# Compound Compositional Data Processes

**J. Bacon-Shone[1] and E.Grunsky[2]**

[1] Social Sciences Research Centre, The University of Hong Kong, Hong Kong; *johnbs@hku.hk*

[2]Department of Earth and Environmental Sciences, University of Waterloo, Waterloo, Canada; *egrunsky@gmail.com*

## Abstract

Compositional data is non-negative data subject to the unit sum constraint. The logistic normal distribution provides a framework for compositional data when it satisfies sub-compositional coherence in that the inference from a sub- composition should be the same based on the full composition or the sub-composition alone. However, in many cases sub-compositions are not coherent because of additional structure on the compositions, which can be modelled as process(es) inducing change.

Sometimes data are collected with a model already well validated and hence with the focus on estimation of the model parameters. Alternatively, sometimes the appropriate model is unknown in advance and it is necessary to use the data to identify a suitable model. In both cases, a hierarchy of possible structure(s) is very helpful. This is evident in the evaluation of, for example, geochemical and household expenditure data.

In the case of geochemical data, the structural process might be the stoichiometric constraints induced by the crystal lattice sites, which ensures that amalgamations of some elements are constant in molar terms. The choice of units (weight percent oxide or moles) has an impact on how the data can be modelled and interpreted. For simple igneous systems (e.g. Hawaiian basalt) mineral modes can be calculated from which a valid geochemical interpretation can be obtained.

For household expenditure data, the structural process might be how teetotal households have distinct spending patterns on discretionary items from non-teetotal households.

Measurement error is an example of another underlying process that reflects how an underlying discrete distribution (e.g. for the number of molecules in a sample) is converted using a linear calibration into a non-negative measurement, where measurements below the stated detection limit are reported as zero. Compositional perturbation involves additive errors on the log-ratio space and is the process that does show sub-compositional coherence.

The mixing process involves the combination of compositions into a new composition, such as minerals combining to form a rock, where there may be considerable knowledge about the set of possible mixing processes.

Finally, recording error may affect the composition, such as recording the components to a specified number of decimal digits, implying interval censoring, which implies error is close to uniform on the simplex.

In summary, analysis of compositional data may require accounting for the following sequence of processes:

Underlying generation process -> Measurement error -> Closure -> Compositional perturbation -> Mixing process - >Recording error.

**Key words:** compound compositions, mixing process, closure, compositional data, detection limits

# 1 Introduction

Compositional data is non-negative data subject to the unit sum constraint. (Aitchison 1982) provides a framework based on the logistic Normal distribution, which is appropriate for data that satisfies sub-compositional coherence, i.e., where conclusions about a sub-composition should be the same based on the full composition or the sub-composition alone. This approach has been strongly criticised by (Scealy and Welsh 2014)), who have argued that this principle unreasonably excludes other appropriate models for compositional data. If we could identify a reasonable process that led to the models proposed by (Scealy and Welsh 2011) and (Butler and Glasbey 2008), we would consider their approaches to be reasonable, however they have not proposed any such process. We certainly agree (as, ironically, does Aitchison himself who highlights amalgamation processes in (Aitchison 2005) !) that it is not always the case that the full composition satisfies the principle and it is helpful to consider the complete cycle of processes that yield any specific dataset in order to identify a suitable analysis of the data.

These processes might look like this sequentially:

Underlying generation process -> Measurement error -> Closure -> Compositional perturbation -> Mixing process ->Recording error

In the case of geochemical data, the underlying generation process might be the stoichiometric constraints induced by the crystal sites which ensures that amalgamations of some elements are constant in molar terms, while for household expenditure data, the underlying generation process might include that teetotal households have distinct spending patterns on discretionary items from non-teetotal households. These processes are discussed further in the relevant examples below.

Measurement error is discussed in more detail below – it covers errors introduced in the measurement process.

Closure means imposition of the unit sum constraint through dividing by the total.

Compositional perturbation involves additive errors on the log-ratio space.

Mixing process involves the combination of compositions into a new composition, such as minerals combining to form a rock.

Recording error affects the composition in a linear manner, such as recording the components to a specified number of decimal digits, implying interval censoring, as addressed by (Leung 2015).

# 2 Measurement error

Measurement error affects the raw measurements on the original scale and may reflect an underlying discrete distribution that is converted using a linear calibration into a non-negative measurement, where measurements below the stated detection limit are reported as zero. This can be conceptualised as the measurement process identifying atoms (or molecules) of the specified element, suggesting that the molar fraction is the key determinant for detection, which can be approximated in terms of some scaling of parts per million. If we model this as a Poisson process, this implies that the variance of the measurement process is equal to the mean of the Poisson process, so a mean of 1 has an expected percentage error of 100% (because the mean and standard deviation are both 1), explaining why the “detection limit” may correspond to counts greater than 1 in order to control the percentage error (e.g. a cut-off of 4 corresponds to an expected percentage error of 50%). This non-constant error is contrary to standard Normal distribution assumptions, but for measurements well above the detection limit should map well onto the Lognormal distribution, which is known to approximate the Poisson distribution well as long as the mean is not too small, and hence be consistent with the Logistic Normal distribution for the composition. As will be illustrated in the first example, using a standard multiplicative adjustment for non-detection zeroes may distort any structure in the dataset.

## 3 First expository example

This is an example about mineral composition for samples from a mine recorded as parts per million, with Table 1 showing the minimum, median and maximum compositions.

Table 1 Minimum, median and maximum compositions in parts per million

| Platinum | Gold | Silver | Residual |
|---|---|---|---|
| 0.1 | 1 | 10 | 999,988.9 |
| 1.0 | 10 | 100 | 999,889.0 |
| 10.0 | 100 | 1000 | 998,890.0 |

In this example, the sub-composition of Gold, Silver and Platinum is the same for the 3 samples, but the full composition is very different, because of the large (nearly) constant residual.

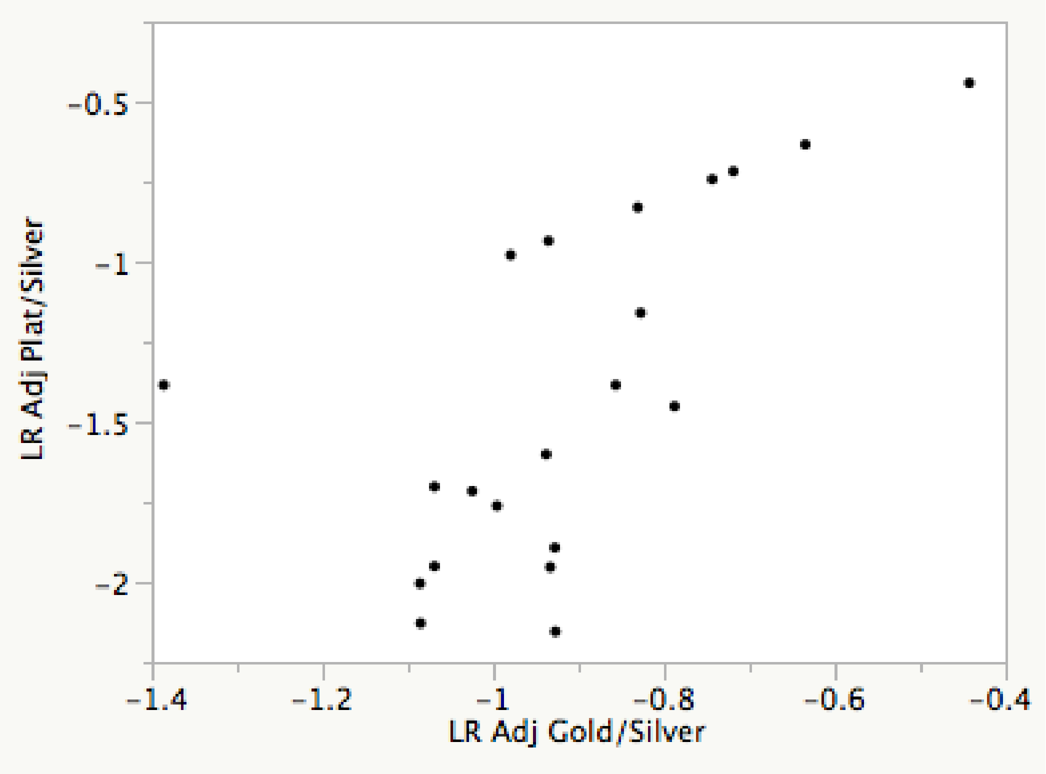
We can see that the important process that determines the viability of these samples for mining is the log-ratio of the aggregate of the three precious metals to the (nearly constant) residual and we can assume the metal ratios are fixed, whether for stoichiometric or other reasons.

However, what we record is perturbed by the measurement process, which is here assumed for the three precious metals to be independent Poisson with the mean in ppm and a detection limit of 5ppm, so any measurements less than 5ppm are rounded down to zero, which corresponds to a percentage error of about 45% at the detection limit, suggesting that a censored Lognormal distribution should be a realistic model.

Clearly, Gold is unlikely to be detected in the minimum sample and may not be detected in the median sample and Platinum will not be detected in the majority of samples.

If we use the standard multiplicative adjustment for zeros as proposed by (MartÌn-Fernandez, Barcelo-Vidal and Pawlowsky-Glahn 2000), (i.e. 0.65 x detection limit), we can see in Figure 1 that there is a distorted picture of the relative proportions, because platinum is measured as a constant for the majority of the samples, rather than adjusting proportional to the other two metals, so the adjustment creates structure that does not exist.

Fig 1 LR Adj Plat/Silver By LR Adj Gold/Silver



This example shows clearly the need to properly model censoring for values below the detection limit rather than making arbitrary zero adjustments, if we are to correctly understand and model the underlying process.

## 4 Second expository example

This is an example about percentages of household expenditure on different categories of expenditure for a limited number of categories, where some households spend no money on alcohol (as they are teetotal, which we assume has no effect on core item spending but is all available for discretionary spending) and others spend no money on white goods (which we assume is occurs 20% of the time for all households and is paid from savings, so it does not affect the relative proportions for the other components). We assume that the core spending is 30% of income, alcohol spending is 5% of income (if not teetotal) and white goods (if any) is 10% of income. Within Alcohol, we can see that the relative proportions relate to whether these are Poor or Rich households, with poor households spending relatively more on Wine and beer. Note that amalgamating alcohol with discretionary spending avoids the zeroes in alcohol expenditure, but this amalgamation loses the information that links the alcohol breakdown to Poor/Rich households.

Table 2 shows four sample observations illustrating the four possible types of expenditures with or without zeroes. There is then some perturbation to reflect random variation, around these values.

Table 2 Household expenditure percentages with presence/absence of zeroes.

| Wine/beer | Spirits | Alcohol | White goods | Core | Discretionary | Poor/Rich |
|---|---|---|---|---|---|---|
| 0.0% | 0.0% | 0.0% | 0.0% | 30.0% | 70.0% | P |
| 4.0% | 1.0% | 5.0% | 0.0% | 30.0% | 65.0% | P |
| 1.0% | 3.5% | 4.5% | 9.1% | 27.3% | 59.1% | R |
| 0.0% | 0.0% | 0.0% | 9.1% | 27.3% | 63.6% | P |

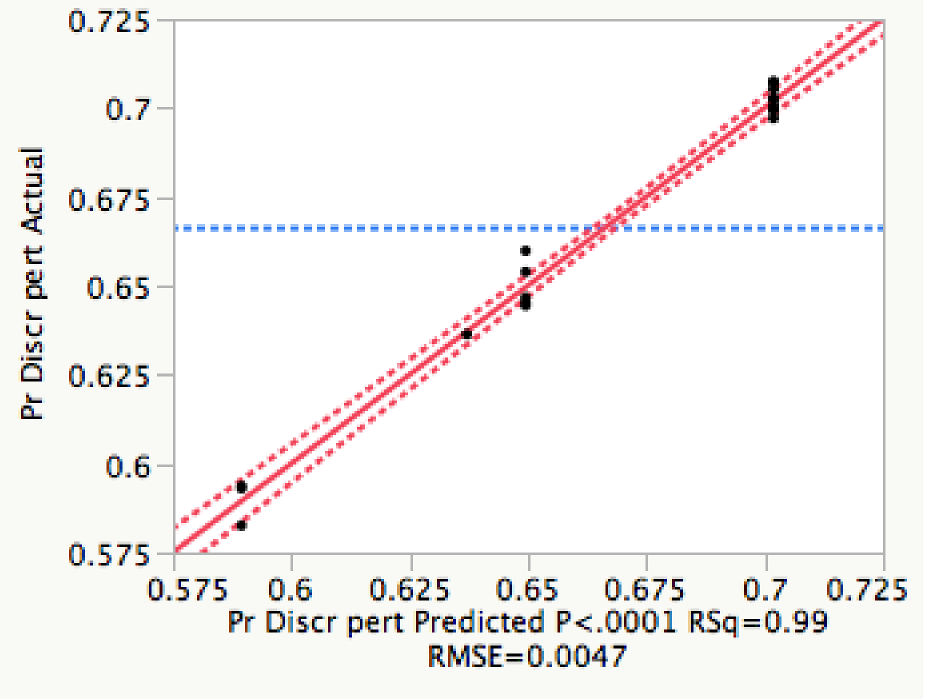
Clearly, the expenditure categories are arbitrary and could be amalgamated, which is contrary to the principles of log-ratio analysis.

In this case, white goods are only purchased rarely, so zeros reflect the rarity of the expenditure rather than failure to ever purchase.

Clearly, any model for this process, must allow for the possibility that teetotal households may have structurally different expenditure patterns, as proposed by (Aitchison 2003) and (Bacon-Shone 2003) and can be handled as a mixture process and we must also examine how white goods purchases affect the relative proportions of other expenses (if at all).

If we have enough data to ensure that the perturbation has little impact, it is easy to show that discretionary expenditure decreases 1% for each 1% in increase in alcohol, but only 0.7% for each 1% increase in white goods (see Figure 2). Similarly, we can see that alcohol does not affect core expenditure, but a 1% increase in white goods yields a 0.3% decrease in core expenditure (Figure 3). Note that these effects are linear on the simplex, not on the log-ratio space. This does not mean that log-ratio modelling is not useful, as it is essential for modelling the perturbation effects.

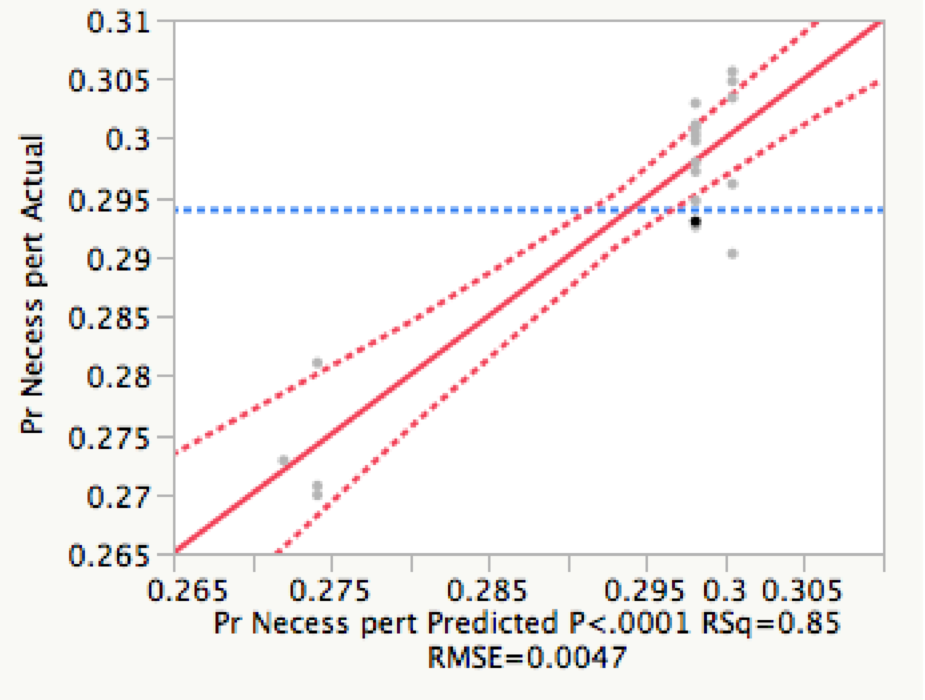
## Fig 2 Pr Discr pert by Pr Alcohol and Pr White Goods

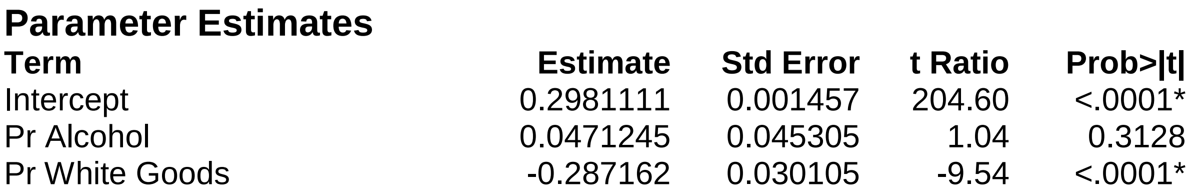


## Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 0.7018889 | 0.001457 | 481.72 | <.0001* |
| Pr Alcohol | -1.047124 | 0.045305 | -23.11 | <.0001* |
| Pr White Goods | -0.712838 | 0.030105 | -23.68 | <.0001* |

## Fig 3 Pr Necess pert by Pr Alcohol and Pr White Goods





Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---|---|---|---|
| Intercept | 0.2981111 | 0.001457 | 204.60 | <.0001* |
| Pr Alcohol | 0.0471245 | 0.045305 | 1.04 | 0.3128 |
| Pr White Goods | -0.287162 | 0.030105 | -9.54 | <.0001* |

## 5 Third expository example

This is an example about mineral composition where there is a stoichiometric constraint (e.g. in a crystal structure, any replacement of one element by another element at a specific lattice site must occur in a integer molar ratio).

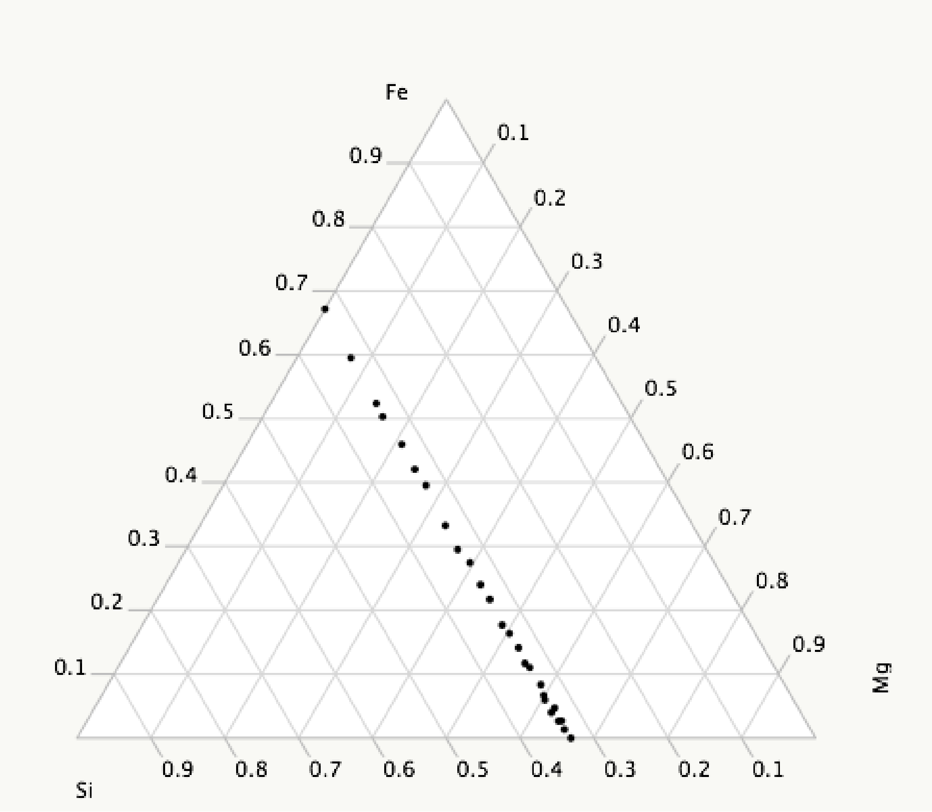
Our simple example, with the minimum, median and maximum data points shown in Table 3, is the mineral olivine in which Fe and Mg substitute into specific lattice sites, with two end-members, fayalite (Fe rich) and forsterite (Mg rich). Si stays close to constant throughout the range of Mg-Fe variability of olivine. Our dataset shows the full range from one end-member to the other. If expressed in weight terms, there is a linear constraint for Mg and Fe, with apparently arbitrary coefficients, while if expressed in molar element terms the constraint is simply Mg + Fe = constant (see Figure 4), while Si varies relative to Mg+Fe and this part of the variability can be easily modelled using log-ratios (see Figure 5).

Table 3 Si,Mg,Fe composition of olivines in molar proportions, showing the end members and median olivine composition

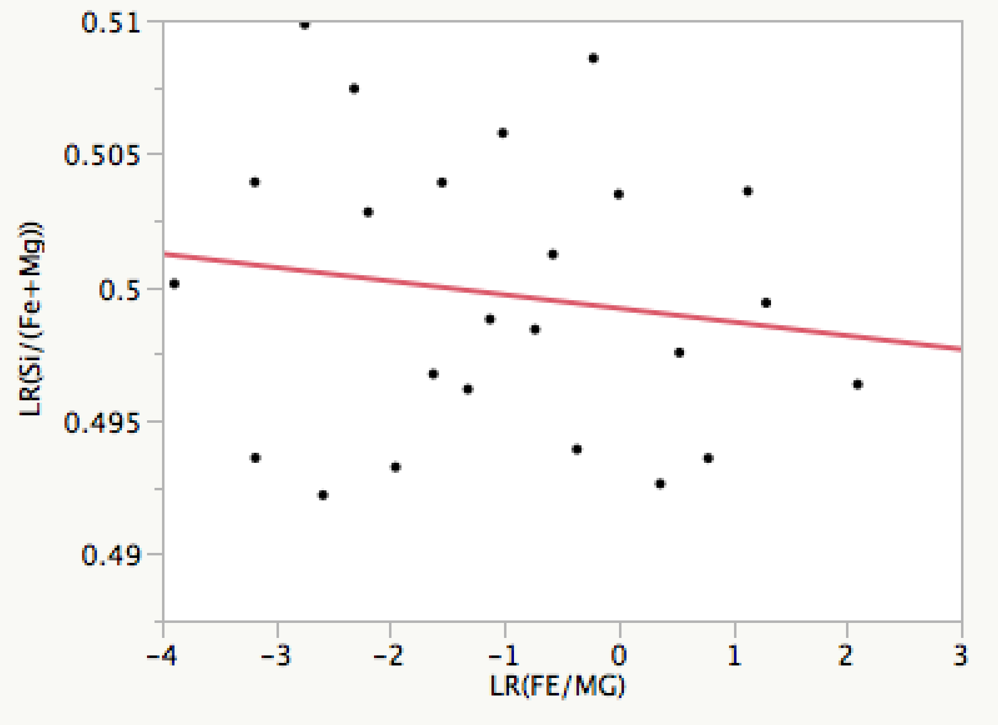
| Si | Mg | Fe | Structural Formula |
|---|---|---|---|
| 0.3333 | 0.3333 | 0.3333 | Olivine |
| 0.3333 | 0 | 0.6667 | Fayalite |
| 0.3333 | 0.6667 | 0 | Forsterite |

As noted in (Grunsky and Bacon-Shone 2011), the linear integer constraints mean that any meaningful analysis must be in molar element terms to see the integer coefficients and must encompass the constraint before modelling the remaining variability in the data, using compositional perturbation, i.e. log-ratio modelling to examine if Si depends on the Mg/Fe ratio.

## Fig 4 Ternary Plot for Molar proportions



## Fig 5 LR(Si/(Fe+Mg)) By LR(FE/MG)



🟥 Linear Fit

## Summary of Fit

| | |
|---|---:|
| RSquare | 0.023011 |
| RSquare Adj | -0.0214 |
| Root Mean Square Error | 0.005377 |
| Mean of Response | 0.499716 |
| Observations (or Sum Wgts) | 24 |

## Parameter Estimates

| Term | Estimate | Std Error | t Ratio | Prob>\|t\| |
|---|---:|---:|---:|---:|
| Intercept | 0.4991969 | 0.001313 | 380.19 | <.0001* |
| LR(FE/MG) | -0.000511 | 0.00071 | -0.72 | 0.4792 |

## 6 Discussion

These three examples all show the need to recognize that although the final dataset we analyse may be compositional data, it is essential in some cases to identify the likely process that generated the dataset, which may include binary, integer and conditional processes, amalgamation and constraints (e.g. stoichiometry), and that we can then avoid the distortions caused by blindly applying zero adjustments and by insisting that all data follow log-ratio processes.

## References

Aitchison, J. 1982. "The Statistical Analysis of Compositional Data." *Journal of the Royal Statistical Society. Series B (Methodological)* 44(2):139-77.

Aitchison, J. 2005. "Some Last Thoughts on Compositional Data Analysis."

Aitchison, J; Kay, J. 2003. "Possible Solutions of Some Essential Zero Problems in Compositional Data Analysis." Paper presented at the CODAWORK 2003.

Bacon-Shone, J. 2003. "Modelling Structural Zeros in Compositional Data Analysis." *CODAWORK03*.

Butler, A and C Glasbey. 2008. "A Latent Gaussian Model for Compositional Data with Zeros." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 57(5):505-20.

Grunsky, Eric and John Bacon-Shone. 2011. "The Stoichiometry of Mineral Compositions." Paper presented at the CODAWORK 2011, Girona, Spain.

Leung, TC; Bacon-Shone J. 2015. "Compositional Data Analysis and the Zero Problem: Interval Censoring Approach." Paper presented at the CODAWORK 2015.

MartÌn-Fernandez, JA, C Barcelo-Vidal and V Pawlowsky-Glahn. 2000. "Zero Replacement in Compositional Data Sets." Pp. 155-60 in *Studies in Classification, Data Analysis and Knowledge Organisation*, edited by H. Kiers, J. Rasson, P. Groenen and M. Shader. Berlin: Springer Verlag.

Scealy, JL and AH Welsh. 2011. "Regression for Compositional Data by Using Distributions Defined on the Hypersphere." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73(3):351-75.

Scealy, JL and AH Welsh. 2014. "Colours and Cocktails: Compositional Data Analysis 2013 Lancaster Lecture." *Australian & New Zealand Journal of Statistics* 56(2):145-69.