Some thoughts on CodaWork 2017, log-ratio transformations and the practice of CoDa

Michael Greenacre

Background

My whole attitude to CoDa has changed since I started to work with a biochemist on fatty acid (FA) compositions. The only common point we had was the use of ratios, which he could accept once I had convinced him about the principle of subcompositional coherence. I could also convince him about using log-transformed ratios. In almost all papers on FA compositions, tables of means and standard deviations of the raw proportions of each FA are given, which are meaningless since each paper deals with a different subset of FAs that have been closed. I used CLR, explaining to the biochemist that this was just a neat way of getting to all pairwise log-ratios for the log-ratio analysis. Also, I now always use, by default, the weighted form of this analysis, which is identical to Paul Lewi's spectral map, published as long ago as the 1970s, as opposed to the unweighted form (Aitchison & Greenacre, 2002), which CodaWorkers generally use, and which can suffer from serious problems with the low-value parts. See Greenacre and Lewi (2009) for the reasons why.

CodaWork 2017

My "pragmatic approach" to CoDa in my presentation is to base the data analysis on simple ratios, very easy to understand, and -- at least in the FA world -- already a familiar idea, since there are specific ratios that researchers often report in their results. I was surprised to see so many papers using ILR transformation, based on ratios of geometric means. As a univariate concept, this is extremely difficult, if not impossible, to explain to a practitioner. I found that some presentations at the conference talked about a balance as if it were a ratio between amalgamations of parts in the numerator and denominator, which it isn't.

Alecos' paper

His presentation impressed me in several respects. First, he clearly stated the sort of questions he needed answering. Second, he interpreted the CLR transformation as if it was a way to interpret individual parts, which is incorrect, even though he liked the results! Third, he presented a scatterplot of two balances (one of which was a ratio) from an article by Antonella (Buccianti 2015) on the FOREGS data, on which clustering was performed in a quite complex way. After his talk I asked why a simple sum rather than geometric mean could not be used for the denominator, and Juanjo replied something about the "geometry not being obeyed". Since I view a scatterplot of two variables as just that, variable *y* plotted against variable *x*, and I can choose what I like for *y* and *x* according to substantive issue at hand, I firmly believed that this analysis could be simplified without resorting to balances, using two variables in the scatterplot that any geologist would understand quite clearly.

My "pragmatic" alternative on the FOREGS stream water data

Buccianti (2015) reports the following scatterplot of two balances (the Gibbs diagram), where the *x* variable is a simple log-ratio (trivially, a balance). [In investigating the Gibb's diagram in this context, most papers show total dissolved solids, TDS, on a log-scale on the vertical axis, and the ratio of two amalgamations, e.g. (Na+Ca)/(Na+K+Ca) or Na/(Na+Ca), on a linear scale on the horizontal axis]. Antonella is probably trying to "Codafy" the Gibbs diagram in this way, using balances, but the benefit to practitioners is not at all clear.

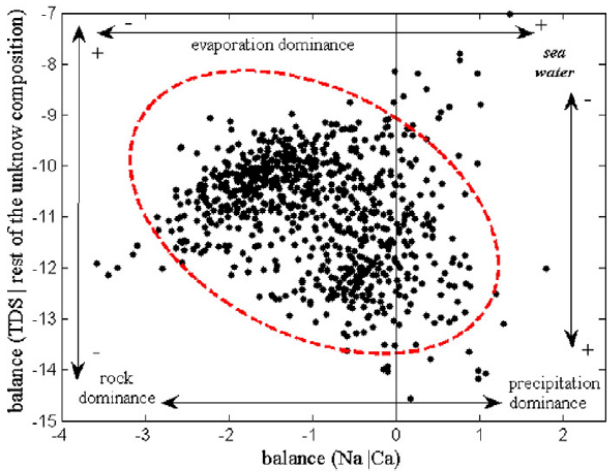
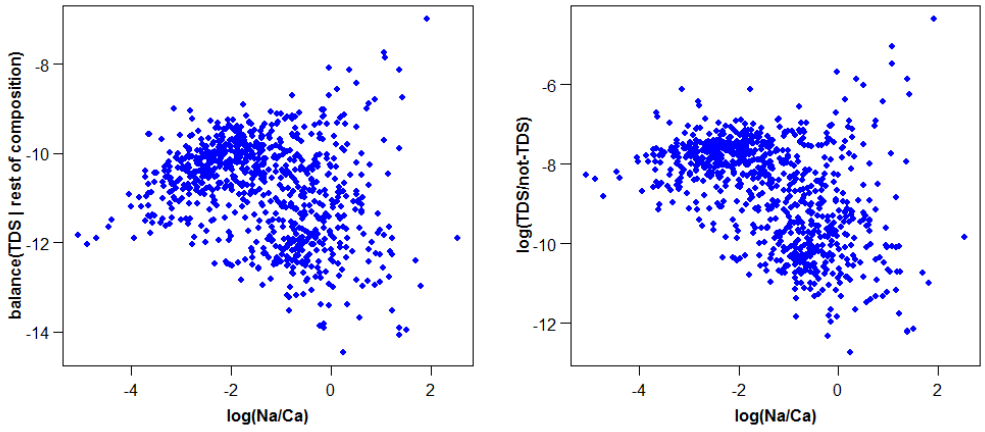


Fig. 6. Gibbs's diagram revised by using isometric log-ratio balances for European stream water samples. The robust confidence ellipse of 97.5% is also reported. Data located outside the ellipse are considered as outliers.

With the data I managed to reconstruct the above scatterplot and, in addition, I redefine the y-variable not as a balance but simply log[TDS/(1-TDS)], which is analogous to a log-odds-ratio, and a quite familiar concept, as opposed to the balance, where the denominator is a geometric mean.

(A: reproduction of scatterplot above) (B: my simpler alternative)



I admit I was quite confused about what the "orthonormality" condition was on balances, but whatever it means, I see no practical reason why this condition should dictate which two variables I use to make a scatterplot in a Gibbs diagram. Clearly, the y-variable is intended to represent total dissolved solids (TDS) somehow as a single concept vertically, and making it in a

form of a ratio is trying to approach Coda principles (but the concept of subcompositional coherence does not seem to be definable when every single part is used in either the balance or the "log-odds-ratio" above). In summary, if I had to choose between A and B above, I would choose B because the practitioner would understand what the two variables are. Since in this particular case TDS is a very tiny part of the whole composition, so that 1-TDS is very close to 1, the log of TDS could be used as the *y*-variable, which is exactly what everyone is doing in the present literature, i.e. representing TDS vertically on a log-scale! As I intimated before, I think that introducing the balance of TDS vs. non TDS [amalgamation in numerator, geometric mean in denominator -- see final remark at the end of this document], to satisfy some obscure (to the practitioner) theoretical principle, is clouding the analysis rather than clarifying it.

If it comes to the cluster analysis in Buccianti (2015), which was using a quite complex combination of methodologies; Mahalanobis distances/robustness/kernel densities, etc..., three clusters were discovered and a classification rule led to 43 misclassifications. From the paper:

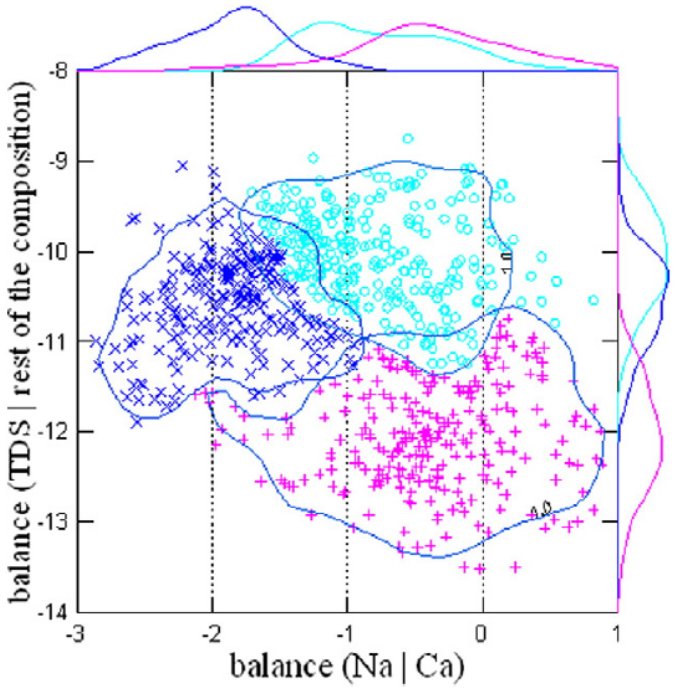
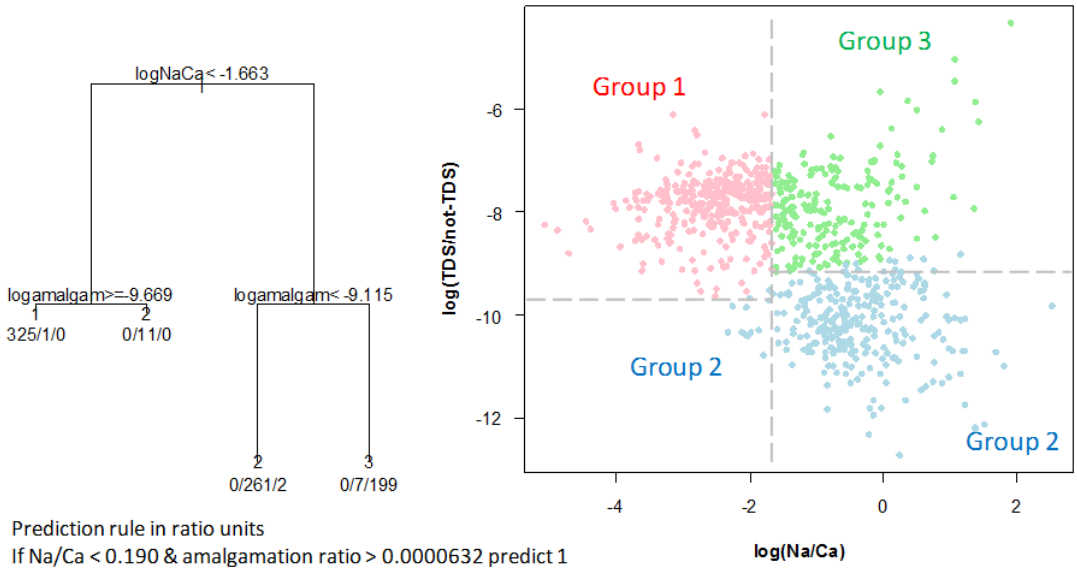


Fig. 10. Gibbs's diagram revised with discrimination of the data by considering the groups obtained by cluster analysis. The continuous lines including the data of the different clusters correspond to the kernel frequency (cluster 1 = light-blue, symbol = o; cluster 2 = blue, symbol = ×; cluster 3 = violet, symbol = +).

Using a much simpler approach, simple Euclidean distance in the log-log scatterplot (version B), for which no further standardization is required, and k-means clustering, in other words a standard procedure in unsupervised learning, three clusters were also identified and a classification rule led to only 10 misclassifications (in the figure below "amalgamation ratio" refers to TDS/(1-TDS), which is the amalgamation of total dissolved solids divided by the amalgamation of all the rest):



Prediction rule in ratio units
If Na/Ca < 0.190 & amalgamation ratio > 0.0000632 predict 1
If Na/Ca < 0.190 & amalgamation ratio < 0.0000632 predict 2
If Na/Ca > 0.190 & amalgamation ratio < 0.0001100 predict 2
If Na/Ca > 0.190 & amalgamation ratio > 0.0001100 predict 3

Only 10/806 misclassified (1.2%) compared to 43/806 in Buccianti (2015)
None misclassified in Group 1, 8 misclassified in Group 2, 2 misclassified in Group 3

The same conclusion would be reached as in Buccianti (2015) in a much simpler way.

Overall conclusion

There is a simpler way to approach compositional data analysis in practice. I believe the fundamental Aitchison principle is that of the ratio. I cannot see why amalgamations are not allowed, and why geometric means have to be used. I see that there is a nice theory behind balances/ILRs, but this theory is out of touch with practical realities. In the case of FA research, biochemists want to amalgamate FAs, e.g. into polyunsaturated FAs, and not combine them in geometric means. This is the practical need.

Final remark about amalgamations at CodaWork 2017

I found it curious, and rather paradoxical, that it is OK to pre-amalgamate parts, e.g. into TDS, and then treat that as a part, rather than make a geometric mean of the parts constituting TDS. So one is not allowed to amalgamate parts, but one can decide to define a part as an amalgamation! (This was also done in the paper by Madalyn Blondes, and no-one objected at all!).

References

Aitchison J, Greenacre MJ (2002) Biplots for compositional data. J R Stat Soc Ser C (Appl Stat) 51(4):375–392

Buccianti A (2015) The FOREGS repository: Modelling variability in stream water on a continental scale revising classical diagrams from CoDA (compositional data analysis) perspective. J Geochem Expl 154:94-104

Greenacre MJ, Lewi PJ (2009) Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J Classif 26: 29-64