Available online http://amq.aiqua.it
ISSN (print): 2279-7327, ISSN (online): 2279-7335

Alpine and Mediterranean Quaternary, 32 (1), 2019, 31-41 https://doi.org/10.26382/AMQ.2020.02

COMPOSITIONAL REGRESSION-BASED METHODS FOR SST RECONSTRUCTION.

Valentino Di Donato [1], Joanna Jamka [1], Josep Antoni Martín-Fernández [2]

[1] Dipartimento di Scienze della Terra, dell'Ambiente e delle Risorse, Università degli Studi di Napoli "Federico II", Naples, Italy.
[2] Departament d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, Spain.

Corresponding author: V. Di Donato <valedido@unina.it>

ABSTRACT: The information in modern or fossil foraminifera assemblages is the relative abundance or percentages of species, i.e., they can be considered as compositional data. In this study we deal with CoDa and regression-based methods as tools to estimate past climatic conditions. We tested standard and robust Partial Least Squares and Principal Component Regression, applied to the log-ratio coordinates of percentage data of Atlantic Ocean and Mediterranean Sea planktonic foraminiferal assemblages. Due to the presence of groups, it was preferred to model separately high latitude and mid to low latitude assemblages. This approach implies the application of cluster analysis, MANOVA and discriminant analysis to the logratio transformed fossil assemblage's compositions. The methods were then applied on marine core assemblages to reconstruct past sea surface temperatures. The obtained results were compared with those formerly obtained by means of compositional modern analogue technique and with the information arising from other paleoclimatic proxies.

Keywords: partial least squares, principal components regression, compositional data analysis, sea surface temperatures, isometric logratio transformation.

1. INTRODUCTION

In the last decades, several methods were proposed to obtain quantitative estimates of past environmental parameters from counts of fossils assemblages (Imbrie & Kipp, 1971, Hutson, 1979; ter Braak & Juggins, 1993; Pflaumann et al., 1996; Waelbroeck et al., 1998; Malmgren et al., 2001; among others). Most methods are applied to percentage data obtained from counting of specimens. *The peculiar* properties of relative abundance data represent however a key issue to be taken into account when developing transfer functions based of fossil assemblages. Percentage data belong to compositional data (CoDa) (Aitchison, 1986): that is, the information contained in a vector of counts $\mathbf{x}$ is the same as in $k \cdot \mathbf{x}$, for any real scalar $k>0$, property known as scale invariance which indicates that a composition is an equivalence class (Barceló-Vidal & Martín-Fernández, 2016). This type of data is very common in Earth Sciences when the constituents and compounds are described in terms their concentration (e.g., Buccianti et al., 2006). In a paper recently published (Di Donato et al., 2018) we revised the modern analogues technique (MAT) (Hutson, 1979; Pflaumann et al., 1996; Waelbroeck et al., 1998) according to the CoDa methodology (Aitchison, 1986). In this study, following the same approach, we deal with regression-based methods, such as Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR). In order to apply PLSR to CoDa, Hinkle & Rayens (1995) proposed logcontrast partial least squares (LCPLS). CoDa refers to vectors of positive components showing the relative weight of a set of parts in a total. Nowadays, there is a general agreement that applying the standard statistical methods to CoDa may yield misleading results (Pawlowsky-Glahn et al., 2015). The log-ratio methodology proposed by Aitchison (1986) and the following developments (i.e. Martín-Fernández & Thió-Henestrosa, 2016a; Martín-Fernández & Thió-Henestrosa, 2016b) represent a powerful set of methods and techniques to apply to CoDa. The approach adopted in this paper follows the principle of working on coordinates (Mateu-Figueras et al., 2011), that is, the standard statistical analysis is conveniently performed after choosing log-ratio coordinates. In particular, we considered to express each D-vector $\mathbf{x} = (x_1, ..., x_D)$ of percentages of species as: 1) a D-dimensional vector $\mathbf{z}=(z_1, ..., z_D)$ of centred log-ratio coordinates ($\mathbf{z} = \mathrm{clr}(\mathbf{x})$) (Aitchison, 1986) and 2) a (D−1)-dimensional real vector $\mathbf{y}=(y_1, ..., y_{D-1})$ of isometric log-ratio coordinates ($\mathbf{y} = \mathrm{ilr}(\mathbf{x})$) (Egozcue et al., 2003) (see Appendix 1 in Di Donato et al. (2018) for definitions and details). To develop our approach, the following points were considered: 1) pre-processing techniques; 2) elaboration of regression-based transfer functions 3) evaluation of the results 4) application to fossil assemblages. As case studies, we considered applications based on the estimation of past sea surface temperatures (SST) from planktonic foraminifera assemblages. However, the described methods may be applied to different palaeoecological contexts. The analyses were carried out with MATLAB codes except for raw-data analysis which was computed with R package rioja

			Matlab codes
Part 1: modelling of modern assemblages			
step 1	Variable selection and (eventually) amalgamation		*raggruppa*
step 2	Zero replacement (on both modern and fossil assemblages)		*zerorep* (for proportion data); *zeroconteggi* (for count data)
step 3	clr- or ilr- transformation (of both modern and fossil assemblages)		*clr* *ilrprogr* *balances*
	Single group analysis	Multigroup analysis	
step 4	Go to step 6	Classification of modern assemblages, evaluation of groups	*pdist* – linkage* -manova1**
step 5		Extraction of subcompositions (if needed)	*raggruppa, estrai*
step 6	Partial Least Squares Regression or Principal Components Regression modelling of modern assemblages (single or multigroup)		*codatransfer*
Part 2: application to fossil assemblages			
step 7	Detection of no-analog assemblages (atypicality index and local outlier factor, as done in CodaMat)		*mat, LOF***
step 8	Go to step 9	Discriminant analysis of fossil assemblages with modern as training groups (on ilr coordinates)	*classify** (called by *codatransfer*)
step 9	Application of the model to fossil assemblages		*codatransfer*

*MATLAB toolbox functions **available at http://dsmi-lab-ntust.github.io/AnomalyDetectionToolbox/

Tab. 1 - Workflow of the analysis

(Juggins, 2017). MATLAB routines expressly written to perform CoDa-regression and CoDaMAT are provided in the supplementary materials. The Robust Partial Least Squares Regression requires the Rsimpls.m matlab code included in the LIBRA package (Verboven & Hubert, 2004), available at https://github.com/duncombe/matlab/blob/master/LIBRA/rsimpls.m.
A workflow of the analysis is shown in Table 1.

2. APPROACHING THE ANALYSIS

2.1. The dataset

The dataset on which our applications are computed is represented by a database consisting of 1252 Atlantic and Mediterranean planktonic foraminifera coretop assemblages, which represent the regressor variables, determined on the >150 µm size fraction (Prell et al., 1999; Hayes et al., 2004; Kucera et al., 2004) (Fig. 1). In this paper we considered SST as response variables. The oceanographical data, consisting of mean annual and seasonal SST refer to Antonov et al. (2010) and Locarnini et al. (2010). Seasonal temperatures are 3 months averages, i.e. January–March for northern (southern) winter (summer) and July–September for northern (southern) summer (winter). The SST values at coretop locations were computed by means of Ocean Data View 4.7.10 (Schlitzer, 2018). Following Kucera et al. (2005), oceanographical data are related to a depth

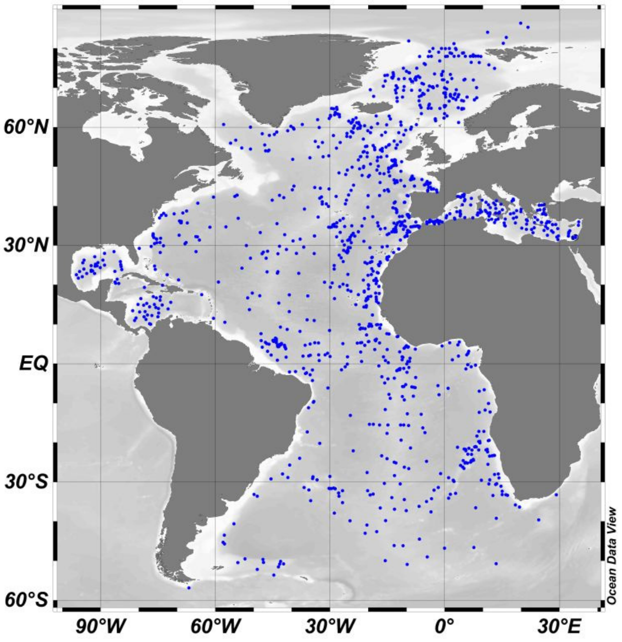


Fig. 1 - Location of modern planktonic foraminifera coretop samples adopted for application of CoDa-MAT. Drawn with Ocean Data View software (Schlitzer, 2018).

of 10 m.

2.2 Data pre-processing: subcomposition, amalgamation and zero replacement

The data pre-processing which is needed to accomplish the analysis is fully explained in Di Donato et al. (2018) and will not be detailed here. In short, our approach requires data to be strictly positive (Aitchison, 1986). To reduce the number of zero values to be replaced, rarer species, which carry inevitably low signal to noise ratios (Kucera et al., 2005) can be excluded from the assemblages, by considering a subcomposition (Aitchison, 1986) of the original assemblages. Moreover, an amalgamation of taxa characterised by similar ecological requirements can be also considered (Aitchison, 1986). To manage the zeros occurring in the data, we adopted a mixed Bayesian-multiplicative estimation approach, which is recommended when the compositional data arise from counts (Martín-Fernández et al., 2003; Martín-Fernández et al., 2015) (see also Appendix 1 in Di Donato et al., 2018).

After the zero replacement, the clr- and ilr-coordinates' vectors for the fossil and modern data are obtained. Thus, fossil assemblages are represented by its logratio-coordinates, whereas the modern database is consisting of logratio-coordinates together with the environmental parameters measured at each location. The ilr coordinates can be computed by means of balances (Egozcue et al., 2005). These are logcontrasts obtained by means of a Sequential Binary Partition (SBP) matrix. For each order of the partition, it is possible to define the balance between the two sub-groups formed at that level: if $i_1, i_{2,\dots,} i_r$ are the r parts of the subgroup coded by +1, and $j_1, j_{2,\dots,} j_s$ the s parts of the subgroup (coded by -1), a balance is defined as:

$$y = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_{i_1} x_{i_2} \dots x_{i_r})^{1/r}}{(x_{j_1} x_{j_2} \dots x_{j_s})^{1/s}} \qquad (1)$$

From a *D* part composition, *D-1* balances (ilr-coordinates) can be obtained.

In our case, for the sake of simplicity, we adopted a matrix of the type:

order	x_1	x_2	x_3		x_{D-1}	x_D
1	1	-1	-1	….	-1	-1
2	0	1	-1	….	-1	-1
3	0	0	1	….	-1	-1
….	….	….	….	….	….	….
D-1	0	0	0	….	1	-1

As explained in Di Donato et al. (2018), two datasets were generated, consisting of respectively 19 and 15 taxonomical groups (see appendix 2 in Di Donato et al., 2018).

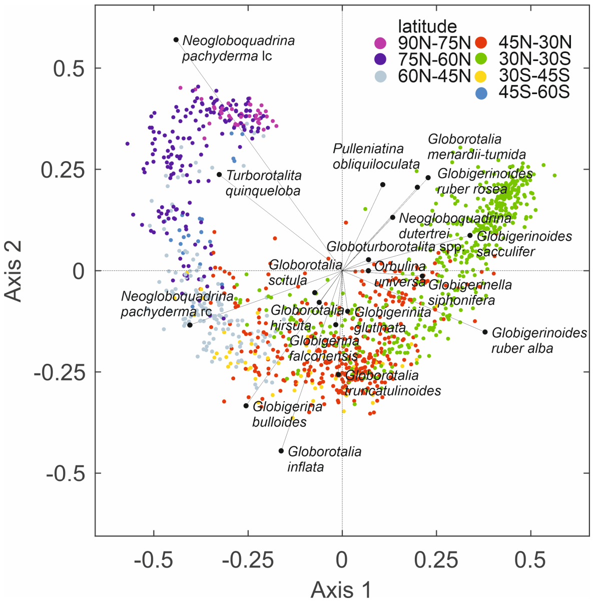


Fig. 2 - Relative variation biplot (RVB) of planktonic foraminiferal assemblages included in the modern dataset. Row points are grouped according to their latitude.

2.3. Multiple regression methods

The PLSR and PCR methods applied to perform the palaeo-estimates are models recommended to predict a response variable when there are a large number of predictor variables highly correlated (ter Braak & Juggins; 1993 Juggins, 2017). Both methods construct new orthogonal, hence not correlated, predictor variables as linear combinations of the original predictor variables. However, while PCR creates components to maximise the observed variability in the predictor variables, without considering the response variable at all, PLSR creates components by maximising the covariance between predictors and response variables. For PLSR we adopted the SIMPLS algorithm (de Jong, 1993). We also tested a robust PCR (RPCR) and robust PLS (RSIMPLS) (Hubert & Branden, 2003). In order to comply with the CoDa approach, both analyses should be computed on ilr coordinates. It can be noted, however, that clr- and ilr-coordinates, provide the same results for both PCR and PLSR (Filzmoser et al., 2018). Hereafter we denoted with CoDa-PLSR and CoDa-PCR the analysis performed on log-ratio coordinates. The main advantage of clr-coordinates is that they are logcontrasts more easily interpretable. However, ilr-coordinates are advantageous when the analysis requires full rank data, such as discriminant analysis.

In order to evaluate the number of PLS components to be taken into account, we considered, as usual, the percent of variance explained in the response variable and the mean-squared errors (see section 2.4) as a function of the number of components. For CoDa-PCR we considered the percent of total variance of regressor variables.

The Relative variation biplot (RVB) of the planktonic foraminifera assemblages included in the modern dataset is shown in Figure 2. A RVB consists of a standard principal components biplot applied to the clr coordinates of the assemblages. The first two axes accounts for about 63% of total variability (a total of 71% is reached if a third axis is added). The location of the data points in the RVB highlights the well-known broad latitudinal distribution of assemblages. It can be noted that the origin of the RVB corresponds to a low-density area. Moreover, the distribution of data points in the RVB seems to indicate the existence of two groups within the dataset, the first of which is represented by high latitude assemblages. Principal Components Analysis (PCA) is properly defined for homogeneous, normally-distributed data from a single population (Tolosana-Delgado & Mc Kinley, 2016). Thus, we considered evaluating, apart from a regression model built with the whole dataset, two separate models for high latitude and low to middle latitude assemblages. Separate regional models were considered, among others, in Prell (1995), Ortiz & Miz (1997) and Kucera et al. (2005). For our purposes, a cluster analysis (Ward's method on ilr-coordinates) was applied to the dataset, discarding then some observations that did not belong to either of the two main defined groups. A MANOVA test indicates that the two groups are significantly distinct (with p~0). Since at high latitude only few planktonic foraminifera species occur, the regression models for the high latitude group was built by considering a sub-composition of the dataset made of only 5 species. The application of this two-step procedure to fossil assemblages requires to first perform a linear discriminant analysis (LDA) in order to classify them into one of the 2 defined groups. Apart from theoretical considerations, the application of LDA to percentage data faces problems related to the singularity of the within groups variance/covariance matrix. This problem also occurs with clr-coordinates, while ilr-coordinates are not affected by this problem. The LDA misclassification rate provided a leave-one-out cross validation method for the two groups defined in the modern dataset is 0.0085.

2.4. Evaluation of the quality of the estimates

The evaluation of the quality of the estimated results provided by CoDa-PCR and CoDa-PLSR techniques has been carried out by means of sensitivity analysis of leave-one-out cross validation method (ter Braak & Juggins, 1993; Barrows & Juggins, 2005). The indices we took into account were: the Coefficient of determination R^2 and the Mean Square Error of Prediction (MSEP) (with its square root, the RMSEP) (Wallach & Goffmet, 1989; Birks, 1995) and as multivariate counterpart, the root of mean squared distances (RMSD) (e.g., Martín-Fernández et al., 2003).

The MSD is a multivariate index of quality of the estimates of *k* environmental parameters defined as:

$$MSD = \sum_{i=1}^{n} d_e^2 \left(\mathbf{P}(i), \hat{\mathbf{P}}(i) \right)^2 \Big/ n \qquad (2)$$

where *n* is the number of modern samples, $\mathbf{P}$ represents the vector of *k* measured environmenal parameters, and $\hat{\mathbf{P}}$ the corresponding vector of estimated values. This approach is equivalent to analyse the mean of the norm of the residual rows (difference between measured and estimated parameters). The best result of RMSD is obtained when RMSD=0, i.e, all the estimates are equal to the measured values. The multivariate approach may be applied when different paleoclimatic parameters are estimated (i.e. temperature, seasonal or annual precipitation, potential evapo-transpiration, as done in pollen analysis). For the evaluation of the results obtained from grouped assemblages, we considered both a "pooled" R^2, i.e. for 2 groups, R^2=(SSR gr1+SSR gr2)/(SST gr1+SST gr2), together with the squared correlation coefficient r^2 obtained by comparing the overall (both groups) measured and fitted SST values. Squared correlation coefficients are reported in the supplementary materials.

2.5. Application on fossil assemblages

Probably the most important difficulty of any proxy-based reconstruction is represented by the no-analogue problem, occurring when the palaeoenvironmental conditions represented in the fossil assemblages do not have a correspondence in modern environments (Hutson, 1977). It can be considered that the application of regression methods to no-analogue samples, may be regarded as extrapolation rather than interpolation (Conn et al., 2015). In Di Donato et al. (2018), we adopted atypicality index (e.g., Aitchison, 1986), both standard and robust, and Local Outlier Factor (LOF) (Breunig et al., 2000) of assemblages as tools to detect no-analogue conditions. Details on the computation of atypicality index and robust methods to outlier detection (Peña & Prieto, 2001) for CoDa sets (Filzmoser & Hron, 2008) are reported in Di Donato et al., 2018.

3. TESTING THE METHOD

3.1. Results

The number of components from which compute the PLS was determined by considering the percent variance explained for the response variables and the MSE of response variable as functions of the number of PLS components (Fig. 3) and, for CoDa-PCR, the amount of total variance of regressor variables accounted by the principal components (Fig. 4). For both the tested datasets, 4 components seem to provide an adequate CoDa-PLSR model. For CoDa-PCR, we adopted a 6 components model. As regards the two-groups approach, the subcomposition obtained for high latitude assemblages is made of only 6 parts. In the other group, for both CoDa-PLSR and CoDa-PCR, we considered a full-components model.

We tested both annual SST and seasonal SST. The relationships between measured and estimated seasonal SST obtained with PCR and PLSR under different conditions are reported in the supplementary materials. In general, the 15 taxa and 19 taxa datasets provided quite similar results. It can be noted, that, for both CoDa-PCR and CoDa-PLSR, higher squared correlation coefficients and lower RMSEP are obtained with a two-groups analysis. The highest r^2 of 0.9718, with a RMSE of 1.33°C was obtained for mean annual SST with CoDa-PLSR. Figure 5 shows the relationship between measured and cross-validation-estimated SST for

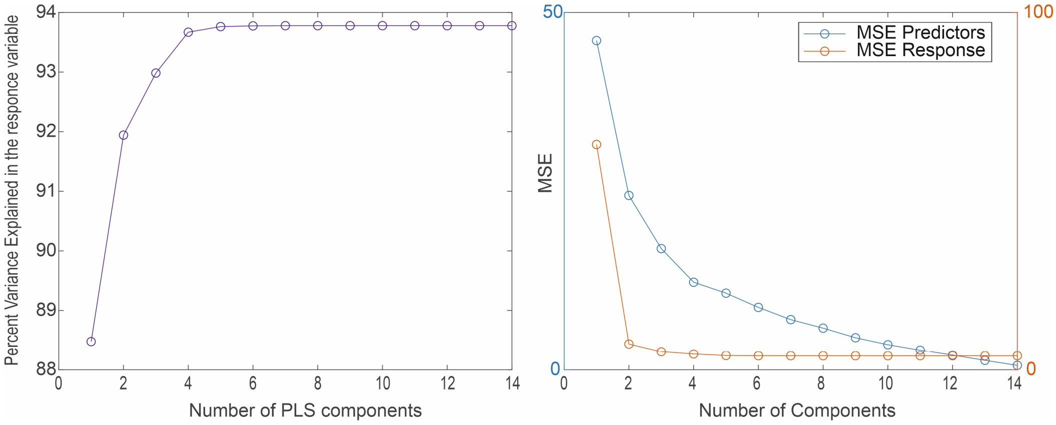


Fig. 3 - The figure shows for the 15 taxonomic groups dataset plots (from which 14 predictors represented by ilr-variables are obtained) of a) the percent variance explained for a single response variable (annual SST) and b) the MSE of response variable as functions of the number of PLS components.

a 1 group and for 2-groups analysis. It can be noted that, the 1-group model, a “plateau” for low SST, which can be also observed, in the Imbrie & Kipp (1971) method and CoDa-MAT validation, and which becomes much less pronounced, with a 2-groups modelling.

For a single group analysis, PLSR performs better than PCR. However, for a 2 groups analysis, PLSR and PCR provided quite similar results. Robust version of CoDa-PLSR and CoDa-PCR did not improve the fitting with respect to standard CoDa-PLSR and CoDa-PCR. In comparison with the CoDa-MAT method, CoDa-PLSR and CoDa-PCR provides slightly lower R^2 and higher RMSEP. A slightly better performance of MAT with respect to a regression method (i.e. the Imbrie & Kipp transfer function) was also found by Ortiz & Mix (1997) by working on raw percentage data. As far as the comparison with raw data regression-based methods, on our dataset, the Imbrie & Kipp (1971) Q-mode regression method yields, with 5 components an R^2=0.8956 and an RMSE=2.47°C. The Weighted Averaged Partial Least

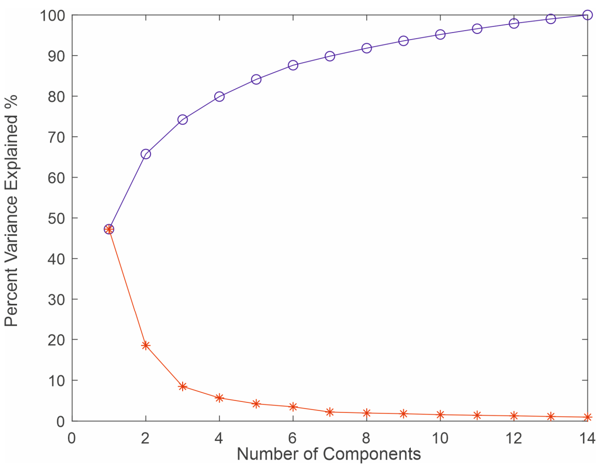


Fig. 4 - Plot of cumulative percentage accounted (blue line) and variance contribution of each component (red line) in CoDa-PCR.

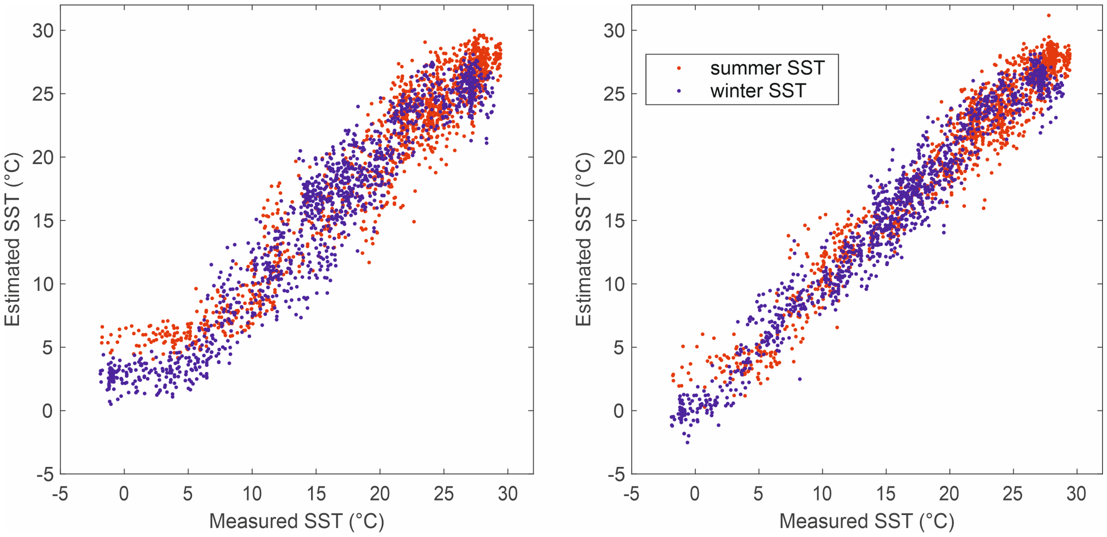


Fig.5 - Plots of observed versus estimated SST (annual and seasonal) obtained with a 1-group (left) and a 2-groups CoDa-PLSR modelling.

Squares Method (WAPLS) (ter Braak & Juggins, 1993) yields an R^2=0.9544 and an RMSE=1.70°C.

3.2. Application examples

As an application example, the CoDa-MAT method was applied to planktonic foraminifera records which were also considered to evaluate the performing of CoDa-MAT. The first one is a literature dataset, consisting of the record of planktonic foraminifera assemblages of the core MD95-2040 (de Abreu et al., 2003; Voelker & de Abreu, 2011), recovered in the Atlantic Ocean off the Iberian margin, the second is that of GNS84-C106 core recovered in the Tyrrhenian sea (Buccheri et al, 2002; Di Donato et al., 2008; 2009). For the Mediterranean Sea, we also considered the planktonic foraminiferal record of the Core TEA-C6 (Di Donato et al., 2019), from which an estimate of past SST for the last 15 ka, was obtained with the CoDa-MAT method. All datasets are obtained from >150 μm size fractions. The location of the cores is shown in Figure 6.

A discussion of the possible drawback represented by the excessive loss of small sized species in the >150 μm size fraction can be found in Di Donato et al. (2015). It can be noted that >150 micron and >106 μm datasets, if analysed with CoDa methods, provide the same covariance structure. This suggests that regression-based methods based on CoDa, may be quite robust with respect to treatment changes such as the analysed size fraction.

3.2.1. Atlantic Ocean

The foraminiferal record of MD95-2040 core covers the last 210 ka (de Abreu et al., 2003; Voelker, & de Abreu, 2011). SST for this interval were formerly reconstructed (de Abreu et al., 2003) from planktonic foraminifera with SIMMAX28 method (Pflaumann et al., 1996). The dataset consists of 732 assemblages. As regards the atypicality of assemblages, in relation to the 99.5 percentile only 4% of the samples have Mahalanobis distances are above the xi-square critical value of 31.32 (see Figure 6 in Di Donato et al., 2018). As for the LOF it can be noted that glacial assemblages are characterised by higher values of up to 2, while most interglacial assemblages have LOF values not exceeding 1.5. On the basis of LDA computed on ilr-coordinates, 98 assemblages were classified into the high latitude assemblage group, and 634 assemblages into the low to middle latitude assemblage group. The output of the LDA with the indication of the group to which each assemblage was assigned with the posterior probabilities is provided in the supplementary materials.

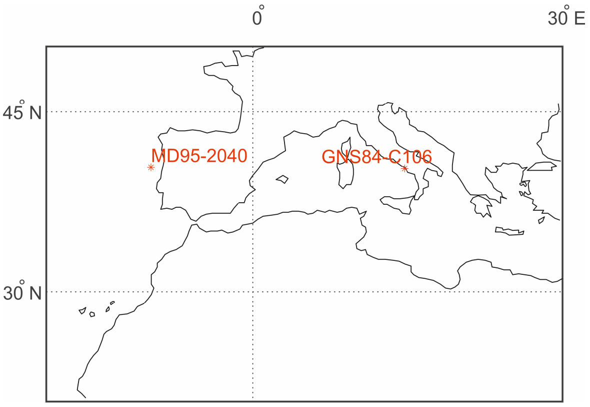


Fig. 6 - Location of cores considered in this paper.

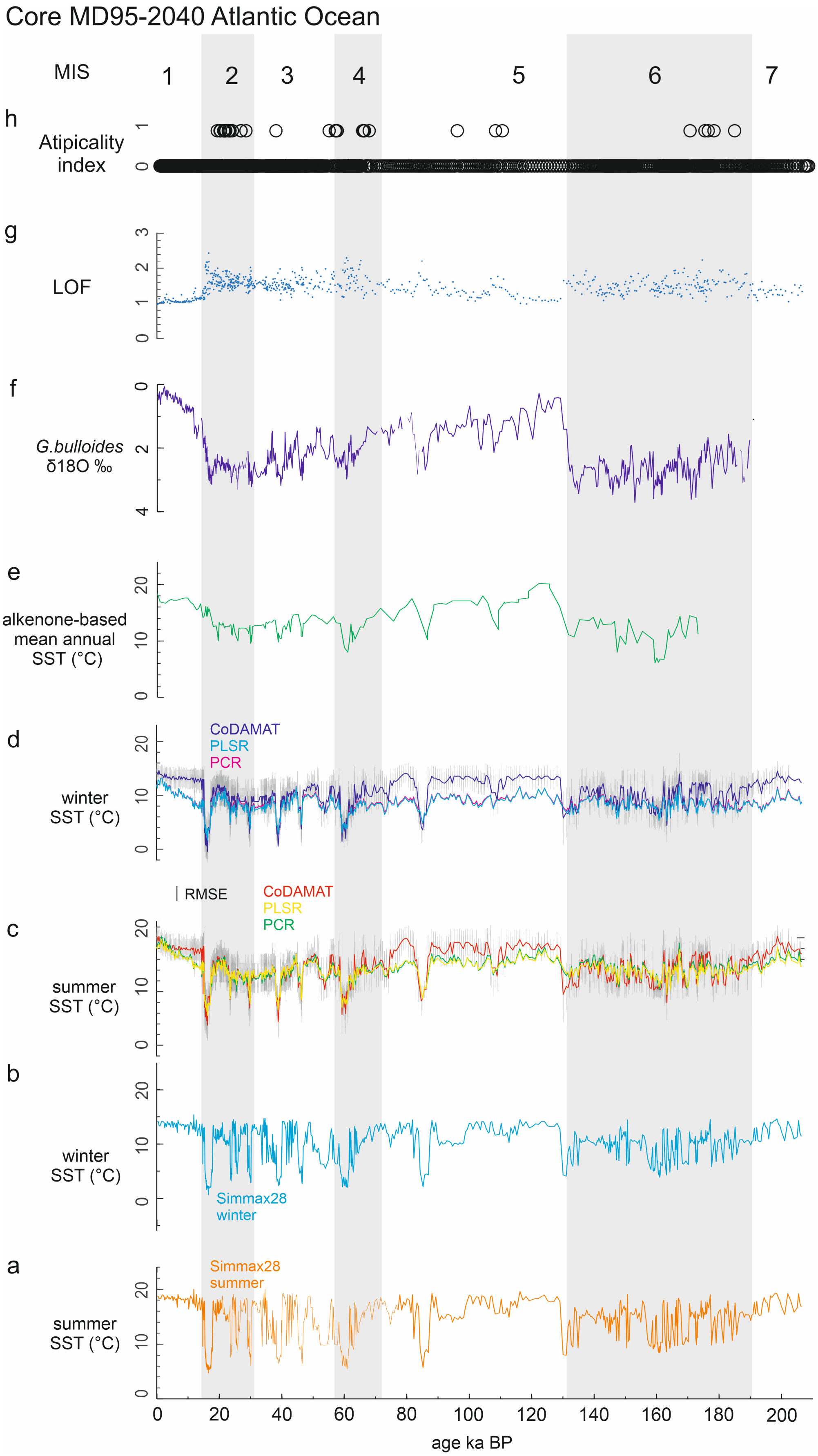
A comparison between the values reconstructed for summer and winter SST by means of SIMMAX28 (de Abreu et al., 2003), CoDa-MAT, CoDa-PLSR and CoDa-PCR is shown in Figure 7. The stronger coherence of CoDa-MAT reconstruction with Alkenones and the stable isotope record with respect to SIMMAX28 has been already highlighted in Di Donato et al. (2018). Here we note that CoDa-PLSR and CoDa-PCR reconstruction are largely overlapping and show a same general trend if compared with CoDa-MAT. Several SST minima, which correspond to Heinrich events, are also recorded by CoDa-PLSR and CoDa-PCR with slightly less-deep minima in comparison with CoDa-MAT but more marked in comparison with alkenones record. CoDa-MAT and Regression-based methods also provide different SST reconstructions for the MIS5: the former provides higher SST estimates, while the latter highlight a decreasing trend during the MIS5 which does not appear in the CoDa-MAT reconstruction. The SST reconstructed with the 15 taxa and the 19 taxa datasets, are quite similar, being characterised by a r=0.9865 (r= 0.9868) and by a root mean squared difference of 0.70°C (0.48°C) for summer (winter) SST. The multivariate RMSD between 15 taxa and the 19 taxa reconstructed SST is 0.7177.

3.2.2. Mediterranean Sea

The Core GNS84-C106 recovered in the Gulf of Salerno (Tyrrhenian sea - Western Mediterranean) covers the last 34 ka (Di Donato et al., 2009). This dataset is represented by 228 planktonic foraminiferal assemblages determined on the >150-micron size fraction. The quantitative reconstructions of past climatic conditions for the Mediterranean basin face several problems related to the peculiar hydrological asset of this semi-enclosed basin. Reconstructions became even more problematic for glacial intervals (Sbaffi et al., 2001, among others). As regards the LDA, the whole Core GNS84-C106 dataset was classified into the low to middle latitude group (see supplementary material). However, as shown in Figure 8, a significant atypicality index

------ --->>>>>

Fig. 7 - Reconstruction of seasonal SST for the last 210 ka off the Iberian margin from core MD95-2040 and comparison between SIMMAX28 (de Abreu et al., 2003, Voelker & de Abreu, 2011), CoDa-MAT, CoDa-PLSR and CoDa-PCR reconstructed SSTs. a) summer and b) winter SIMMAX28 reconstruction. c) summer and d) winter SST CoDa-based reconstructions. Grey lines indicate the standard deviation of CoDa-MAT estimates e) Alkenone based SST reconstruction (Pailler & Bard, 2002). f) Globigerina bulloides stable isotope record and Marine Isotopic Stages (MIS) (Abreu et al., 2003; Schönfeld et al., 2003): grey-shaded dots: distance of fossil assemblages from each of the 6 closest modern analogues. Full line: mean values. g) LOF values (see Di Donato et al., 2018) h) atypicality index: 0: not significant; 1: significant.


Core MD95-2040 Atlantic Ocean
MIS
1
2
3
4
5
6
7
h
Atipicality index
0
1
g
LOF
0
1
2
3
f
G.bulloides δ18O ‰
0
2
4
e
alkenone-based mean annual SST (°C)
0
10
20
d
winter SST (°C)
CoDAMAT
PLSR
PCR
0
10
20
RMSE
CoDAMAT
PLSR
PCR
c
summer SST (°C)
0
10
20
b
winter SST (°C)
Simmax28 winter
0
10
20
a
summer SST (°C)
Simmax28 summer
0
10
20
0
20
40
60
80
100
120
140
160
180
200
age ka BP

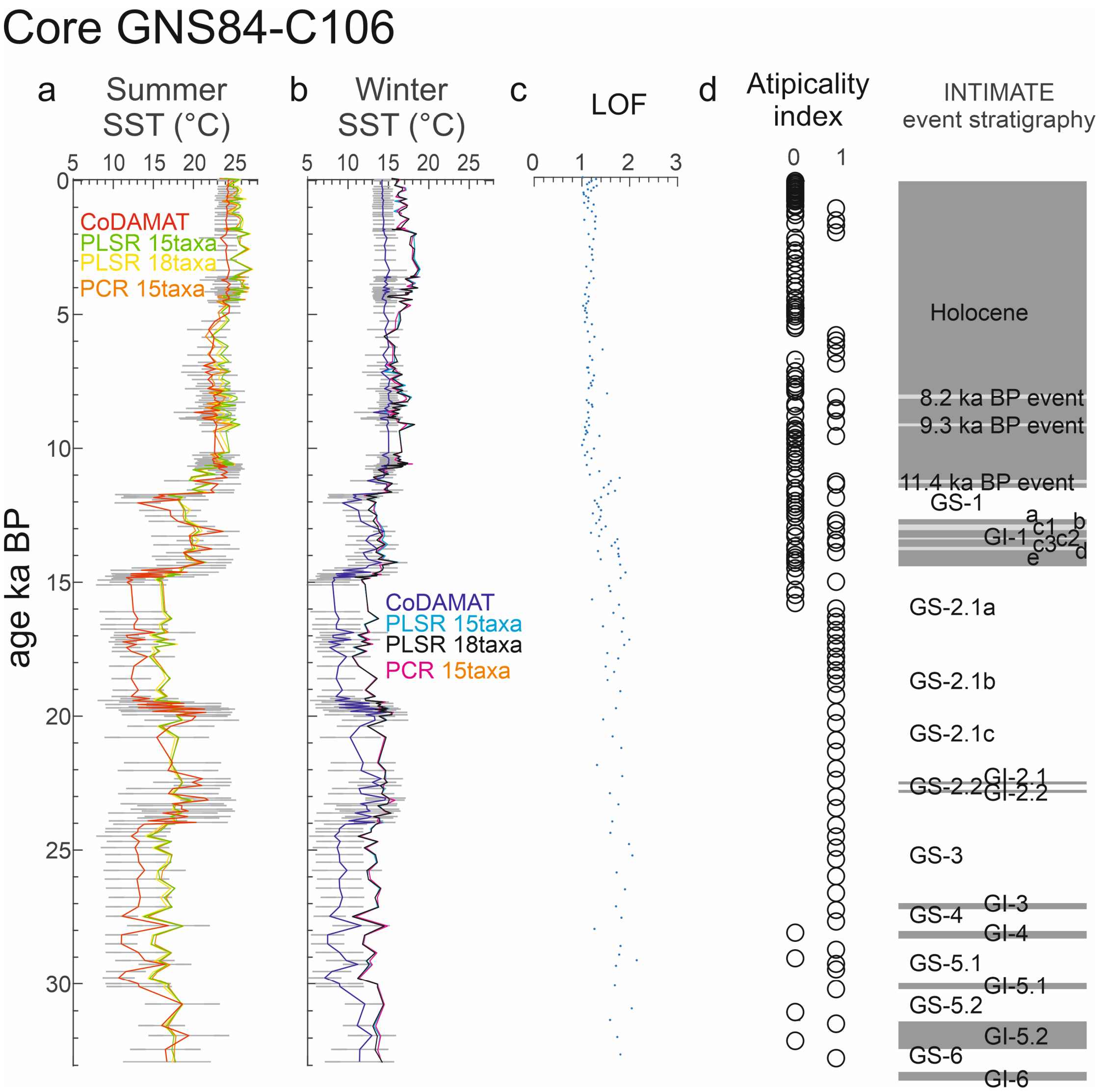


Fig. 8 - Reconstruction of seasonal SST obtained from GNS84-C106 Core. a) summer b) winter. The grey error bars indicate the standard deviation of each reconstructed value. c) LOF values e) atypicality index: 0: not significant; 1: significant. The INTIMATE Greenland event stratigraphy is reported from Rasmussen et al. (2014).

was found for several planktonic foraminiferal assemblages of the Core GNS84-C106. As in the previous case study, CoDa-PLSR and CoDa-PCR provide quite similar results. For this Core, however, within the same general trend, regression-based methods and CoDa-MAT provide quite different results. In this case study, raw percentage data MAT and, to a lesser degree, CoDa-MAT provided summer SST which seem quite high for the GI-1 interval and for time intervals of the Last Glacial Period centred around 24 and 20 ka BP. This problem is likely partly related to the adopted size fraction (Di Donato et al., 2015). Regression based methods seem to provide a more coherent SST trend, i.e. Last Glacial Period lower than Holocene, and intermediate SST values during the Late Glacial. However, it can be noted that CoDa-PLSR and CoDa-PCR provide higher SST estimates for the colder intervals of the glacial period. As an example, between 15 and 17 ka BP, CoDa-MAT reconstruct winter SST of even 8.5 °C, while CoDa -PLSR and CoDa-PCR indicate lower values of about 10°C. For this core, we do not have, at present alternative proxy-based reconstructions. However, the regression based reconstructed SSTs seem more coherent with alkenone-based reconstructions obtained for the Southern Tyrrhenian Sea (Sbaffi et al, 2001), which provided higher SST values if compared with the MAT reconstructions. During the Holocene, CoDa-MAT, CoDa-PLSR and CoDa-PCR indicate an SST rise around 5 ka BP. However, CoDa-PLSR and CoDa-PCR indicate warmer than present SST for an interval centred around 4 to 3 ka BP, which coincides with a peak in the warm species *Globigerinoides sacculifer* widely recognized in the Mediterranean Sea (Capotondi et al., 1999; among others) and a decreasing trend afterwards. This trend is quite similar to the alkenones record of the Core BS79-38 recovered in the Southern Tyrrhenian Sea (Sbaffi et al, 2001). However, alkenones provided higher than present SST values for the early to middle Holocene which are not confirmed by planktonic foraminifera.

4. CONCLUSIVE REMARKS

Following out previous paper focused on CoDa and modern analogue technique, in this article we developed a transfer function based on multivariate regression methods in a fashion coherent with a CoDa approach. The main advantages of CoDa-MAT, being a non-parametric method (Guiot & De Vernal, 2011a;b), is its flexibility and the fact that the quality of each single reconstruction can be evaluated by means local outlier factor and mean distances. By contrast, once the model has been built, we must accept the reconstructed SST "as they are". Likely, CoDa-PLSR might be more sensitive than CoDa-MAT to random effects, since limited random effects should not strongly influence the choice of the best modern analogs for a fossil assemblage. However, CoDa-PLSR and CoDa-PCR, may be more robust with respect to treatment changes such as the size fraction adopted for the analysis, which represent a critical point in the analysis of foraminifera assemblages. Whatever the approach, it is important to evaluate the atypicality of the fossil assemblages in comparison with modern ones. In this article we provided application examples for an Atlantic Ocean and a Mediterranean Sea core. CoDa-MAT and regression-based methods seem to provide quite coherent reconstructions for the Atlantic Ocean, while for the Mediterranean Sea, the obtained reconstructions are, as expected, more problematic. Together with CoDa-MAT, CoDa-PLSR and CoDa-PCR provide the basis for more extensive reconstructions which will be the focus of future investigations.

ACKNOWLEDGEMENTS

We would like to thank the two reviewers for their careful reading of the paper and their helpful comments and prof. Adele Bertini for her support in the revision of the article. This work has been supported by the project "CODAMET" (Ministerio de Ciencia, Innovación y Universidades; Ref: RTI2018-095518-B-C21).

REFERENCES

Aitchison J. (1986) - The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd. (Reprinted 2003 with additional material by The Blackburn Press), London (UK), pp. 416.

Antonov J.I., Seidov D., Boyer T.P., Locarnini R.A., Mishonov A.V., Garcia H.E. (2010) - World Ocean Atlas 2009 Volume 2: Salinity. S. Levitus, Ed., NOAA Atlas NESDIS 69, U.S. Government Printing Office, Washington, D.C., pp. 184.

Barceló-Vidal C., Martín-Fernández J.A. (2016) - The mathematics of compositional analysis. Austrian Journal of Statistics, 45(4), 57-71.

Barrows T.T, Juggins S. (2005) - Sea-surface temperatures around the Australian margin and Indian Ocean during the last glacial maximum. Quaternary Science Reviews, 24, 1017-1047.

Birks H.J.B. (1995) - Quantitative palaeoenvironmental reconstructions. In: Maddy, D., Brew, J.S. (Eds.), Statistical Modelling of Quaternary Science Data. Technical Guide 5. Quaternary Research Association, Cambridge, 116-254.

Breunig M.M., Kriegel H.-P., Ng R.T., Sander J. (2000) - LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data (SIGMOD '00). ACM, New York, NY, USA, 93-104. Doi: 10.1145/342009.335388

Buccheri G., Capretto G., Di Donato V., Esposito P., Ferruzza G., Pescatore T., Russo Ermolli E., Senatore M.R., Sprovieri M., Bertoldo M., Carella D., Madonna, G. (2002) - A high resolution record of the last deglaciation in the southern Tyrrhenian Sea: environmental and climatic evolution. Marine Geology, 186, 447-470.

Buccianti A., Mateu-Figueras G., Pawlowsky-Glahn V. (Eds.) (2006) - Compositional Data Analysis in the Geosciences: From Theory to Practice. In: Special Publications, vol. 264. Geological Society, London. pp. 207.

Capotondi L., Borsetti A.M., Morigi C. (1999). Foraminiferal ecozones, a high resolution proxy for the late Quaternary biochronology in the central Mediterranean Sea. Marine Geology, 153, 253-274.

Conn P.B., Johnson D.S., Boveng P.L. (2015) - On extrapolating past the range of observed data when making statistical predictions in ecology. PLoS ONE 10(10): e0141416.
Doi: 10.1371/journal.pone.0141416

de Abreu L., Shackleton N.J., Schonfeld J., Hall M., Chapman M. (2003) - Millennial-scale oceanic climate variability off the Western Iberian margin during the last two glacial periods. Marine Geology, 196 (1-2), 1-20.

de Jong S. (1993) - SIMPLS: an alternative approach to partial least squares regression. Chemometrics Intell. Lab. Syst., 18, 251-263.

Di Donato V., Esposito P., Russo Ermolli E., Scarano A., Cheddadi, R. (2008) - Coupled atmospheric and marine palaeoclimatic reconstruction for the last 35 ka in the Sele Plain-Gulf of Salerno area (southern Italy). Quaternary International, 190, 146-157.

Di Donato V., Esposito P., Garilli V., Naimo D., Buccheri G., Caffau M., Ciampo G., Greco A., Stanzione D. (2009) - Surface-bottom relationships in the Gulf of Salerno (Tyrrhenian Sea) over the last 34 kyr: Compositional data analysis of palaeontological proxies and geochemical evidence. Geobios, 42, 561-579.

Di Donato V., Martin-Fernandez J.A., Daunis-i-Estadella J., Esposito, P. (2015) - Size fraction effects on planktonic foraminifera assemblages: a compositional contribution to the golden sieve rush. Mathematical Geosciences, 47(4), 455-470.

Di Donato V., Martín-Fernández J.A., Comas-Cufí M., Jamka J. (2018) - Palaeoenvironmental reconstructions through Compositional Data analysis.

Alpine and Mediterranean Quaternary, 31(1), 59-73.

Di Donato V., Insinga D.D., Iorio M., Molisso F., Rumolo P., Cardines C., Passaro S. (2019) - The palaeoclimatic and palaeoceanographic history of the Gulf of Taranto (Mediterranean Sea) in the last 15 ky. Global and Planetary Change, 172, 278-297.

Egozcue J.J., Pawlowsky-Glahn V. (2005) - Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7), 795-828.

Egozcue J.J., Pawlowsky-Glahn V., Mateu-Figueraz G., Barceló-Vidal C. (2003) - Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3), 279-300.

Filzmoser P., Hron K. (2008) - Outlier detection for compositional data using robust methods. Math. Geosciences, 40, 233-248.

Filzmoser P., Hron K. Templ, M. (2018) - Applied Compositional Data Analysis. Springer, pp. 280.
Doi: 10.1007/978-3-319-96422-5

Guiot J., de Vernal A. (2011a) - Is spatial autocorrelation introducing biases in the apparent accuracy of paleoclimatic reconstructions? Quaternary Science Reviews 30, 1965-1972.

Guiot J., de Vernal, A. (2011b) - QSR Correspondence "Is spatial autocorrelation introducing biases in the apparent accuracy of palaeoclimatic reconstructions?" Reply to Telford and Birks. Quaternary Science Reviews, 30, 3214-3216.

Hayes A., Kucera M., Kallel N., Sbaffi L., Rohling E. (2004) - Compilation of planktic foraminifera modern data from the Mediterranean Sea. Pangaea.
Doi: 10.1594/PANGAEA.227305

Hinkle J., Rayens W. (1995) - Partial least squares and compositional data: problems and Alternatives. Chemometrics and Intelligent Laboratory Systems 30(1), 159-172.

Hubert M., Branden K.V. (2003) - Robust methods for partial least squares regression. J. Chemometrics, 17, 537-549.
Doi: 10.1002/cem.822

Hutson W.H. (1977) - Transfer functions under no-analog conditions: Experiments with Indian Ocean planktonic foraminifera, Quaternary Research, 8, 355-367.

Hutson W.H. (1979) - The Agulhas Current during the Late Pleistocene: Analysis of modern faunal analogues. Science, 207(1), 64-66.

Imbrie J., Kipp N.G. (1971) - A new micropaleontological method for paleoclimatology: Application to a Late Pleistocene Caribbean core. The Late Cenozoic Glacial Ages. New Haven, Yale University Press, 71-181.

Juggins S. (2017) - rioja: Analysis of Quaternary Science Data, R package version (0.9-21).
http://cran.r-project.org/package=rioja

Kucera M., Weinelt M., Kiefer T., Pflaumann U., Hayes A., Weinelt M., Chen M.-T., Mix A.C., Barrows T., Cortijo E., Duprat J., Juggins S., Waelbroeck C. (2004) - Compilation of planktic foraminifera census data, modern from the Atlantic Ocean. Pangaea.
Doi: 10.1594/PANGAEA.227322

Kucera M., Weinelt M., Kiefer T., Pflaumann U., Hayes A., Weinelt M., Chen M.-T., Mix A.C., Barrows T.T., Cortijo E., Duprat J., Juggins S., Waelbroeck C. (2005) - Reconstruction of the glacial Atlantic and Pacific sea-surface temperatures from assemblages of planktonic foraminifera: multi-technique approach based on geographically constrained calibration datasets. Quaternary Science Reviews, 24, 951-998.

Locarnini R.A., Mishonov A.V., Antonov J.I., Boyer T.P., Garcia H.E. (2010) - World Ocean Atlas 2009, Volume 1: Temperature. S. Levitus, Ed., NOAA Atlas NESDIS 68, U.S. Government Printing Office, Washington, D.C., pp. 184.

Malmgren B. A., Kucera M., Nyberg J., Waelbroeck C. (2001) - Comparison of statistical and artificial neural network techniques for estimating past sea surface temperatures from planktonic foraminifer census data. Paleoceanography, 16(5), 520-530.
Doi: 10.1029/2000PA000562.

Martín-Fernández J.A., Thió-Henestrosa, S. (Eds.) (2016a) - Compositional Data Analysis: CoDaWork, L'Escala, Spain, June 2015. Springer Proceedings in Mathematics & Statistics, 187. Springer International Publishing, New York (USA), pp. 209.

Martín-Fernández J.A., Thió-Henestrosa, S. (Guest eds.) (2016b) - Compositional Data Analysis. Austrian Journal of Statistics, 45(4), pp.95.

Martín-Fernández J.A., Barceló-Vidal C., Pawlowsky-Glahn V. (2003) - Dealing with zeros and missing values in compositional datasets using nonparametric imputation. Mathematical Geology, 35(3), 253-278.

Martín-Fernández J.A., Hron K., Templ M., Filzmoser P., Palarea-Albaladejo J. (2015) - Bayesian-multiplicative treatment of count zeros in compositional datasets. Statistical Modelling, 15(2), 134-158.

Mateu-Figueras G., Pawlowsky-Glahn V., Egozcue J.J. (2011) - (Pawlowsky-Glahn V., Buccianti A., eds). John Wiley & The Principle of Working on Coordinates, in Compositional Data Analysis: Theory and Applications Sons, Ltd, Chichester, UK, 29-42.

Ortiz J.D., Mix A.C. (1997) - Comparison of Imbrie-Kipp transfer function and modern analog temperature estimates using sediment trap and core top foraminiferal faunas. Paleoceanography, 12 (2), 175-190.

Pailler D., Bard É. (2002) - High frequency palaeoceanographic changes during the past 140000 yr recorded by the organic matter in sediments of the Iberian Margin. Palaeogeography, Palaeoclimatology, Palaeoecology, 181(4), 431-452.

Pawlowsky-Glahn V., Egozcue J.J., Tolosana-Delgado R. (2015) - Modeling and analysis of compositional data. John Wiley & Sons, Chichester, pp. 378.

Peña D., Prieto F. (2001) - Multivariate outlier detection and robust covariance matrix estimation. Technometrics, 43(3), 286-310.

Pflaumann U., Duprat J., Pujol C., Labeyrie, L.D. (1996) - SIMMAX: a modern analogue technique to deduce Atlantic sea surface temperatures from planktonic foraminifer in deep-sea sediments.

Paleoceanography, 11, 15-36.

Prell W.L. (1985) - The stability of low-latitude sea-surface temperatures: An evaluation of the CLIMAP reconstruction with emphasis on the positive SST anomalies, Rep. TR025, 60 P., U.S. Dep. of Energy, Washington DC, pp. 60.

Prell W., Martin A., Cullen J., Trend M. (1999) - The Brown University Foraminiferal Data Base. IGBP PAGES/World Data Center-A for Paleoclimatology Data Contribution Series # 1999-027. NOAA/NGDC Paleoclimatology Program, Boulder, CO, USA.
https://www.ncdc.noaa.gov/paleo/metadata/noaa-ocean-5908.html

Rasmussen S.O., Bigler M., Blockley S.P., Blunier T., Buchardt S.L., Clausen H.B., Cvijanovic I., Dahl-Jensen D., Johnsen S.J., Fischer H., Gkinis V., Guillevic M., Hoek W.Z., Lowe J.J., Pedro J.B., Popp T., Seierstad I.K., Peder Steffensen J., Svensson A.M., Vallelonga P., Vinther B.M., Walker M.J.C., Wheatley J.J., Winstrup M. (2014) - A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records: refining and extending the INTIMATE event stratigraphy. Quaternary Science Reviews, 106, 14-28.

Sbaffi L., Wezel F.C., Kallel N., Paterne M., Cacho I., Ziveri P., Shackleton N. (2001) - Response of the pelagic environment to palaeoclimatic changes in the central Mediterranean Sea during the Late Quaternary, Marine Geology, 178(1-4), 39-62.

Schlitzer R. (2018) - Ocean Data View.
http://odv.awi.de

Schönfeld J., Zahn. R., de Abreu. L. (2003) - Stable isotope ratios and foraminiferal abundance of sediment cores from the Western Iberian Margin. Doi:10.1594/PANGAEA.733303, supplement to: Schönfeld, J., Zahn R., de Abreu L. (2003): Surface to deep water response to rapid climate changes at the western Iberian Margin. *Global and Planetary Change*, 36(4), 237-264.
Doi:10.1016/S0921-8181(02)00197-2

ter Braak C.J.F., Juggins S. (1993) - Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia, 269-270(1), 485-502.

Tolosana-Delgado R., McKinley J. (2016) - Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland). Applied Geochemistry, 75, 263-275.

Verboven S., Hubert M. (2004) - LIBRA: a MATLAB library for robust analysis.
https://github.com/duncombe/matlab/blob/master/LIBRA/rsimpls.m

Voelker A.H.L., de Abreu L. (2011) - A Review of Abrupt Climate Change Events in the Northeastern Atlantic Ocean (Iberian Margin): Latitudinal, Longitudinal and Vertical Gradients. In: Rashid H., Polyak L., Mosley-Thompson E. (Eds), Abrupt Climate Change: Mechanisms, Patterns, and Impacts, Geophysical Monograph Series (AGU, Washington D.C.), 193, 15-37.

Waelbroeck C., Labeyrie L., Duplessy J.-C., Guiot J., Labracherie M. (1998) - Improving past sea surface temperature estimates based on planktonic fossil faunas. Paleoceanography, 13, 272-283.

Wallach D., Goffmet B. (1989) - Mean squared error of prediction as a criterion for evaluating and comparing system models. Ecological Modelling, 44, 299-306.

Ms. received: July 23, 2019
Accepted: January 26, 2020
Revised: January 25, 2020,
Available online: February 6, 2020