Stoch Environ Res Risk Assess (2008) 22:139–141
DOI 10.1007/s00477-007-0193-3

BOOK REVIEW

A. Buccianti, G. Mateu-Figueras and V. Pawlowsky-Glahn (eds): Compositional data analysis in the geosciences: from theory to practice

The Geological Society Publishing House, Special Publication 264, Bath, UK, October 2006, Hardback, pp 224, 150.00 USD, ISBN: 10-1-86239-205-6, 13-978-1-86239-205-2,

John Bacon-Shone

Published online: 26 October 2007

It is already 25 years since the publication of John Aitchison's seminal paper on the statistical analysis of compositional data, which finally provided a solution to the problem identified by Karl Pearson in the nineteenth century of spurious correlation when there is a constant sum constraint. Log-ratio transformations provide a way to "escape" the constraint of a constant sum in a coherent way, which enables the consistent analysis of sub-compositions. It is surprising that such a fundamental advance has still not been fully embraced by scientific disciplines that deal frequently with compositional data, such as geology. Given this lack of adoption of such a fundamental advance, a book like this one is especially invaluable as it provides an excellent balance between the theoretical advances of the 20 years since John Aitchison's book and detailed analysis of practical problems that shows that log-ratio methods provide more meaningful solutions in practice.

Overall, there are four theoretical chapters, eight chapters that apply the methods to real datasets and three chapters, which cover the software available for compositional data analysis. Although the book was initially based on conference contributions, the chapters are quite consistent in notation and level, suggesting hard work by the editors!

The introductory chapter by Pawlowsky-Glahn and Egozcue explains how log-ratio transformations solve the spurious correlations of Pearson and covers the advances since the Aitchison book, including isometric-log-ratio (ilr) transformations, which provide an alternative to the additive-log-ratio (alr) and centred-log-ratio (clr) transformations. However, arguably, the major benefits of the ilr transformations are mathematical (an orthogonal basis) and geometric (neat link with the implied geometry on the simplex) rather than statistical as the results transformed back to the simplex are not affected by the choice of transformation and the ilr transformations can be harder to interpret. Centering the data at the geometric mean, as a special case of the perturbation operation, is shown to help in visualization, together with biplots, which provide a useful graphical summary of the variation in the simplex.

The second theoretical chapter by Egozcue and Pawlowsky-Glahn provides a detailed description of the geometry on the simplex including the distance metric, which is simple for the clr and ilr representation/transformations. While this chapter is interesting and important from a mathematical and geometrical perspective it is not clear that it provides additional statistical or geological insight.

The third theoretical chapter by Martin-Fernandez and Thio-Henestrosa discusses possible solutions to the problem of zero components, treating the zeros as though the true value is missing due to detection limits rather than a complete absence of the component. They examine the possibility of using non-parametric imputation following the methods of Rubin and Little for missing data and conclude that multiplicative adjustments are superior to additive or simple adjustments, although they emphasize the necessity of sensitivity analysis to examine the robustness in the context of the analysis selected. They also claim that artificial correlation is caused when there is more than one component with zeros and non-parametric methods of imputation are used, but it seems that this outcome reflects the use of a specific non-parametric method, rather than a weakness of non-parametric methods

J. Bacon-Shone (✉)
Social Sciences Research Centre,
The University of Hong Kong, Pokfulam Road,
Hong Kong
e-mail: johnbs@hku.hk

in general. One question is why they assume the data is missing, when it is really censored, which suggests that some of the relevant information is being ignored and may explain why the proposed method does not work well under all circumstances.

The final theoretical chapter by Barrabes and Mateu-Figueras considers the simplex from a topological perspective and answers the question as to whether the simplex is an open or closed space from that perspective.

The first software oriented chapter by Thio-Henstrosa and Martin-Fernandez discusses a set of Excel macros developed by the authors called CoDaPack, which facilitates compositional data analysis within Microsoft Excel. The full range of analyses and graphical representations for compositional data are covered, which suggests that this free software is an invaluable tool for those who are already familiar with Excel.

People unable or unwilling to pay the 'Microsoft tax' will find the second software chapter by Van Der Boogaart and Tolosana-Delgado of great interest as they present an open-source (under the GNU public license) package called 'Compositions' that is available as an add-on for the open-source statistical environment "R". For those at home with open-source packages that may need more knowledge in order to install without problems, this approach provides not only financial savings, but the ability to combine these methods with all the other methods and tools included in "R" directly or through other add-on packages. It also allows use of this software on any computer with a standard C compiler.

The final software chapter by Bren and Batagelj presents an alternative package for "R" called "MixeR". This package seems to include similar capability to "Compositions", with the possible addition of attractive 3D display using tetrahedrons.

While choice is good and I applaud the hard work on display in these two "R" packages, as an end-user I would frankly prefer it if these two packages could be combined to provide one even better supported and more capable package!

The first chapter dealing with real datasets is by Kovacs et al., which applies compositional exploratory analysis (biplots and centred ternary diagrams) and linear discriminant analysis (LDA) to a Hungarian vulcanite dataset. This analysis shows two major sub-compositional trends that are consistent with the LDA results and with the general understanding of magma evolution.

Thomas and Aitchison use a similar methodology to investigate the geochemistry of limestones from the Dalradian Supergroup in Scotland. It is interesting to see the full lattice of hypotheses which helps identify the simplest subcomposition that provides good discrimination between the two types of limestone: Inchrory and Dufftown. The Fe_2O_3-MgO-CaO (FMC) subcomposition proves to be good statistically and meaningful geologically, while being different from the siliciclastic composition which Thomas had previously identified as important using non log-ratio methods. The FMC results are then applied using non-parametric statistics to other Scottish limestones and provide further insights consistent with their proposed lithostratigraphical positions.

Gorelikova et al. first apply clr transforms then apply the standard multivariate methods of Ward cluster analysis and hierarchical binary logistic discriminant analysis to the geochemical composition of cassiterite in Asian Russia. Biplots provide additional graphical summaries. The discriminant groups correspond to geodynamic models already accepted but separation using the Be/V ratio also suggests some additional ideas for further investigation.

Reyment looks at using the idea of ridge (shrinkage) estimators, which are known from Campbell's research to work well in multiple regression, discriminant and canonical variate analysis, to compositional data analysis after applying alr or clr transformations. He examines the application to canonical variate analysis of Lithianian Silurian sediments and immunoassay and amino acid analysis of fossil and living brachiopods. In both examples, he concludes that redundant directions of variation affect the reliability of the analysis, suggesting that shrinkage may be helpful. Interestingly, the clr transform seems to yield more reliable data for shrinkage, although it requires handling the singular covariance matrices.

Buccianti et al. look at the analysis of compositional changes in a fumarolic field in Italy. This dataset is interesting as there is also temperature data for each of the seven fumaroles and the data is measured on 10–15 occasions over a 5-year period. The temperatures of the fumaroles each vary randomly over time, but the mean temperatures are significantly different. The alr-transformed data for each location shows that only two log-ratios are time-dependent and only for three fumaroles. However, the small sample size calls into question the elaborate interpretation given to the PCA of these log-ratios given in the paper. In short, log-ratios have facilitated an exploratory analysis, but more data is required before definitive conclusions can be reached.

Weltje takes on a fascinating task in his paper, namely to try and evaluate a popular model (Dickinson Model, or DM) that relates the composition of sandstones to the plate-tectonic setting of the sedimentary basin in which they were deposited. Unfortunately, the raw compositional data is not available in the DM database, requiring him to work with mean compositions that contain many zeros in the standard size part compositions. After addressing zero replacement and sample size weighting, predictive regions are calculated for the four provenance associations. This is

followed by partitioning, using iso-density lines for each pair of predictive distributions, and simulation to assess the efficiency of the partitioning. The analysis showed that the simple additive logistic normal model worked in most but not of all of the regions and the average probability of assigning the correct provenance varied from 65 to 78%, lower than had been estimated previously using inferior models and analysis. Overall Weltje shows that the DM needs better data if it is to move beyond being a successful exploratory tool, but this chapter shows how powerful log-ratio models can be in evaluating geological models.

Daunis-I-Estadella et al. apply the exploratory tools of log-ratio models (i.e., centering and the biplot) to soil compositions from Tuscany, using major oxides and trace elements. They show graphically how the compositional variability relates to the three bedrock types. Finally they illustrate how the PCA axes provide a natural and meaningful connection to linear trends such as weathering.

Buccianti et al. look at the geochemical composition of natural waters in wells in Italy. They look at probability plots of the log-ratios assuming normal distributions, unfortunately without confidence intervals, but it is still clear that some of the distributions are not normal. Indeed the histograms make it clear that while the skew-normal fits better than the normal distribution, some log-ratios clearly follow bimodal distributions. As pointed out by the authors, these results may reflect multiple processes, as the simplex constraint cannot be the explanation for the multimodality displayed. It would be interesting to try and identify an explanation that would remove the multimodality by identifying two (or more) distinct populations. Unfortunately, the ilr transformation chosen serves only to make this more difficult as all except the first component involve more than two chemicals. There is also the strange claim that alr transformations do not allow standard multivariate analysis.

In conclusion, I highly recommend this very useful book to any geologist (or indeed any scientist) interested in how log-ratio methods can facilitate better statistical analysis.