CODAMAT: a Modern Analogue Technique for compositional data

V. Di Donato[1], J. A. Martín-Fernández[2]

[1]Dipartimento di Scienze della Terra, Università degli Studi di Napoli "Federico II", Italia; *valedido@unina.it*
[2]Departamento d'Informàtica i Matemàtica Aplicada, Campus Montilivi, Edif. P-4, 17071- Girona

Abstract

The quantitative estimation of Sea Surface Temperatures from fossils assemblages is a fundamental issue in palaeoclimatic and paleooceanographic investigations. The Modern Analogue Technique, a widely adopted method based on direct comparison of fossil assemblages with modern coretop samples, was revised with the aim of conforming it to compositional data analysis. The new CODAMAT method was developed by adopting the Aitchison metric as distance measure. Modern coretop datasets are characterised by a large amount of zeros. The zero replacement was carried out by adopting a Bayesian approach to the zero replacement, based on a posterior estimation of the parameter of the multinomial distribution. The number of modern analogues from which reconstructing the SST was determined by means of a multiple approach by considering the Proxies correlation matrix, Standardized Residual Sum of Squares and Mean Squared Distance. This new CODAMAT method was applied to the planktonic foraminiferal assemblages of a core recovered in the Tyrrhenian Sea.

Kew words: Modern analogues, Aitchison distance, Proxies correlation matrix, Standardized Residual Sum of Squares.

1 Introduction

The quantitative estimation of sea surface temperatures (SST) is a fundamental issue in palaeoclimatic and paleooceanographic investigations. The most widely adopted methods to obtain quantitative estimates from fossil assemblages are the transfer functions (Imbrie and Kipp, 1971) and the modern analogue technique (MAT) (Hutson, 1979). In both methods the palaeoestimates are calibrated with respect to modern data sets – usually consisting of core top faunas. Transfer functions are based on factor analysis and multiple regression, while the MAT is based on the comparison of a fossil assemblage with a set of modern samples through the computation of a distance measure or a similarity coefficient. The paleoenvironmental estimates are obtained from the mean value of environmental parameters measured at the location of the most similar - usually 8 or 10 - modern assemblages. For both methods the main limit is represented by the non-analogue problem, occurring when the palaeoenvironmental conditions represented in the fossil assemblages have no correspondence in modern environments. Traditional MAT application is based on squared chord distance or cos theta similarity index.

Fossil assemblages are usually expressed as relative abundance or percentages of species. Therefore, it seems logical to develop a MAT consistent with the statistical approach developed in the last years for the compositional data (Aitchison, 1986; Barceló-Vidal and others, 2001, among others).

2 The CODAMAT Method

The revised Modern Analogue Technique, here presented, was developed by taking into account three main points: a) choice of the distance measure; b) zero replacement; 3) number of analogues.

Aitchison (1992) proposed a distance measure between compositions which is equivalent to euclidean distance between log centered compositions. Martín-Fernández and others (1998) showed that the Aitchison distance and Mahalanobis distance computed on log centered data meet the requirements of permutation invariance, perturbation invariance and subcompositional dominance that are needed in order to achieve a meaningful statistical analysis of compositional data. They also showed that angular distances does not have a compositional coherent behaviour. An obvious consequence of the above mentioned statements is that a MAT conform to compositional data analysis could be based on either Aitchison distance or Mahalanobis distance between log centered data. At this stage the analysis is based on the Aitchison distance.

The compositional data analysis requires the data to be strictly positive. Due to the strong relationship existing between planktonic foraminifera and water masses properties, modern coretop datasets are characterised by a large amount of zeros. In order to reduce the number of replacements, rarer species were excluded or grouped into informal taxonomical units prior to any replacement procedure. A meaningful analysis of fossil assemblages requires the data to be expressed in the form of relative abundances. The original dataset, however, are made of counts. Following Martín-Fernández and Di Donato (2007) the zero replacement was carried out by adopting a Bayesian approach to the zero replacement, based on a posterior estimation of the parameter of the multinomial distribution with Jeffreys and Uniform priori (Walley, 1996).

Another fundamental problem concerns the selection of a proper number of modern analogues. This subject was handled by Martin-Fernández and Di Donato (2007) with a multiple approach by considering the Proxies correlation matrix, Standardized Residual Sum of Squares and Mean Squared Distance.

3 An application example

A check of the validity of this new CODAMAT approach for the estimation of summer and winter sea surface temperatures (SST) on modern conditions was carried out by means of leave-one-out validation. The results indicate, for both Jeffreys and Uniform priori, a very high correlation between the measured and estimated summer and winter SST, the correlation coefficient being respectively 0.9895 and 0.9922.

The CODAMAT was then applied to the planktonic foraminiferal assemblages of a core previously analysed with a traditional MAT approach (Di Donato and others, in press). The GNS84-C106 core (14°42’24” E; 40°28’52” N) was recovered in the Tyrrhenian sea at a depth of 292 m. The chronological framework of this core is based on eight ^{14}C AMS dates and on the occurrence of the 79 AD Pompeii pumice. This core, spanning the last 35 ka, provides a high resolution record not only of Last Glacial Period and the Holocene, but also of Late Glacial events such as the Bølling-Allerød and the Younger Dryas. Figure 1 shows the comparison between the reconstruction formerly obtained with the traditional MAT and the new ones obtained with the CODAMAT. A controversial result yielded by the traditional MAT approach rested in the high winter SST reconstructed for a Last Glacial time interval spanning from

25 ka BP and 18 ka BP. These apparently high estimates con be due to both non-analog conditions and problems related to the analysed size fraction. As shown in Figure 1, the temperature reconstruction obtained with the CODAMAT for the Last Glacial Period are lower in respect to the former ones, indicating that the new method works more efficiently in unfavourable conditions.

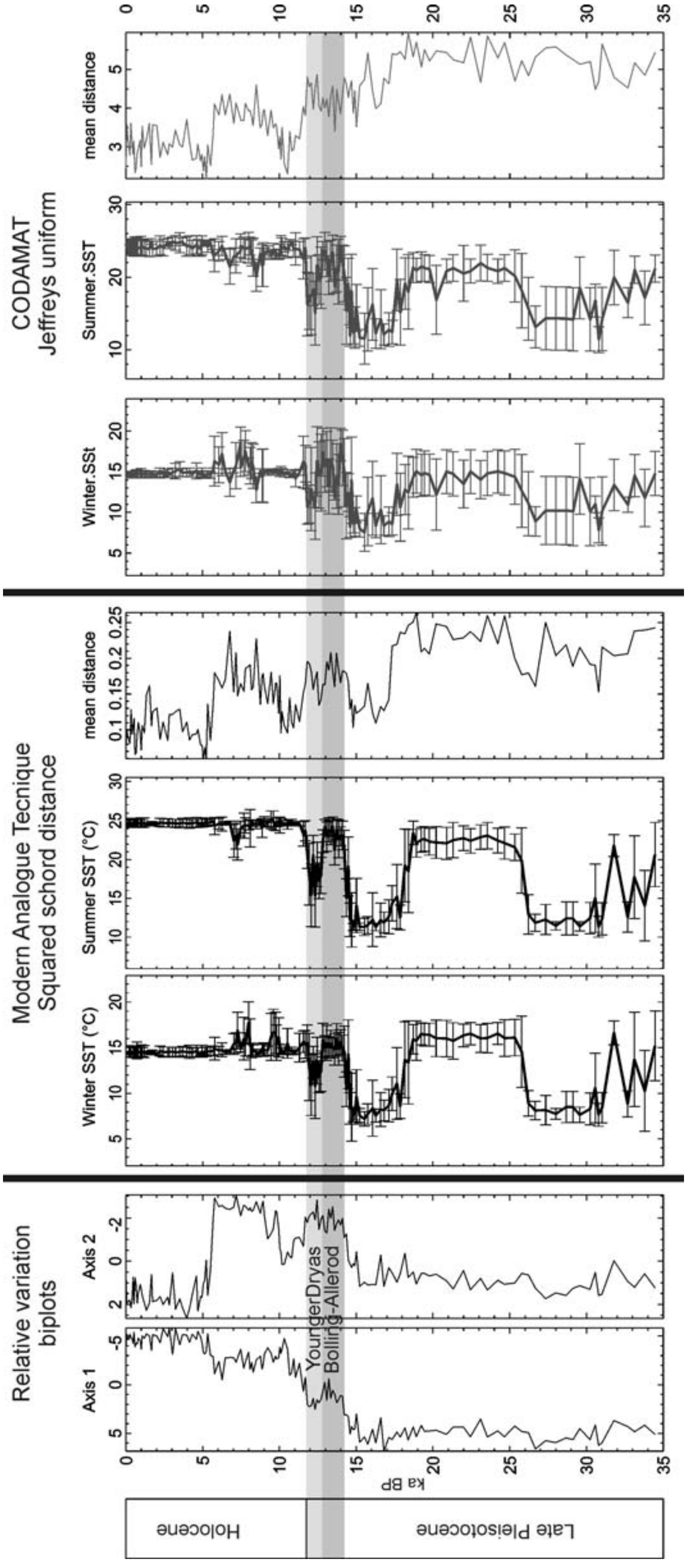

Figure 1: On the left: first and second axis form biplot scores related to planktonic foraminiferal assemblages of GNS84-C106 core. The first axis of relative variation biplot is related to Sea Surface Temperatures (Di Donato and others, in press). The following graphs show winter and summer sea surface temperatures (°C) reconstructed with MAT and CODAMAT. The bars indicate the standard deviation of the environmental parameters measured on the best modern analogs. For each reconstruction are also reported the mean distances of the selected modern analogs from the fossil ones. Higher values indicate non-analog conditions.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Aitchison, J. (1992). *On criteria of measures of compositional difference*. Mathematical Geology 24:365–379.

Barcel ó-Vidal, C., Martín-Fern ández, J.A. and Pawlowsky-Glahn, V., 2001, *Mathematical Foundations of Compositional Data Analysis*, in Proceedings of the Annual Conference of the International Assotiation for Mathematical Geology, Cancun (Mexico), CD-ROM, 20p.

Di Donato, V., Esposito, P., Russo-Ermolli, E., Cheddadi, R., Scarano, A. (in press). *Coupled atmospheric and marine palaeoclimatic reconstruction for the last 35 kyr in the Sele Plain-Gulf of Salerno area (southern Italy).*

Hutson, W.H. (1979), *The Agulhas Current during the Late Pleistocene: Analysis of modern faunal analogs*. Science, vol. 207, no. 1, p. 64-66.

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

Martín-Fernández, J.A., Barceló, C. and Pawlowsky, V. (1998), *Measures of Difference for Compositional Data and Hierarchical Clustering Methods*, in: Proceedings of the Fourth Annual Conference of the International Association for Mathematical Geology, Ed. Buccianti, A. , Nard, G. and Potenza, R. Naples (I), Part 2, pp. 526-531.

Martin-Fernández, J.A. and Di Donato, V. (2007). *New tools in Modern Analogue Techniques: the number of analogues*, In: Proceedings of XXX Congreso Nacional de Estadística e Investigación Operativa, Valladolid, Spain, September 25-28, p. 117. CD-ROM.

Walley, P. (1996). *Inferences from Multinomial Data: Learning about a Bag of Marbles*. Journal of the Royal Statistical Society. Series B (Methodological), Vol. 58, No. 1. (1996), pp. 3-57.