Reply to "**A Tale of Two Logratios** (is this a counter-example to using ILRs?)" by M. Greenacre

Martín-Fernández

20 February 2018

	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

First of all, thanks to Michael for introducing this discussion and preparing the example. After reading your comments and Alex's ideas, I realize that we need to improve our explanations about CoDa-techniques.

Let us start … my answer to the question in the title "*is this a counter-example to using ILRs?*" is "***absolutely not, this is not a counter-example to using ILRs, this is a perfect example of poor statistical analysis***". In what follows, I justify this answer.

As far as I know, when one starts an analysis the most important thing is the objective, the goal, that is, the "question". In this case, the question given is

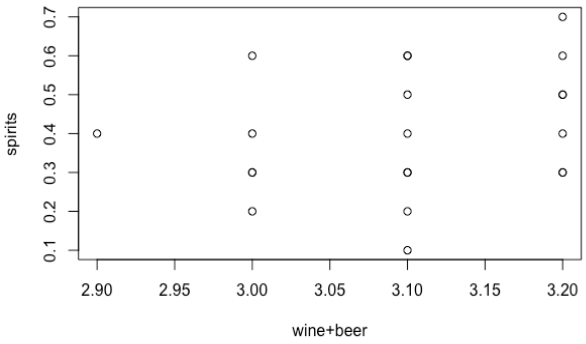
"Researchers are interested in relationships between these variables, and they have heard about CoDa and the ILR transformation. In particular they are interested in how the high alcohol spirits are related to the lower alcohol wine & beer."

The first decision of the researcher is about the "total". They must decide if the "total" is informative or not. That is, we have to decide if the first row (1.2, 1.8, 0.6) is equivalent or not to any composition in its equivalence class, where one representative is (33.33%, 50%, 16.67%). In other words, for instance, do we want to analyse if people who drink a large (absolute) quantity of spirits also drink a large (absolute) quantity of wine & beer? Or, perhaps, do we want to analyse if people who consume high percentage spirits drink more wine than beer? Section A gives the analysis where the "absolute" quantity matters. Section B is "our" CoDa-section.

- **Section A**: "the absolute quantity matters"

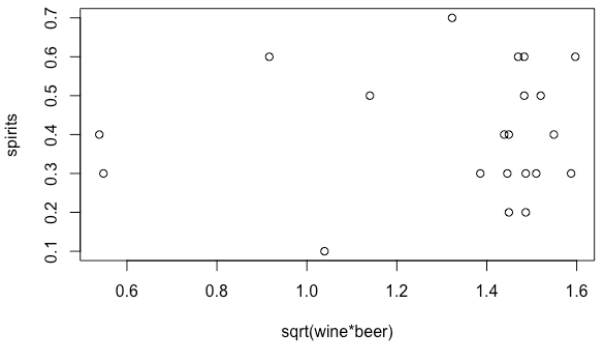
In this section, when one compares row #9: (1.4, 1.5, 0.4) with row #19: (1.5,1.7,0.6) one states that #19 drinks more wine, beer and spirits. We can assume that the sample space of these samples is R3. The Euclidean distance between both is d_e(#9,#19)=0.30. When one compares row #19 with row #16: (1.6,1.5,0.4) one states that #19 drinks more beer and spirits, but less wine. The Euclidean distance between both is again d_e(#16,#19)=0.30. Note that in both cases #19 drinks +0.2 more spirits than #9 and #16.
When we amalgamate wine and beer, we are taking R2 as sample space. Again, we respectively obtain #9*=(2.9,0.4), #16*=(3.1,0.4), and #19*=(3.2,0.6). Now we state that #19* drinks more wine & beer and more spirits than #9*, and #16* as well. However in this case, d_e(#9*,#19*)= 0.36, a larger difference than before, and d_e(#16*,#19*)= 0.22, less than before. That is, when we amalgamate variables, the distances can decrease or can increase despite of working in a sample space of reduced dimension. Note that the concept of variability is related with the Euclidean distance, therefore the amalgamation operation affects the variability.

If the “absolute” quantity matters, the pairwise scatterplots of the raw data describe the relation between pairs of variables in an absolute scale, taking the Euclidean distance in real space as reference for interpretation. The only relation is between wine and beer (r=-0.99), that is “more wine less beer”.



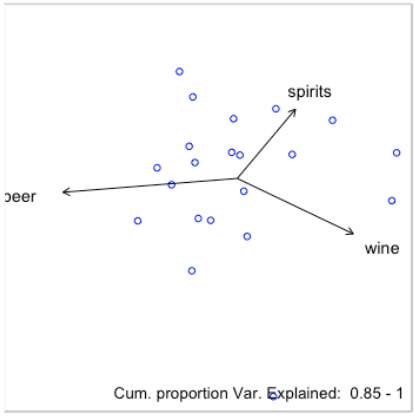
What happens with amalgamated data?

The plot shows no relation between spirits and wine & beer: regardless of the quantity of wine & beer, people drink more or less spirits; there is no pattern in the plot.
Note that the sum wine+beer is proportional to (wine+beer)/2, the arithmetic mean. Therefore the plot “spirits” against “mean(wine,beer)” will show the same pattern.



What happens with the geometric mean?

The plot is not so “nice”, but the interpretation is the same: no relation between the amount of spirits and wine & beer is shown.
The plot is not “nice” because the data set was created with an “additive” pattern (4 levels of sum). However, one can easily create a similar example, where the nice plot is obtained for the geometric mean.



Using PCA for the original (raw) data we get the loadings

	Comp.1 (**85%**)	Comp.2 (≈**15%**)	Comp.3 (≈0%)
wine	**0.721**	0.054	0.690
beer	**-0.692**	0.086	0.716
spirits	0.020	**0.995**	-0.100

PC1 (85%) represents “wine” against “beer”, that is, an ordination of the samples from (small amount of wine & large amount of beer) to (large amount of wine & small amount of wine).
PC2 (15%) represents “spirits”, that is an ordination according the amount of “spirits”.

Conclusion: when "absolute" information matters, the quantity of spirits has no association neither with wine, nor with beer, nor with wine+beer. The quantity of wine has relation with the quantity of beer.

- **Section B**: "the relative information is the issue"

When our goal is the analysis of the relative information, each sample is an equivalence class. A representative of the class can be taken to be proportions, percentages, or the like, to facilitate interpretations. Now, the 3-part simplex can be taken as the sample space, where the representatives of all equivalence classes are. As is well known, this "closure" operation is not essential when working with log-ratio techniques. However, sometimes it is helpful to interpret the results.

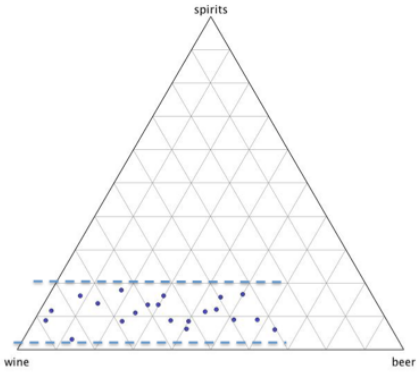
In this section, one can consider that the representatives of rows #9, #16 and #19 in percentages are the values shown in the column "%" of the following table:

#row	raw	%	d_e	d_a
#9	(1.4, 1.5, 0.4)	(42.42%, 45.45%, 12.12%)	d_e(#9,#19)= 0.3	d_a(#9,#19)= 0.25
#16	(1.6,1.5,0.4)	(45.71%,42.86%,11.43%)	d_e(#16,#19)= 0.3	d_a(#16,#19)= 0.33
#19	(1.5,1.7,0.6)	(39.47%,44.74%,15.79%)		

num	%wine	%beer	%spirits
1	33,33%	50,00%	16,67%
2	51,43%	40,00%	8,57%
3	75,68%	8,11%	16,22%
4	72,22%	13,89%	13,89%
5	64,10%	17,95%	17,95%
6	88,24%	2,94%	8,82%
7	33,33%	57,58%	9,09%
8	55,88%	35,29%	8,82%
9	42,42%	45,45%	12,12%
10	53,13%	40,63%	6,25%
11	39,39%	51,52%	9,09%
12	68,57%	22,86%	8,57%
13	56,76%	29,73%	13,51%
14	30,30%	63,64%	6,06%
15	54,05%	29,73%	16,22%
16	45,71%	42,86%	11,43%
17	84,38%	12,50%	3,13%
18	63,89%	25,00%	11,11%
19	39,47%	44,74%	15,79%
20	59,46%	27,03%	13,51%
21	85,29%	2,94%	11,76%

When we compare #9 and #19 one states that #19 drinks RELATIVELY LESS wine and beer and RELATIVELY MORE spirits. The Aitchison distance between both is d_a(#9,#19)=0.25. When one compares row #19 with row #16 one states that #19 drinks RELATIVELY MORE beer and spirits but RELATIVELY LESS wine. The Aitchison distance between both is d_a(#16,#19)=0.33. Note that the comparison using RELATIVE information is different from the ABSOLUTE case (previous Section A) … But anyway … there is nothing unexpected, isn't it?

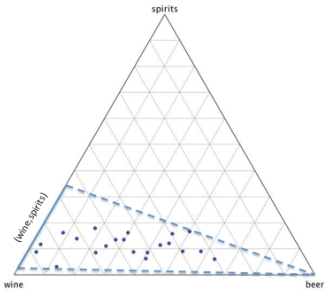
After amalgamation we are working in the 2-part simplex; the parts are sprits and wine & beer.
We then obtain #9*=(87.88%,12.12%), #16*=(88.57%,11.43%), and #19*=(84.21%,15.79%), respectively. Now we state that #19* drinks RELATIVELY LESS wine & beer and RELATIVELY MORE spirits than #9* and #16* as well. Now d_a(#9*,#19*)= 0.22 and d_a(#16*,#19*)= 0.26, which is in both cases less than before. In this example, the Aitchison distances after amalgamation are smaller than without amalgamation. However, this is not true for all CoDa sets, that is, one can find examples where the amalgamation operation "distorts" the Aitchison distances. When we amalgamate parts, the distances (both Euclidean and Aitchison) can decrease or can increase; that is, here we do not have the "dominance property" of distances.



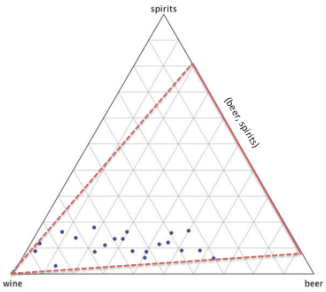
Because we are working with 3-part compositions (equivalence class) we can plot them in a ternary diagram.
One quick look at the diagram suggests that spirits do not vary too much. On the other hand, the variability is mainly due to the relation between wine and beer. This simple plot suggests that no relation exists between the relative consumption of spirits and the relative consumption of wine and of beer. Note that both high and low levels of wine consumption have simultaneously high and low levels of consumption of spirits. The same effect happens with beer and spirits.

But, what about the variability for the subcompositions?

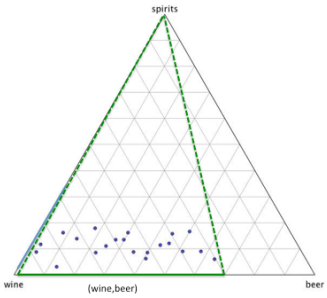
Subcomposition (wine,spirits)



Subcomposition (beer,spirits)



Subcomposition (wine,beer)



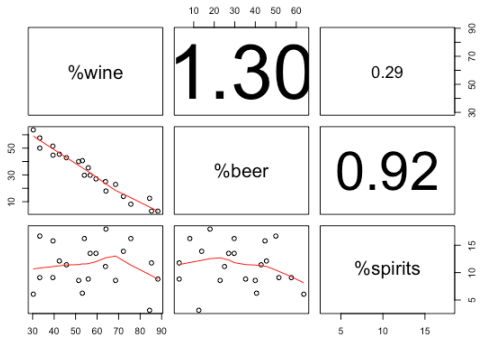
These diagrams suggest that the subcomposition [wine, spirits] has the lowest variability (range), whereas the other two have a similar variability. One must be cautious with this interpretation because we are looking these diagrams with our “Euclidean eyes”. However, in this case, the variation array supports these interpretations.

Variation array:

Variance ln(Xi/Xj)

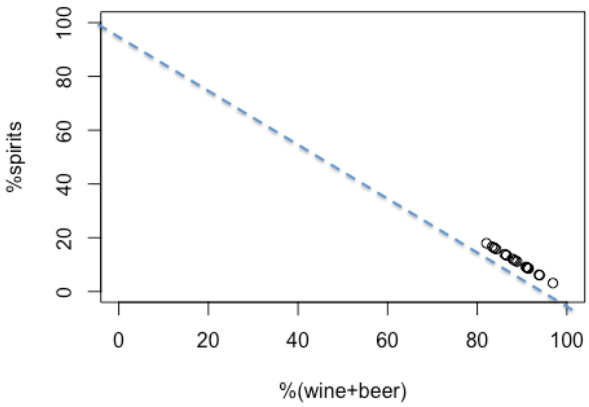
Xi\Xj	wine	beer	spirits	clr variances
wine		1.3046	0.2858	0.2508
beer	-0.7995		0.9236	0.4634
spirits	-1.6377	-0.8382		0.1238
Mean ln(Xi/Xj)				0.8380 Total Variance

Spirits shows the lowest clr-variance (0.1238). The subcomposition (wine, beer) has the largest log-ratio variance (1.3046). No log-ratio variance is “close” to zero, suggesting no-proportionality for the pairs of parts. The negative signs in the log-ratio means suggest that, on average, people drink more wine than beer, more wine than spirits, and more beer than spirits.

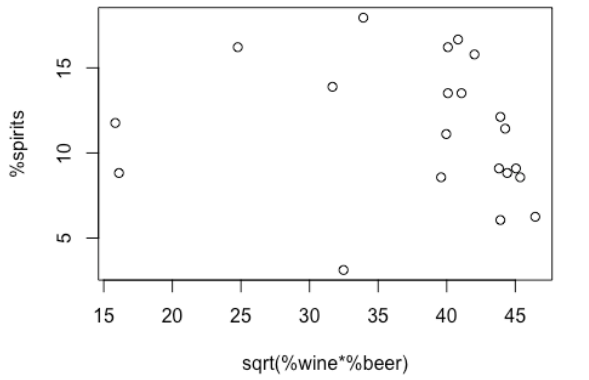


Using a pairwise scatterplot we can investigate how the numerator and the denominator in the log-ratios by pairs are varying. Note that the log-ratio variance of log(wine/beer) is due to a linear relationship between both parts: when one part increases in relative terms, the other diminishes relatively. On the other hand, no relation is suggested between spirits and wine, none between spirits and beer.

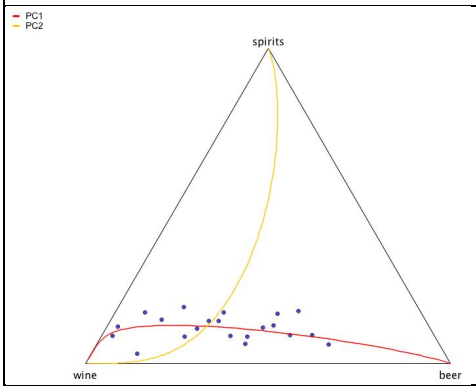
What about the amalgamated data?
Because after amalgamation only two parts remain, we can represent the data on the simplex S^2.



The pattern is as expected, showing that we must avoid the confusion with linear correlation here!!
Obviously, there is only one log-ratio here: log(spirits/(wine+beer)), with mean -2.13 and variance 0.21. We can interpret that people, on average, drink more (wine+beer) than spirits, with a not "very large" variance. Replacing amalgamation with the arithmetic mean (wine+beer)/2 in the denominator, the mean changes to -1.43. Therefore, interpretation is the same: people also drink less spirits than the average of wine and beer. Note that a plot using the arithmetic mean will show again a "perfect" linear relation.



Using the geometric mean of (wine,beer) and forming the corresponding log-ratio coordinate, the results are: mean=-1.01 and var=0.19. That is, the same interpretation as before. However, when we plot spirits against the sqrt(wine*beer) we obtain a very similar pattern as for the "absolute" case (Section A): no relation between the relative spirit consumption and the (wine&beer) consumption.

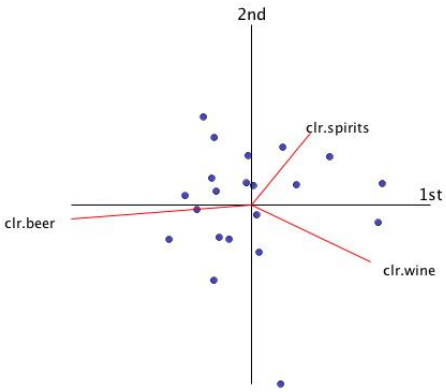


Regarding the log-ratio PCA for the original 3-part CoDa set, one can plot both the PCs and the biplot.

The loadings are:

Principal Components:

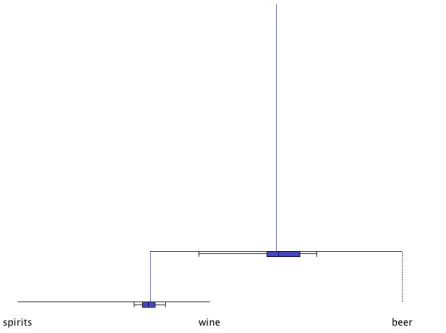
	spirits	wine	beer	Cum.Prop.Exp.
PC1	0.3776	0.4929	0.1296	0.8547
PC2	0.1328	0.5324	0.3349	1.0000



Principal Components:

	clr.wine	clr.beer	clr.spirits	Cum.Prop.Exp.
PC1	0.5342	−0.8019	0.2677	0.8547
PC2	−0.6175	−0.1538	0.7714	1.0000

The loadings suggest that the main variability (85%) corresponds to the log-ratio beer against (wine, spirits). The second (15%) is in the log-ratio spirits against (wine, beer), where beer has the smallest coefficient.



When the optimal algorithm for Principal Balances is applied the result is

```
$bal
           wine       beer    spirits
bal1 0.4082483 -0.8164966  0.4082483
bal2 0.7071068  0.0000000 -0.7071068
```

That is, the SBP suggested is **SBP1**

ILR binary partition:

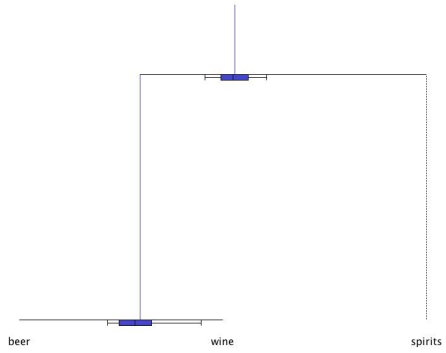
wine	beer	spirits
−1	1	−1
1	0	−1

The corresponding ilr-coordinates have variances:

Variance:

Balance 1	Balance 2
0.6951	0.1429

corroborating the information in the coda-dendrogram: the largest variance corresponds to the log-ratio between beer and (wine, spirits).



However, we are interested in the relation between spirits and (wine, beer); therefore, we define a second SBP, **SBP2**:

ILR binary partition:

wine	beer	spirits
−1	−1	1
1	−1	0

The corresponding ilr-coordinates have the statistics

Mean:

Balance 1	Balance 2
−1.0108	0.5654

Variance:

Balance 1	Balance 2
0.1857	0.6523

Here we can see that people drink relatively more wine&beer than spirits, but the variability is concentrated in the logratio between wine and beer.
We corroborate that people drink “relatively” more (wine, beer) than spirits, and a little bit more wine than beer. Moreover, by far, the largest variability is in the log-ratio between wine and beer. This variability represents 0.6523/0.8380 ≈ 78% of total variance.

I have defined two different log-ratio basis (SBP1 and SBP2), which address different goals (retained variance and spirits *vs* wine&beer, respectively). For each of the two log-ratio coordinate sets we have the corresponding covariance matrix, of course both can be related to the variation matrix. Therefore, we can also consider the correlation matrix for each SBP. In particular for SBP2 we obtain the correlation matrix

Correlation:

	ilr.1	ilr.2
ilr.1	1.0000	0.5291
ilr.2	0.5291	1.0000

The correlation r_{12}=0.5291 is interpreted by M. Greeancre as

> "The conclusion is that as the ratio of wine to beer increases, so the ratio of spirits to the group wine&beer of lower alcohol drinks increases.
>
> Now there is an ongoing debate about using amalgamations rather than geometric means to simplify the interpretation. So, as a check, the ILR balance on the y-axis is plotted against the logratio of spirits divided by the sum (i.e. amalgamation) of wine and beer (Fig. 2), what I call an *amalgamation balance*. The two look very similar, apart from some differences at the upper end, so the researchers feel justified in their conclusions. But just to check even more, they plot the amalgamation balance against log(wine/beer) and get Fig. 3, a surprise! There is no relationship: correlation = 0 (p=0.99)."

Here my detailed answers:

- "as the ratio of wine to beer increases, so the ratio of spirits to the group wine&beer of lower alcohol drinks increases": this is not a surprise, it is natural; it is the value of the correlation matrix, which comes from the variance matrix, i.e., from the variation matrix. HOWEVER, when the ratio spirits to the group wine&beer increases we have to consider different possibilities. The table shows all the different scenarios in which a ratio increases

Numerator: spirits	+	+	++	=	-
Denominator: wine&beer	-	=	+	-	--

  Where the symbols "+" and "-" mean increase and decrease, respectively, and a double symbol means stronger variation. After the analysis of the data, I can conclude that the case for this example is as follows: "numerator: spirits: =" and "denominator: wine&beer: -". That is, spirits has no important variation, whereas wine&beer decreases. Because the interpretation starts by "as the ratio of wine to beer increases", we state that "wine: increase: +" and "beer: decrease: -". Due to the fact that beer diminishes faster than wine increases, it holds that "wine*beer" has the trend to diminish. IMPORTANTLY, all the above discussion about the different scenarios for the ratio, is also valid for the log-ratio where the denominator is the "amalgamation".
- "using amalgamations rather than geometric means to simplify the interpretation": to be honest, I don't see at all any reason why the amalgamation in the denominator "simplifies" the interpretation. As I have shown in my analysis above, the use of amalgamation hides interpretations. In addition, amalgamation affects distances, i.e., variability.
- "...the ILR balance on the y-axis is plotted against the logratio of spirits divided by the sum (i.e. amalgamation) of wine and beer (Fig. 2), what I call an *amalgamation balance*. The two look very similar apart from some differences at the upper end": no surprise here. Note that this correlation is due to the fact that both log-ratios include spirits in the numerator. The correlation between both logratios is r= 0.82 (p-value = 6.137e-06), significant. HOWEVER, if you remove the numerator "spirits" of both logratios then the correlation between log(wine+beer) and log(sqrt(win*beer)) is only 0.18 (p-value= 0.43 !!), not significant. Amalgamation is proportional to the arithmetic mean. Consequently, the arithmetic mean of wine$beer is not related to the geometric mean.
- "so the researchers feel justified in their conclusions. But just to check even more, they plot the amalgamation balance against log(wine/beer) and get Fig. 3, a surprise! There is no relationship: correlation = 0 (p=0.99)": this is not a surprise at all!!! **Correlation is not "transitive"**!!! As already pointed out by McNemar in 1949 [*McNemar, Q. (1949). Psychological statistics. New York: John Wiley and Sons*], this non-transitivity property implies that, given three quantita-

tive random variables X, Y, and Z, a positive correlation between X and Y and a positive correlation between Y and Z (in terms of Pearson's correlation coefficients, not necessarily mean that X and Z will be positively correlated. In fact X and Z might be uncorrelated or even negatively correlated. To understand this effect we have to remember that correlation is the cosine of the angle between the (centred) column vectors. For example, take as X and Z two orthogonal vectors and Y a non-orthogonal one. One can find lot of examples about this issue in the literature.

In our case, X= log(wine/beer) is related to Y= log(spirits/(wine*beer)). In particular, log(wine/beer) is related to log(wine*beer) (r=-0.89, p-value= 4.166e-08), when wine increases, beer decreases and the product diminishes (see explanations above). And log(wine/beer) is not related to log(spirits) (r=0; p-value=0.98),), not a strange result (see ternary diagram above). In addition, the Y= log(spirits/sqrt(wine*beer)) is related to the Z= log(spirits/(wine+beer)), via the spirits in the numerator (see explanation above). However, the X= log(wine/beer) is not related to Z= log(spirits/(wine+beer)). In addition, log(wine/beer) is not related to log(spirits) (r=0, p-value=0.98), and it is not related to log(wine+beer) (r=0.13; p-value=0.58), that is, THE RATIO wine/beer IS NOT RELATED TO THE TOTAL wine+beer (very common case in CoDa, by the way).

I enjoyed a lot working with these data and I find they will be very useful for students in our CoDa-courses. Once more, thanks a lot!