IAMG'06
Int. Assoc. for Mathematical Geology
XIth International Congress
Université de Liège - Belgium

Concepts for handling of zeros and missing values in compositional data

K.G. van den Boogaart[1], **R. Tolosana-Delgado**[2], **M. Bren**[3]
1 Institute for Mathematics and Informatics, Greifswald University, Germany
2 Department of Sedimentology and Environmental Geology, GeoZentrum of the Göttingen University, Germany
3 Faculty of Organisational Sciences, University of Maribor, Slovenia Institute of Mathematics, Physics and Mechanics, Slovenia
Corresponding author: boogaart@uni-greifswald.de

ABSTRACT: In compositional data, such as e.g. geochemical data, missing values are quite common. As for real data, there are several kinds of missing values such as "missing at random", "below detection limit" and "structural zero". Several imputation methods have been proposed to replace zeros, but they all rely on specific models, introduce some sort of randomness or virtual dependency, or depend on arbitrary choices.
This paper puts forward a new strategy to treat zeros and missing values based on the vector space approach for the simplex. The observed variables define an observed subspace, and the unobserved variables specify some qualitative information on the orthogonal complement of the observed subspace. This approach allows a consistent handling, without the usual problems introduced by imputation techniques.
A consequent elaboration of this concept shows how to introduce missing values into graphics, descriptive statistics, statistical computation and statistical models for compositional data. This approach is implemented in the last version of our software, R package 'compositions', and ready for use. This contribution focuses the unbiased estimation of mean and variance.

KEYWORDS: clr, composition, lost values, zero values, 'compositions' R package

1. The types of missing values

There are several types of missing values and zeros typically present in compositional data. A given composition can contain any of the following types of missing values:

BDT *Below Detection Limit*
This is the most-commonly found case. The value of a part is artificially rounded to zero because it falls below the detection limit. To some extent such a censored value is actually missing, since no full quantitative information is available. But on the other side, the information on the existence of the limit can be used to replace the zero by a function of the detection limit (Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn, 2003). However, if the geometry of the analysis is a relative one (like the postulated by Aitchison, 1986, and followers), then BDT can correspond to extremely different ratios. Even a clever imputation replacing the values will yield very different ratios depending on the rest of the data set (Rehder and Zier, 2001).

SZ *Structural Zero*
In some cases it makes no sense to talk about, or to even measure, a specific component (e.g. water content in dried out material, quartz in a foidolite) since it has been removed as an artifact, or it *cannot* be included intrinsically. One could argue that in this case we have two different populations. However, if the part affected by this structural zero is unrelated to the process to be investigated, the relative geometry of the simplex is well able to understand that the observed parts form just a subcomposition of the original one, thanks to its subcompositional coherence (Aitchison, 1997).

MAR *Missing At Random*
If the measurement process just failed at random (i.e., unrelated in any way to the lost variable), we have no data for a part of that composition. This is, at first sight, very near to the structural zero, although the understanding is very different: the former says there exists a unknown true value, and the later neglects the possibility of talking about it.

NMAR *Not Missing At Random*
A part value was lost due to a mechanism which depends on the true value. This is a problematic case in all statistics. It could arise e.g. from measurement procedures with a different probability to fail stochastically dependent to their actual value. Theoretically, an analysis based on such data is only possible with a clear model for the loss mechanism. However, pragmatically we would analyze the data in a MAR fashion and report the NMAR character.

2. Mathematical basic concepts of missing values

An important basic concept for this contribution is that a composition $\mathbf{x} = (x_i)$ (where $i = 1,\ldots,D$ and $\sum_{i=1}^{D} x_i = 1$) viewed in the Aitchison geometry of the simplex (Aitchison et al., 2002) can be isometrically mapped to a subspace of $\mathbb{R}^D$ by the centered log ratio transform $\mathrm{clr}(\mathbf{x}) \in \mathbb{R}^D$:

$$\mathrm{clr}(\mathbf{x})_i = \ln x_i - \frac{1}{D}\sum_{j=1}^{D} \ln x_j \tag{1}$$

The image of this mapping is called the *clr-space*. It is an hyperplane spanned by the non orthogonal, non basis vectors $\mathbf{w}_i = \mathbf{e}_i - \frac{1}{D}\mathbf{1}$, one associated to each part, and the whole hyperplane is orthogonal to the vector $\mathbf{1}$ (the axis of the positive orthant). The components of the spanning vectors sum up to zero (as all clr transformed values do).
According to Aitchison (1986), a composition is (over-)specified by all its pairwise ratios. A BDT, SZ or MAR value in a component means that we do not known *any* of the ratios of this part with any other. Thus we are only informed about a subcomposition. According to Egozcue and Pawlowsky-Glahn (2005), a subcomposition can be seen as a projection P of the clr-transformed composition into the null space of the vectors $\mathbf{w}_i, i \in M$, where M contains the indices of the missing parts. The dimension of this observed subspace is in general $D - 1 - \mathrm{cardinal}(M)$. In other words, one observes only a projection of the true composition, a projection onto the orthogonal complement of the $\mathbf{w}_i$ vectors associated to the lost parts. Note that this is not so different from the classical multivariate case: the added difficulty of the compositional case lies on the fact that the reference axis are neither orthogonal nor a basis, thus one has to work always on the orthogonal complements of the non-observed parts.
According to Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn (2001), compositions are equivalence classes. Such definition can be generalized to accommodate a composition with missing values by a looser equivalence relation:

Definition 1 Composition with missing values
Let M be a set of indices of parts of a composition. Two compositions $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ are called equivalent as compositions of missing values in *M, denoted by $=_{A|M^C}$, if*

$$\mathbf{x} =_{A|M^C} \mathbf{y} :\Leftrightarrow \exists \lambda > 0 : \forall i \notin M : x_i = \lambda y_i$$

Theorem 2 Missing values correspond to projections
Let be P_M the orthogonal projector onto the null space of $\{\mathbf{w}_j, j \in M\}$. Then

$$P_M \mathrm{clr}(\mathbf{x}) = P_M \mathrm{clr}(\mathbf{y}) \Leftrightarrow \mathbf{x} =_{A|M^C} \mathbf{y},$$

In words, this theorem states that two compositions equivalent with missing values on M have the same *clr* transformed values once projected onto the orthogonal directions to M. In short, the idea of this abstract is to represent the information obtained from a composition with missing values by the projected values $P_M\text{clr}(\mathbf{x})$ and the projection mapping P_M itself. So, an efficient way to compute $P_M\text{clr}(\mathbf{x})$ is needed.

Theorem 3 Calculation of $P_M\text{clr}(\mathbf{x})$
Let M^C be the set of indices of observed parts, thus not included in M. Let $\mathbf{x}^*$ be the subcomposition of M^C. For this subcomposition,

$$(P_M\text{clr}(\mathbf{x}))_{i\notin M} = \text{clr}(\mathbf{x}^*)_{i\in M^C}.$$

For the missing parts, $(P_M\text{clr}(\mathbf{x}))_{i\in M} = 0$.

Sketch of Proof: Note that the log-ratios in M^C are preserved, since the clr-transformed composition and the clr-transformed subcomposition do only differ by a constant term (Eq. 1), which is canceled computing the log-ratio. Note also that $\sum_{i=1}^{D} P_M\text{clr}(\mathbf{x}))_i = 0$, thus $P_M\text{clr}(\mathbf{x}))$ belongs to the clr-plane. which is trivial due to the clr on the reduced space and that the whole things is a clr and perpendicular to all $\mathbf{w}_j,\ j \in M$.

3. Unbiased estimation of mean and variance in the presence of missing values

A simple idea is now to use the transformation $P_M\text{clr}(\mathbf{x})$ instead of Eq. (1) itself. If $\mathbf{x}_i$ denotes the i-th composition of a data set,

$$E\left[\sum_i P_{M_i}\text{clr}(\mathbf{x}_i)\right] = \left(\sum_i P_{M_i}\right)\mu$$

Theorem 4 *An unbiased estimation of the mean with missing values in M_i for $\mathbf{x}_i$ is obtained by:*

$$\text{clr}(\hat{\mu}) := \left(\sum_i P_{M_i}\right)^{-} \sum_i (P_{M_i}\text{clr}(\mathbf{x}_i))$$

where A^- denotes the Moore-Penrose Inverse of matrix A.

From now on, the following symbols are used for these two matrices: $N = \sum_i P_{M_i}$ and $S = \sum_i P_{M_i}\text{clr}(\mathbf{x}_i)$. The computation of a variance follows a quite similar strategy, but rather more complex. A first inspection of the traditional estimator leads to very complex expressions for the expectation of an empirical variance. Taking into account that $P_M^t = P_M$, and denoting $\mathbf{v}_i = \text{clr}\mathbf{x}_i$,

$$\begin{aligned} & E\left[\sum_i \left(P_{M_i}(\mathbf{v}_i - N^- S)\right)\left(P_{M_i}(\mathbf{v}_i - N^- S)\right)^t\right] \\ = & \sum_i P_{M_i}\mathbf{v}_i\mathbf{v}_i^t P_{M_i} - \sum_i P_{M_i}\mathbf{v}_i\mathbf{v}_i^t N^- P_{M_i} \\ & - \sum_i P_{M_i}N^-\mathbf{v}_i\mathbf{v}_i^t P_{M_i} + n\sum_i P_{M_i}N^-\mathbf{v}_i\mathbf{v}_i^t N^- P_{M_i} \end{aligned}$$

which differs from univariate problems by the matrix N^-, which is simply n^{-1} in an univariate estimation. Instead, the idea is to start from the paired differences: $E\left[P_{M_i\cup M_j}(\text{clr}(\mathbf{x}_i) - \text{clr}(\mathbf{x}_j))\right] = 0$ and

$$\text{var}\left(P_{M_i\cup M_j}(\text{clr}(\mathbf{x}_i) - \text{clr}(\mathbf{x}_j))\right) = P_{M_i\cup M_j}\text{var}\left[\text{clr}(\mathbf{x})\right]P_{M_i\cup M_j}^t,$$

where A^t denotes the transposed matrix of A. Computation of the mean in this context needs a little technical trick. Let $c(A)$ denote the mapping of a matrix $A \in \mathbb{R}^{p\times q}$ to the vector with the same entries but stored column by column. If $\otimes$ represents the Kronecker product, it holds that $c(P_M A P_M) = (P_M \otimes P_M)c(A)$.

Theorem 5 *Let* $\mathbf{v}_k = c((\mathrm{clr}(\mathbf{x_{i_k}}) - \mathrm{clr}(\mathbf{x_{j_k}}))(\mathrm{clr}(\mathbf{x_{i_k}}) - \mathrm{clr}(\mathbf{x_{j_k}}))^t)$ *and* $M_k = M_{i_k} \cap M_{j_k}$*, and k of* $\{(i,j) : i \neq j\}$*. An unbiased estimation of the clr-variance of a compositional data set in the presence of missing values is given by:*

$$c(\hat{var}(\mathrm{clr}(\mathbf{x}))) := \left(\sum_{i\neq j} P_{M_i\cap M_j} \otimes P_{M_i\cap M_j}\right)^{-} \times$$
$$\times \sum_{i\neq j} c\left(P_{M_i\cap M_j}(\mathrm{clr}(\mathbf{x}_i) - \mathrm{clr}(\mathbf{x}_j))(\mathrm{clr}(\mathbf{x}_i) - \mathrm{clr}(\mathbf{x}_j))^t P_{M_i\cap M_j}\right)$$

The Moore-Penrose-Inverse has to be taken since all the matrices are singular and all vectors orthogonal with respect to all vectors of the form $c(a\mathbb{1} \otimes b\mathbb{1})$.

4. Final words

The treatment of missing values in compositional data is different and theoretically more difficult than in the classical multivariate case, where the missing of values typically occurs parallel to the axis. In contrast, compositional missing values occur along the clr-directions, which are not an orthonormal reference system. Therefore, a proper treatment of missing values must begin during the generation and recording of the data and has to convey each step of the analysis, from the first plot to the last test. The estimation of mean and variance showed in this paper are just a first step. The R package 'compositions' (van den Boogaart and Tolosana-Delgado, 2005; Bren and Batagelj, 2005) contains a much more complete treatment of missing values and zeros based on the ideas presented here.

Acknowledgments: The 'compositions' R package was programmed during a stage of Tolosana with van den Boogaart funded by the 2005 IAMG Student Grant Program.

REFERENCES:

Aitchison, J., 1986. *The Statistical Analysis of Compositional Data.* New York: Chapman and Hall. 416p.

Aitchison, J., 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In *Proceedings of IAMG'97* (ed. by V. Pawlowksy-Glahn), 1: 3-35.

Aitchison, J., Barceló-Vidal, C., Egozcue, J.J. and Pawlowsky-Glahn, V., 2002. A coincise guide for the algebraic-geometric structure of the simplex, the sample space for the compositional data analysis. In *Proceedings of IAMG'02* (ed. by U. Bayer, H. Burger, W. Skala) Terra Nostra, 3: 387-392.

Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. 2001. Mathematical foundations for compositional data analysis. In *Proceedings of IAMG'01* (ed. by G. Ross) CD-Rom. 20 p.

Bren, M. and Batagelj, V., 2005. Compositional data analysis with R. In: *Compositional Data Analysis Workshop - CoDaWork'05.* ISBN 84-8458-222-1.
CD-Rom `http://ima.udg.es/Activitats/CoDaWork05/` (ed. by g. Mateu-Figueras and C. Barceló-Vidal). Software available at
`http://vlado.fmf.uni-lj.si/pub/MixeR`.

Egozcue, J.J., and Pawlowsky-Glahn, V., 2005. Groups of parts and their balances in compositional data analysis. *Mathematical Geology* 37(7): 795-829.

Van den Boogaart, K. G. and Tolosana R., 2005: The 'compositions' R Package. Available at `http://cran.r-project.org/src/contrib/Descriptions/compositions.html`