Reply to "**A Tale of Two Counter Examples**" by M. Greenacre (doc number 16)

A Tale of Two Counter Examples and more

J.A. Martín-Fernández

5 April 2018

1. ***A Tale of Two Counter Examples…..***

This series of letters (from doc number 11 to this one) deals with the advantages and disadvantages (*pros* & *cons*) of the log-ratios of amalgamated parts with respect to the isometric log-ratios (ILR). To illustrate his ideas Michael Greenacre introduced an example (*Michael's data*, Table 1). In short, he argued that the ILR ($\propto$ln(spirits/g(beer,wine)) creates an artificial correlation with the log-ratio ln(wine/beer). This artificial correlation doesn't exist when using the log-ratio with the amalgamated parts (ln(spirits/(beer+wine)). He launched the challenge to provide an example where the effect was the opposite. I replied to the initial ideas about the log-ratio with amalgamated parts (docs number 13 and 15) and prepared a simple data set where the effect is the opposite (*Martin's data, Table 1*).

Table 1 Two data sets used in the series of letters

Michael's data

	wine	beer	spirits
[1]	1.2	1.8	0.6
[2]	1.8	1.4	0.3
[3]	2.8	0.3	0.6
[4]	2.6	0.5	0.5
[5]	2.5	0.7	0.7
[6]	3.0	0.1	0.3
[7]	1.1	1.9	0.3
[8]	1.9	1.2	0.3
[9]	1.4	1.5	0.4
[10]	1.7	1.3	0.2
[11]	1.3	1.7	0.3
[12]	2.4	0.8	0.3
[13]	2.1	1.1	0.5
[14]	1.0	2.1	0.2
[15]	2.0	1.1	0.6
[16]	1.6	1.5	0.4
[17]	2.7	0.4	0.1
[18]	2.3	0.9	0.4
[19]	1.5	1.7	0.6
[20]	2.2	1.0	0.5
[21]	2.9	0.1	0.4

Martin's data

```
### SIMULATION ### EXAMPLE
# means ilr1, ilr2 from M.Greenacre example
set.seed(1)
rilr1=rnorm(50,0.5653601,0.2)
rilr2=rnorm(50,-1.010777,0.06)
############ ilr data set
rilr=cbind(rilr1,rilr2)
cov(rilr)
cor(rilr) # independent ILRs
########## contrast matrix "wine/beer" and "spirits/(wine&beer)"
FI=rbind(c(sqrt(0.5),-sqrt(0.5),0),c(-sqrt(1/6),-sqrt(1/6),sqrt(2/3)))
########## clr coordinates
rclr=rilr%*%FI
######### raw data
rX=exp(rclr)/apply(exp(rclr),1,sum)
colnames(rX)<-c("wine",   "beer" ,  "spirits")
head(rX,12)
######### spurious correlation produced by the amalgamation
plot(log(rX[,1]/rX[,2]),log(rX[,3]/(rX[,1]+rX[,2])),xlab="log(wine/beer)",
    ylab="amalgamated logratio")
abline(lm(log(rX[,3]/(rX[,1]+rX[,2]))~log(rX[,1]/rX[,2])), col="red")
cor(log(rX[,1]/rX[,2]),log(rX[,3]/(rX[,1]+rX[,2])))# -0.5260762
cor.test(log(rX[,1]/rX[,2]),log(rX[,3]/(rX[,1]+rX[,2])))$p.value # 8.718905e-05
#########  correlation between ilr and amalgamated logratio
plot(log(rX[,3]/sqrt(rX[,1]*rX[,2])),log(rX[,3]/(rX[,1]+rX[,2])),xlab="ILR",
    ylab="amalgamated logratio")
abline(lm(log(rX[,3]/(rX[,1]+rX[,2]))~log(rX[,3]/sqrt(rX[,1]*rX[,2]))), col="red")
cor(log(rX[,3]/sqrt(rX[,1]*rX[,2])),log(rX[,3]/(rX[,1]+rX[,2])))# 0.8645423
cor.test(log(rX[,3]/sqrt(rX[,1]*rX[,2])),log(rX[,3]/(rX[,1]+rX[,2])))$p.value # 4.440e-16
######### correlation between ilr-coordinates
plot(log(rX[,1]/rX[,2]),log(rX[,3]/sqrt(rX[,1]*rX[,2])),xlab="log(wine/beer)",
    ylab="ILR")
abline(lm(log(rX[,3]/sqrt(rX[,1]*rX[,2]))~log(rX[,1]/rX[,2])), col="red")
cor(log(rX[,1]/rX[,2]),log(rX[,3]/sqrt(rX[,1]*rX[,2])))# -0.03908718
cor.test(log(rX[,1]/rX[,2]),log(rX[,3]/sqrt(rX[,1]*rX[,2])))$p.value # 0.7875446
```

Table 2 shows a summary of the Pearson correlation coefficient and its p-value for the different log-ratios considered in each example. I am not including the corresponding scatterplot, they are provided in previous letters. In any case, the values of "r" show that both examples represent the opposite situation.

Table 2 Pearson correlation coefficients (and p-values) for the corresponding log-ratio

Log-ratios	*Michael's data*	*Martin's data*
ln(wine/beer) *vs* ln(spirits/(wine+beer))	**r = 0.00 (p = 0.99)**	**r = -0.53 (p = 8.72e-05)**
ln(spirits/(wine+beer)) *vs* ln(spirits/g(wine,beer))	**r = 0.82 (p = 6.137e-06)**	**r = 0.86 (p = 4.44e-16)**
ln(wine/beer) *vs* ln(spirits/g(wine,beer))	**r = 0.53 (p = 0.01)**	**r = -0.04 (p = 0.79)**

In addition, in my reply (doc number 13) I analysed the "artificial" correlation between ln(wine/beer) *vs* ln(spirits/g(wine,beer)) with Michael's data (r=0.53; p=001). When the ratio spirits to the group wine&beer increases, one has to consider different possibilities (please see my doc number 13). After the analysis of the data, I concluded that the case for this example is "numerator: spirits: =" and "denominator: wine&beer: -". That is, spirits has no important variation, whereas wine&beer decreases. Because the interpretation starts by stating "as the ratio of wine to beer increases", I stated that "wine: increases: +" and "beer: decreases: -". Given that the decrease in beer is stronger than the increase in wine, it holds that "wine*beer" has the trend to diminish. Importantly, all the above discussion is also valid for the log-ratio where the denominator is the "amalgamation" and it applies to the example with Martin's data (doc number 13). That is, both examples suffer the same effect, but they correspond to opposite situations.

2. ***…and more***

In my opinion once one amalgamates some parts, one should be very cautious to preserving in the same study the original parts. This idea originates from the definition of "amalgamation operation", when John Aitchison (1986, page 37) was dealing with the simplex as sample space:

> *Definition 2.9*
> If the parts of a D-part composition are separated into C (≤ D) mutually exclusive and exhaustive subsets and the components within each subset are added together, **the resulting C-part composition** is termed an amalgamation.

That is, the 3-composition [wine, beer, spirits] transforms to the 2-composition [wine+beer, spirits], according the amalgamation operation from the simplex S^3 to the simplex S^2 (Aitchison 1986, Property 2.8, page 38). More recently van den Boogaart and Tolosana-Delgado (2013, page 19) describe the issue more accurately:

> Amalgamation, though very commonly done, is a quite dangerous manipulation: a fair amount of information is lost in a way that we may find ourselves unable to further work with the amalgamated dataset. For instance, if saturated and unsaturated fats have different energy content, we cannot compute the total fat mass proportion from the total fat mass energy content, or vice-versa. Amalgamation thus should only be applied in the "definition of the problem" stage, when choosing which variables will be considered and in which units.

Table 3 shows one example to illustrate how the amalgamation can be considered a "point of no return". For example, consider that one wants to calculate the centre of Michael's data set to evaluate in average the spirits consumption with respect to the low alcohol group consumption (wine+beer). One has two different options. On one side, we can compute the centre (geometrical mean) of the 3-compositional parts [wine, beer, spirits], namely [g_{wine}, g_{beer}, $g_{spirits}$] and afterwards make the amalgamation [g_{wine}+g_{beer}, $g_{spirits}$]. On the other side, we can amalgamate all the samples of the data set to create the 2-part compositional data set [wine+beer, spirits], and afterwards compute the centre [$g_{wine+beer}$, $g_{spirits}$]. Table 3 shows that one obtains different results.

Table 3 Centres (in %) of the amalgamated Michael's data according to the procedure

	Procedure: (first step) ----> (second step)	wine+beer(%)	spirits(%)
1st	centre of data set ----> amalgamation of the centre	88.17	11.83
2nd	amalgamation of each sample ----> centre of data set	89.34	10.66

Following Aitchison (1986), Egozcue and Pawlowsky-Glahn (2005) and van den Boogaart and Tolosana-Delgado (2013), the 2nd procedure (amalgamation ----> centre) is the more consistent way. In fact, this procedure provides a result also obtained from the log-ratio coefficients with amalgamated parts (Table 4, first row). In contrast, the 1st procedure is consistent with the results obtained with the ILR coordinates (Table 4, last row).

One of the principal ideas in CoDa is to identify any composition with a vector of log-ratio coefficients. For example, alr, clr, or ilr. Michael Greenacre is proposing the use of log-ratios with

amalgamated parts. That is, for example, to identify each 3-composition [wine, beer, spirits] with the 2-vector of log-ratio coefficients (log(wine/beer), log(spirits/(wine+beer)).

Following Aitchison (1986, page 122)

> Any attempt, however, to find a tractable form for the amalgamation **t** of the partition is fraught with difficulty. The reason for this difficulty is simply that there is no way of expressing the logarithm of a sum of components in terms of the logarithms of the components.

I pointed out this idea in my previous letters: the log-ratio coefficients alr, clr and ilr are log-contrasts, but a log-ratio with amalgamated parts is not. Regarding the centre of the compositional data set, one finds some difficulties. For example, suppose that a 3-compositional data data set expressed with the log-ratios with amalgamated parts (log(wine/beer), log(spirits/(wine+beer)) has the centre in the origin of the 2-dimensional real space, i.e. (0, 0). When one back transforms to the simplex S^3 the centre, this data set [wine, beer, spirits] has the centre in the 3-composition [1/4, 1/4, 1/2]. Note that this composition is not the origin or baricentre of the simplex. This difficulty complicates the basic and critical centring operation, both in log-ratio coefficients and as compositions, related to the perturbation operation.

Another simple problem that can be detected is the "permutation invariance" issue, largely analysed in Aitchison (1986). Table 4 and Figure 1 show that the centre of the Michael's data depends on the log-ratio coefficients considered. In other words, one calculates the log-ratio coefficients of each sample; afterwards one computes the arithmetical mean of these coefficients and, finally, one back-transforms the result so as to get the representative of the centre in the simplex. The conclusion is that the centres obtained depend on the log-ratio with amalgamated parts (it is not invariant under permutations) and, in addition, all of them are different from the centre of the data obtained using alr, clr or ilr coefficients. All these log-contrast coefficients provide the same centre, the representative of the geometrical means, that is, the composition that minimises the Aitchison distance to all the samples, related to the concept of total variance. As a consequence, a first simple problem one has to face when using amalgamation and returning to the original parts is the decision about which is the centre of the data set.

Table 4 Centres of Michael's data according to the different log-ratio coefficients considered

Log-ratio coefficients	wine(%)	beer(%)	spirits(%)
(log(wine/beer), log(spirits/(wine+beer)))	61.63	27.71	10.66
(log(wine/spirits), log(beer/(spirits+wine)))	61.12	27.00	11.88
(log(beer/spirits), log(wine/(spirits+beer)))	58.44	29.01	12.55
(any) ALR & CLR & (any) ILR	60.83	27.34	11.83

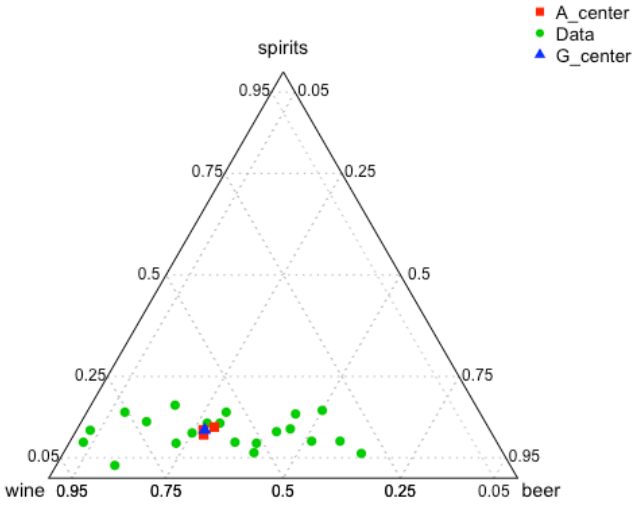

Figure 1 Ternary diagram: Michael's data (green circles) and the centres according to the log-ratio coefficients considered: log-ratio with amalgamated parts (red squares) and ILR (blue triangle).

Once more, thanks a lot for your attention!

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall Ltd. (Re-printed 2003 with additional material by The Blackburn Press), London, UK

Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7):799-832

van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing Compositional Data with R. Springer-Verlag, Berlin Heidelberg, Germany