Alpine and Mediterranean Quaternary, 31 (1), 2018, 59 - 73 https://doi.org/10.26382/AMQ.2018.05

ISSN 2279-7335

Supplementary Appendices

Appendix 1: *ilr-coordinates and Bayesian-multiplicative treatment of zeros*

In our work, each vector represents the composition of species in both fossil and modern data assemblages. Such compositions can be formalised as a vector $\mathbf{x} = [x_1,\ldots,x_D]$ of non-negative elements representing proportions with sum-constraint $x_1+\ldots+ x_D = 1$. Compositions belong to the simplex S^D, the sample space: $S^D=\{\mathbf{x} \in R^D / x_i>0, i=1..D; x_1+...+x_D= 1 \}$. Log-ratio coordinates are obtained by one-to-one correspondences between vectors of percentages from CoDa set $\mathbf{X}\in S^D$ and the log-ratio vectors from the coordinates dataset $\mathbf{Y} \in R^{D-1}$ in real space. This correspondence allows the use of standard multivariate techniques on the coordinates dataset **Y**. Typical log-ratios coordinates are the centered log-ratio coefficients (clr_j):

$$\mathrm{clr}(\mathbf{x})= \left(\mathrm{clr}_1(\mathbf{x}), \ldots,\mathrm{clr}_D(\mathbf{x})\right) \text{ with } \mathrm{clr}_j(\mathbf{x})= \ln\frac{x_i}{g(\mathbf{x})},$$

where $g(\mathbf{x})$ is the geometric mean of the composition. Because $\sum_{j=1}^{D} \mathrm{clr}_j(\mathbf{x})=0$, then (i.e. the centred log-ratio covariance matrix is singular) the dimension of the clr-space is *D*-1. One can construct an orthonormal basis in the clr-space to obtain orthonormal log-ratio coordinates. To do this one can calculate the isometric log-ratio coordinates (ilr) of the percentages of species. The ilr-coordinates' vector is defined by

$$\mathrm{ilr}(\mathbf{x}) = \mathbf{y} = [y_1,\ldots, y_{D-1}]\in R^{D-1}, \text{ where } \quad \mathbf{y}_i = \frac{1}{\sqrt{i(i+1)}}\ln\left(\frac{\prod_{j=1}^{i}\mathbf{x}_j}{(\mathbf{x}_{i+1})^i}\right)$$

From among the log-ratio coefficients, the orthornormal coordinates are preferred because of their advantageous theoretical and practical properties (Pawlowsky-Glahn et al., 2015). The log-ratio coordinates obviously cannot be obtained for CoDa with zero values. Consequently, before proceeding with the log-ratio coordinates, observations with zero values had to be pre-processed with appropriate compositional techniques. The zeros can be present for various reasons in a CoDa set. Martín-Fernández et al. (2011) presented a comprehensive description of this difficulty, known in the literature as the *zero problem*, and the several types of zeros and its recommended treatments. In our work we dealt with so-called *count* zeros, i.e. compositional count datasets that contain zero values resulting from insufficiently large samples. We modelled our observations with zero values by a replacement strategy following a *Bayesian-multiplicative* approach (Martín-Fernández *et al.*, 2015). This approach preserves the true total number of counts per row and the ratios between the observed species. In this way, the distortion of the covariance structure for the observed part of the dataset is minimized. This property, firstly introduced in Martín-Fernández *et al.* (2003), is based on a "readjustment" of the non-zero values in a multiplicative way.

In the Bayesian-multiplicative method we assume that the counts vector **c** of assemblages derives from a multinomial distribution with D categories, i.e., number of different species. Let N be the total count in **c** and let **θ** be the parameter vector of probabilities, where we assume that $\theta_k>0$. This assumption is crucial because it indicates that the zero values observed in the vector $\mathbf{x}= \mathbf{c}/N$ are due the sample size. Note that vector **x** is an estimate of vector **θ**. Using a Bayesian approach, the prior distribution for **θ** is the conjugate distribution of the multinomial: a Dirichlet

Alpine and Mediterranean Quaternary, 31 (1), 2018, 59 - 73 https://doi.org/10.26382/AMQ.2018.05

ISSN 2279-7335

Supplementary Appendices

distribution of parameter vector $\boldsymbol{\alpha}$, where $\alpha_k = st_k$, k=1,..., D. The parameter s is the scalar "strength" of the prior; and vector **t** is the "prior" expectation for $\boldsymbol{\theta}$. After one sample is collected, the posterior distribution for $\boldsymbol{\theta}$ is a Dirichlet distribution of parameter vector $\boldsymbol{\alpha}^*$, where $\alpha_k^* = c_k + st_k$. Here c_k is the observed counts in the category k of vector **c**. Therefore, Bayes theorem gives the posterior estimate for θ_k

$$\hat{\theta}_k = \frac{c_k + st_k}{N + s}.$$

For each percentage vector **x** in **X** the replacement of the zeros transforms in a vector $\mathbf{x}^* = (x_1^*, \ldots, x_D^*)$ where

$$x_k^* = \begin{cases} \frac{st_k}{N+s} & if\ x_k = 0 \\ x_k\left(1 - \sum_{x_r=0} x_r^*\right) & if\ x_k > 0 \end{cases}.$$

When one assumes a prior non-informative model, the value of tk is equal to 1/D. s is a parameter that controls for the effect (or weight) that the prior distribution of probability has on the posterior distribution of probability. Note that if s = 0, then the posterior is equal to ck/N and only depends on the observed data in the trial and the prior t has no effect on the posterior distribution. Following Palarea-Alabaldejo and Martín-Fernández (2015), in our work we selected the Jeffreys prior (JBZR), where s = D/2 and tk = 1/D. Other different priors were checked with no significant differences (Martín-Fernández et al., 2015) because the values of Pearson correlation coefficient between the results provided by different priors were all greater than 0.99.

After the replacement a new dataset **X*** without zeros is available and one could also make the dataset **C*** (pseudo-counts) without zeros if we multiply each row of **X*** by its total count. The ilr-coordinates can be obtained from the new dataset, **X*** or **C***, obtaining the same coordinates dataset **Y**. Because these ilr-coordinates vectors y are equivalent to the coordinates of the composition x* from a particular orthonormal basis then any typical multivariate technique can be consistently applied. The unique requirement is that this technique should be invariant under change of orthonormal basis. The CoDa-MAT, based on Aitchison distances, verifies this requirement.

Alpine and Mediterranean Quaternary, 31 (1), 2018, 59 - 73 https://doi.org/10.26382/AMQ.2018.05

ISSN 2279-7335

Supplementary Appendices

Appendix 2: *Planktonic foraminifera taxonomical groups*

19 taxonomical groups dataset

Globigerina bulloides d'Orbigny 1826
Globigerina falconensis Blow, 1959
Globigerinella siphoniphera (d'Orbigny 1839)
Globigerinita glutinata (Egger 1893)
Globigerinoides ruber (d'Orbigny 1839) var. *alba*
Globigerinoides ruber (d'Orbigny 1839) var. *rosea*
Globigerinoides sacculifer (Brady 1877)
Globorotalia hirsuta (d'Orbigny 1839)
Globorotalia inflata (d'Orbigny 1839)
Globorotalia menardi-tumida group
Globorotalia scitula (Brady 1882)
Globorotalia truncatulinoides (d'Orbigny 1839)
Globoturborotalita spp.
Neogloboquadrina dutertrei (d'Orbigny 1839)
Neogloboquadrina pachyderma (Ehremberg 1861) left-coiled
Neogloboquadrina pachyderma (Ehremberg 1861) right-coiled
Orbulina universa d'Orbigny 1839)
Pulleniatina obliquiloculata (Parker & Jones, 1862)
Turborotalita quinqueloba (Natland 1938)

15 taxonomical groups dataset

Globigerina bulloides d'Orbigny 1826
Globigerina falconensis Blow, 1959
Globigerinita glutinata (Egger 1893)
Globigerinoides ruber (d'Orbigny 1839)
Globigerinoides sacculifer (Brady 1877)
Globorotalia hirsuta (d'Orbigny 1839)
Globorotalia inflata (d'Orbigny 1839)
Globorotalia menardi-tumida group
Globorotalia scitula (Brady 1882)
Globorotalia truncatulinoides (d'Orbigny 1839)
Neogloboquadrina dutertrei (d'Orbigny 1839)
Neogloboquadrina pachyderma (Ehremberg 1861) left -coiled
Neogloboquadrina pachyderma (Ehremberg 1861) right-coiled
Turborotalita quinqueloba (Natland 1938)

AWS: includes *Globigerinella siphoniphera* (d'Orbigny 1839); *Globoturborotalita rubescens* (Hofker 1956); *Globoturborotalita tenella* (Parker 1958); *Orbulina universa* d'Orbigny 1839).

Supplementary references:

Martín-Fernández J.A., Palarea-Albaladejo J. and Olea R.A. (2011) - Dealing with zeros (Chapter 4). In: Pawlowsky-Glahn and Buccianti, Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Ltd., Chichester, UK, 47-62.

Palarea-Albaladejo J., Martín-Fernández J.A. (2015) - zCompositions - R package for multivariate impu-tation of nondetects and zeros in compositional datasets. Chemometrics and Intelligent Laboratory Systems, 143, 85-96.