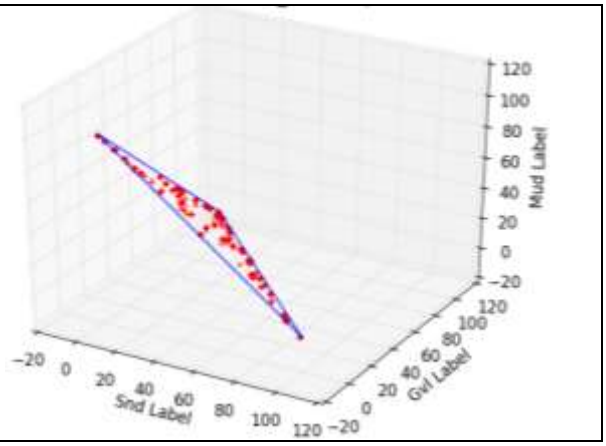
Compositional Data Analysis (CoDA): Handling the Gravel-Sand-Mud Simplexes

dbSEABED Project
October 2020



General.
CoDA data refers to data dealing with fractions – e.g., gravel:sand:mud or mineral components of a sediment. Gravel:sand:mud, of course, is one of the central parameter-groups for seabed characterization.

These sets of values – simplexes - are ratios, not absolute values, and accordingly are often plotted on ternary diagrams. Correct mathematical handling of ratios can be awkward. The usual treatment (Aitchison 1986) is a transformation into lognormal space. We choose the Centered Log-Ratio transformation CLR: $\mathrm{clr}(x) = \left[ln\left(\frac{x_1}{g(x)}\right), ln\left(\frac{x_2}{g(x)}\right), ln\left(\frac{x_3}{g(x)}\right)\right]$ with g(x) being $\sqrt[D]{x_1\, x_2\, x_3}$ where D is the number of fractions in the simplex (3). Conventional math and statistical operations can be performed accurately using CLR(x). The inverse, CLR^{-1}, is $C[\exp(z)]$ – C[] denoting the simplex.

Implementation Issues.
Unfortunately there are some practical problems in the implementation. First, zero abundances (x_i=0) are barred, so a very small epsilon ≈0.001 is used instead, with necessary corrections also to all other x.

Secondly, with legacy (and modern !) seabed data, many samples have just one of the three vectors reported – e.g., just mud %. These cases are too many and too important to ignore for mapping and statistics, so some work-around is needed. Other cases have two of three components reported but perhaps, because of uncertainties, the simplex should not be closed. The Cartesian {x_1,x_2,x_3} plot above shows full simplexes, but in most datasets coming in to dbSEABED, errors and null vectors ('-99' in dbSEABED) cause points to lie off-triangle. Those points are not compliant with CoDA.

To cure these problems: (i) simplexes with one gap are closed to 100%; (ii) simplexes with 2 gaps are filled in withvalues from the closest other full simplex in terms of spatial distance, water depth and the given fraction's value; this is done per mapping project in *dbS_COVERAGES_*.py*; (iii) the infilled values are marked with an "F" in the audit code of their data lines; (iv) the infilled simplexes are transformed to CLR; (iv) the "F" flagged CLR values are masked out, not used, for map compilation/interpolation. At all stages visual inspection of orthogonal data plots (see the figure) are used to inspect the degree of compliance, and to detect problems and null cases.

Conclusion. Unless CoDA methods are used errors of up to 25% are caused in the handling of gravel:sand:mud data, such as with averaging, resampling and interpolating. The present method is a practical implementation that optimizes accuracy by using CoDA, while maintaining the full input dataset of seabed G:S:M.

References.

Aitchison J. (1986), *The Statistical Analysis of Compositional Data*, Chapman & Hall; reprinted in 2003, with additional material, by The Blackburn Press.
CoDaWeb - *Compositional Data Website*. [Available at: "http://www.compositionaldata.com/"]