only** interested in proportions, not total amounts!

Aitchison geometry

Numerical example: We start with (total amounts, not closed yet)

$$a = (10, 50, 100)$$

these components decay and after one unit of time we have the following proportions of the orignal amounts left:

$$b = (0.9, 0.8, 0.9)$$

So after one unit of time we have

$$(a_1 b_1, a_2 b_2, a_3 b_3) = (9, 40, 90)$$

Aitchison geometry

$$a = (10, 50, 100)$$
$$clo(a) = (0.0625, 0.3125, 0.6250)$$
$$b = (0.9, 0.8, 0.9)$$

So after two units of time we have

$$(a_1 b_1^2, a_2 b_2^2, a_3 b_3^2) = (8.1, 32.0, 81.0)$$

If we close this, we get $(0.0625, 0.3125, 0.6250)$.
After t units of time:

$$(a_1 b_1^t, a_2 b_2^t, a_3 b_3^t).$$

Aitchison geometry

Since we are only interested in the proportions and close the final result anyway, we might as well close a and b from the beginning.

Suppose now a and b are closed, then the proportions after $0, 1, 2, ..., t$ are given by

$$
\begin{aligned}
a \\
a \oplus b \\
a \oplus (2 \odot b) \\
\dots \\
a \oplus (t \odot b)
\end{aligned}
$$

which is a straight line in Aitchison geometry (and corresponds to an euclidean straight line in ilr transformed space).

Aitchison geometry

There are also a few standard distributions for this type of data:

- **Normal distribution on the simplex:** distribution that is mapped to multivariate normal under ilr transformation
- **Dirichlet distribution:** Constructed by closing the sum of D independent gamma distributed rvs with same λ but potentially different $\alpha_1, \ldots, \alpha_D$
- **Aitchison distribution:** Family of distributions on the simplex that includes normal and Dirichlet as special cases.

(live demo)

Aitchison geometry

Otehr concepts have their Aitchison equivalents as well:

- geometric shapes like ellipsoids
- summary numbers like expectations, variances etc.

One usually can construct these by applying standard versions in ilr transformed space and then transforming back to the simplex.

Some of these are not very intuitive at first, but usually have some interesting inetrpretations.

Criticism

Main proponents of the approach presented here claim it is the **only** reasonable way to deal with compositional data.

It can be shown it is the only one consixtent with the four axioms, but then not all of them are equally convincing.

Criticism

There are a few equally reasonable requirements the approach does not fulfill:

- **Cannot cope with zeros** (because of logs). Common denfense: In reality, no component is exactly zero but only "below detection limit". Some approaches exist, but none is totally convincing.
- **Amalgamation**, i.e. making a new component from two or more components by summation is **not** compatible with simplex geometry.
- **Mixtures**, i.e. compound made from two ingredients are on a straight line between points given by the two ingredients in standard geometry. In Aitchison geometry they do not.
- Some physical principles like **mass conservation** are violated (which might or might not be relevant depending on the context).

Criticism

Alternatives include:

- Take square roots of components and then use Euklidean distances. This amounts to mapping to the sphere. Has an interpretation of compositions as probability distributions on a discrete set with Hellinger distance

$$d(x,y) = \sqrt{\sum_{i=1}^{D} (\sqrt{x_i} - \sqrt{y_i})^2}.$$

- Do not close and interpret compositions as rays from the origin in $\mathbb{R}^{\mathbb{D}}$. Use cosine-distances (angles between rays).
- Ignore the special structure and hope for the best.

Some selected references

- van den Boogaart, K.G., Tolosana, R., Bren, M. (2014): `compositions`: Compositional Data Analysis. `R` package version 1.40-1.
  `https://CRAN.R-project.org/package=compositions`
  Package for compositional data analysis, provides relevant data types and supports many methods and approaches. A few other packages built on `composition`
- van den Boogaart, K.G., Tolosana-Delgado, R. (2013): *Analyzing Compositional Data with R*, Springer, Berlin and Heidelberg.
  Not only a book on the package but an accesible text book on compositional data analysis. Gives a good overview. Mainly advocates the Aitchison approach but also discusses drawbacks and alternatives.

Some selected references

- Scealy J.L., Welsh, A.H. (2014): Colours and Cocktails: Compositional Data Analysis 2013 Lancaster Lecture, *Australian & New Zealand Journal of Statistics*, 56 (2), 145-169
  `http://onlinelibrary.wiley.com/doi/10.1111/anzs.12073/abstract`
  Very critical review, disputes the claim that the Aitchison approach is the only reasonable one. Discusses history and alternatives.