Mathematical Geosciences (2025) 57:1305–1326
https://doi.org/10.1007/s11004-025-10205-5

A Constrained Divisive-Compositional Approach to Micropalaeontological Ecozonation

Valentino Di Donato[1] · **Josep Antoni Martín-Fernández**[2]

Received: 18 July 2024 / Accepted: 18 May 2025 / Published online: 13 June 2025

Abstract

Zonation of stratigraphic successions is a key practice for identifying intervals characterised by stability or, conversely, by palaeoenvironmental changes. Stratigraphically constrained agglomerative algorithms have been commonly adopted to obtain zonation based on quantitative palaeontological data. Here we explore constrained divisive algorithms aiming at obtaining a zonation that meets the principle of maximizing coefficients commonly used to evaluate the effectiveness of clustering algorithms. In particular, a constrained version of Cavalli Sforza's method was applied, together with an algorithm conceived to maximise, at each division, the average silhouette width of the observations. The results were compared, following the compositional data analysis properties, with those obtained with a commonly adopted agglomerative method. When evaluated on artificial data, the divisive algorithms show stability and a tendency to identify the boundary between intervals at the midpoint of transitions, consistently with common stratigraphic practice. Overall, the application to real micropalaeontological data, consisting of percentages of planktonic foraminifera, provide reasonable zonation patterns with all algorithms considered. For the main partition, the constrained version of Cavalli Sforza's method provides highest values of Calinski–Harabasz and Hartigan indexes, while the average silhouette width method, as expected, performs better in the evaluation of average silhouette width index as well as of Goodman–Kruskal's coefficient and cophenetic correlation. One potential issue to consider is the tendency to define single sample intervals as the number of partitions increases.

Keywords Agglomerative hierarchical clustering · Aitchison distance · Compositional data · Constrained clustering · Divisive hierarchical clustering · Zonation

✉ Josep Antoni Martín-Fernández
josepantoni.martin@udg.edu

Valentino Di Donato
valedido@unina.it

[1] Università degli Studi di Napoli Federico II, Naples, Italy

[2] Department of IMAE, University of Girona, Edifici P4, Campus Montilivi, 17003 Girona, Spain

1 Introduction

Ecozonation methodologies are widely adopted, in both continental and marine successions, to define high-resolution biostratigraphic and chronostratigraphic schemes (Capotondi et al. 1999). The concept behind this approach is that changes in the composition of fossil assemblages are proxies of palaeoenvironmental or palaeoclimatic changes. Thus, the definition of ecozones enables the identification of intervals characterised by relative stability or, conversely, by stratigraphical levels which mark environmental changes. Once their boundaries have been dated, ecozones may be useful to correlate, at regional/sub-basin scale stratigraphic successions. There are several application examples of palaeontological ecozonation of both continental and marine successions (Capotondi et al. 1999; Sprovieri et al. 2003; Siani et al. 2010; Allen et al. 1999). The boundaries between zones can be defined through simple inspection of stratigraphic diagrams, looking for a bioevent that is of interest to the researcher. However, the definition of ecozones may be strengthened by adopting stratigraphically constrained classification algorithms, which allow ecozone boundaries to be objectively defined. In these algorithms, the objects (samples or clusters of samples) can only be joined to immediately preceding or succeeding objects according to stratigraphic order. The CONSLINK method (Gordon and Birks 1972), as an example, is based on a single linkage criterion. The constrained incremental sum of squares (CONISS) method (Grimm 1987) has been widely adopted in micropalaeontological studies and, likely, is the most successful zonation algorithm. Basically, CONISS is a stratigraphically constrained Ward's method (Ward 1963), based on the constraint of minimum increase of within groups (ecozones) variance. As a modified Ward algorithm, the CONISS method can be computed from a distance matrix based on Euclidean distance as the measure of difference between samples. The application of CONISS to micropalaeontological data expressed in terms of percentages requires an approach conforming to the nature of the compositional data (CoDa) (Aitchison 1986). On this basis, Di Donato et al. (2008), Di Donato et al. (2009) defined compositional intervals based on the CONISS method applied to centred log-ratio (clr) data. The rationale behind this approach is that the Ward's method is correctly applied and, at the same time, a distance measure conceived for CoDa is adopted. The CONISS method is a hierarchical agglomerative method because it creates a hierarchical structure of groups from the bottom (each sample forms a group) to the top (a single group contains all samples) (Hennig et al. 2015).

In contrast, a hierarchical divisive algorithm operates from the top to the bottom, recursively splitting a group into two groups. Among them, the Cavalli-Sforza method (Edwards and Cavalli-Sforza 1965) can be considered as the divisive counterpart of Ward's method (Ward 1963), being focused on minimising within-group sum of squares and, by converse, maximising the between-group variance. In a constrained context, Gill (1970) proposes a divisive approach to zonation of univariate stratigraphical data based on analysis of variance, although not developed in a cluster analysis context. Overall, this divisive approach follows a criterion also adopted in time series analysis to identify abrupt changes in signal (Killick et al. 2012). A classical divisive Euclidean approach was also explored by Gordon and Birks (1972) with the SPLITSQ and SPLITINF algorithms.

In recent decades, clustering algorithms based on unusual ratio-type formulas have been proposed. For example, the average silhouette width (ASW) (Rousseeuw 1987; Kaufman and Rousseeuw 1990), which was introduced for measuring cluster quality, has also been considered as a clustering criterion (Batool and Hennig 2021). In a comparative study of divisive and agglomerative clustering algorithms, Roux (2018) highlighted that divisive algorithms based on unusual ratio-type formulas (e.g., the ASW formula) perform efficiently and, in fact, slightly better than their agglomerative counterparts.

In this paper, we explore the behaviour of divisive (DCONSIL) and agglomerative (ACONSIL) constrained algorithms based on the ASW in a context involving CoDa analysis of micropalaeontological data. Furthermore, an algorithm (CONCS) based on the Cavalli-Sforza method (Edwards and Cavalli-Sforza 1965) is also evaluated. The latter can be regarded as a divisive full counterpart of the agglomerative CONISS method. Adoption of log-ratio techniques (Aitchison 1986) overcomes problems associated with adopting typical techniques for CoDa, whose sample space, the simplex, has properties for which classical algebraic/geometric operations are neither subcompositionally coherent nor scaling invariant (Pawlowsky-Glahn et al. 2015).

In the following sections, the algorithms are described together with some pretreatment of micropalaeontological data that should be applied before carrying out the analysis. Examples applying the method to planktonic foraminiferal assemblages of Mediterranean Sea marine cores and to an artificial dataset are used to evaluate the differences between the algorithms.

2 Constrained Hierarchical Clustering Methods

In cluster analysis, one of the objectives is to obtain groups whose entities are as homogeneous as possible, while entities in different clusters are heterogeneous (Hennig et al. 2015). There are more than several algorithms to perform a cluster analysis. Among them, Ward and Cavalli-Sforza methods are distinctive in that they focus on minimising the within-group variance and, by converse, maximising the between-group variance. In fact, the Cavalli-Sforza method is a divisive method which follows, in some ways, the same approach as the Ward method. In this method, a dataset is divided progressively with the constraint of maximising, at each step, the between-groups sum of squares (the general sum of the squared distances is the sum of within-groups and between-groups sum of squares). In practice, the two groups obtained by partitioning an undivided dataset are further divided, while they are made by at least two samples, so that for successive division cycles, an increasing number of clusters is obtained. Overall, this divisive approach follows a criterion adopted in time series analysis to identify abrupt changes in signal (Killick et al. 2012).

To use these clustering algorithms as zonation techniques, they must be modified with the constraint of preserving the stratigraphic order of samples. As pointed out above, based on this concept, Grimm (1987) modified Ward's method to obtain the CONISS algorithm. With the same approach, the introduction of a stratigraphic constraint makes it possible to employ the Cavalli-Sforza method (hereafter CONCS) to divide a stratigraphic succession into zones. Being a divisive algorithm, the CONCS

method is computationally very expensive: as an example, the initial division of an n by D dataset (where n is the number of observations and D the number of variables) would require the examination of all $2^{n-1}-1$ partitions, although this number can be reduced to $(2^D-2)\binom{n}{D}$ (Scott and Symons 1971). Introducing a stratigraphic constraint allows this number to be substantially reduced, since the partitions to consider for the initial division are only n-1 (the succession must simply be divided into two intervals). To perform divisive clustering and construct the dendrogram, we considered two approaches. In the first version of CONCS (CONCS version 1), the algorithm is designed to divide, at each step, the group whose partitioning yields the highest between-groups sum of squares (SSB). In this version, node levels can represent the SSB of the partitioned groups. Alternatively, to reduce the number of inversions often observed in constrained dendrograms, node levels can be defined by the diameter of the successive clusters (measured as the largest dissimilarity between objects within a cluster), as in the DIANA method (Kaufman and Rousseeuw 1990). As a third possibility, the total within-groups sum of squares (SSW) can be considered for node levels. This choice has the advantage of providing monotonic node levels. In the second approach (CONCS version 2), the algorithm is set to partition, at each step, the cluster with the largest SSW, while still maximizing the resulting SSB. In this version, node levels can be based on either the SSW of the cluster being partitioned or the total SSW. This approach has the advantage of producing monotonic node levels and is computationally less expensive, as it does not require a pre-evaluation of which group to partition. For this reason, we preferred this second version in the examples provided. It is worth noting that the two versions do not lead to substantial differences in the resulting zonation, except in some cases where the order of cluster partitioning varies. A comparison of the dendrograms obtained using the two approaches is presented in Sect. 5.

The other divisive algorithm that has been evaluated in a constrained context is based on the principle of maximising, for each partition, the ASW of the observations (hereafter CONSIL). To compute the silhouette width, it is necessary, first, to compute the mean squared dissimilarity of an observation $\mathbf{x}_i$ belonging to a cluster C_k to all other $n_{(k)}-1$ observations of the same cluster, $d_{i,C_k} = \frac{1}{n_k-1}\sum_{j\in C_k, i\neq j} d^2(\mathbf{x}_i,\mathbf{x}_j)$, and second, to compute the smallest value, $d_{i,C} = min_l d_{i,C_l}$, of the average squared dissimilarity of the observation $\mathbf{x}_i$ to observations of any other cluster, C_l, $d_{i,C_l} = \frac{1}{n_l}\sum_{j\in C_l, l\neq k} d^2(\mathbf{x}_i,\mathbf{x}_j)$, by means of which the "closest" cluster C is found. The silhouette width value on an observation $\mathbf{x}_i \in C_k$ is thus defined by $s_i = \frac{d_{i,C}-d_{i,C_k}}{\max(d_{i,C},d_{i,C_k})}$. The values of s_i are constrained in the interval $[-1, 1]$. Values close to 1 indicate that the observation is well classified, values around 0 suggest that the observation is in between two clusters, while values around -1 indicate a wrong classification. The ASW is thus defined by the mean of all silhouette values computed on the n observations in the dataset, $\text{ASW} = \frac{1}{n}\sum_{i=1}^{n} s_i$. The higher this value, the better is the overall classification. We considered both a divisive (DCONSIL) and an agglomerative (ACONSIL) constrained algorithm, both aimed at maximising the ASW. It is worth noting that the ASW values obtained through successive partitions or agglomerations tend to be distinctly non-monotonic, making dendrograms constructed with ASW values at the node levels difficult to interpret (e.g., Fig. 3 in Sect. 5). Therefore, to

construct the dendrograms, we followed the approach used for CONCS, employing the SSW (or alternatively, the diameters) as node levels. The two algorithms differ in their structures. In the divisive approach, the cluster selected for partitioning is the one with the largest SSW (whose value can also be used as the node level), and it is subsequently subdivided to maximise the ASW. In contrast, the agglomerative approach performs aggregation at each step based on achieving the maximum ASW, after which the total SSW is used to determine the node levels.

To evaluate the performance of the algorithms, we considered the cophenetic correlation (CC) and, following Roux (2018), the Goodman–Kruskal coefficient (GK) (Goodman and Kruskal 1954). In general, the hierarchical cluster tree matrices used in the algorithms do not include linkage distances. For instance, in Ward's method, the hierarchical cluster tree is based on the incremental sum of squares. To compute the CC, we reconstructed these distances from the hierarchical cluster tree by calculating the distance of a newly added observation to the centroid of the already-formed cluster. The GK coefficient can be computed in the following way: Let $d(\mathbf{x}_i, \mathbf{x}_j)$ and $u(\mathbf{x}_i, \mathbf{x}_j)$ be, respectively, the input and ultrametric (the level at which objects $\mathbf{x}_i$ and $\mathbf{x}_j$ are linked in the dendrogram) distance between a pair of objects $\mathbf{x}_i$ and $\mathbf{x}_j$. Two pairs of objects $(\mathbf{x}_i, \mathbf{x}_j)$ and $(\mathbf{x}_k, \mathbf{x}_l)$ are said to be concordant if $d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_k, \mathbf{x}_l)$ and $u(\mathbf{x}_i, \mathbf{x}_j) < u(\mathbf{x}_k, \mathbf{x}_l)$. Conversely, they are said to be discordant if $d(\mathbf{x}_i, \mathbf{x}_j) < d(\mathbf{x}_k, \mathbf{x}_l)$ and $u(\mathbf{x}_i, \mathbf{x}_j) > u(\mathbf{x}_k, \mathbf{x}_l)$. If S^+ and S^- are the number of concordant and discordant pairs of distances, then the GK coefficient is given by $GK = \frac{S^+ - S^-}{S^+ + S^-}$. The GK coefficient ranges from 0 (no concordance) to 1 (perfect concordance), with higher values indicating better clustering quality. In general, the zonation of a stratigraphic succession is aimed at defining a relatively restricted number of intervals. Therefore, it is also necessary to consider the effectiveness of the algorithm with regard to the first few partitions only, or with regard to the optimal number of intervals. There are several methods which suggest an optimal number of groups to be retained in cluster analysis. In Roux (2018), as an example, the Dunn index is considered. Among them, we considered that proposed by Mojena (1977), which is based upon statistical stopping rules. These rules utilise $n - 1$ items in the distribution of the criterion α (the criterion by which the algorithm forms the clusters) by calculating the mean and standard deviation of the sample. The values of the criterion can range from α_0 (n clusters) to α_{n-1} (one single cluster). The stopping rule proceeds by defining a significant α as the one that lies in the upper tail of the distribution satisfying $\alpha_j \leq \overline{\alpha} + k \cdot s_\alpha$, where α_j represents the value of the criterion at stage j, k is the standard deviation, and $\overline{\alpha}$ and s_α are the mean and unbiased standard deviation of the distribution, respectively. In Mojena (1977) datasets including 60 and 120 observations were evaluated and, in terms of predicted number of clusters, k values ranging from 2.75 to 3.5 gave the best overall results (in general, with increasing k values, fewer clusters are predicted). We also ran some simulations, and the results suggest a dependence of k on the number of observations. For datasets with n=200 and D=10, k values around 3.5 reasonably predicted the number of clusters considering as criteria both node levels obtained with CONISS and cluster diameters obtained with CONCS and DCONSIL. Instead, when considering the SSW as criterion, a lower k value, around 1.8, provided better predictions.

Other procedures for making cluster number decisions are based on the ratios of between-cluster sum of squares measurements to within-cluster sum of square measurements, such as the Calinski–Harabasz index (Calinski and Harabasz 1974), which provide information on the separation and homogeneity of the clusters. The Calinski–Harabasz index is defined by $\mathrm{CH}_{n_c} = \frac{trace(\mathbf{B}_{n_c})/(n_c-1)}{trace(\mathbf{W}_{n_c})/(n-n_c)}$ in which n_c indicates the number of clusters, and $\mathbf{B}_{n_c}$ and $\mathbf{W}_{n_c}$ are the matrixes of the between- and within-group sums of squares, respectively. This index can be computed for any possible n_c, and its largest value suggests the optimal number of intervals (each of which, obviously, may be divided into subintervals).

It can be noted that indices such as those described above can be used to both define the number of clusters (this is also the case of the ASW), and to evaluate the performance of the algorithms. In effect, in our study the CH_{n_c} index was used twice: first, to define an optimal number of intervals. Next, the maximum values achieved by applying the different algorithms were considered to compare their performance in relation to a reduced number of partitions. Together with the CH_{n_c} index, we considered the ASW index described above and the Hartigan index (H) (Hartigan 1975) defined as $H_{n_c} = \ln \frac{trace(\mathbf{B}_{n_c})}{trace(\mathbf{W}_{n_c})}$. The H_{n_c} values tend to increase as the number of clusters increases, and consequently the within-group sum of squares decreases. The values obtained for only the first few partitions may help to compare the performance of the different algorithms. In doing that, higher performance may be obviously expected for CONSIL in the evaluation of the ASW index. The CH_{n_c} and the H_{n_c} indices provide information on the separation among clusters and their homogeneity. CONISS and CONCS are both based on evaluation of the between/within sums of squares of the clusters. It thus seems interesting to analyse how the two algorithms are evaluated by these indices. Further, we explored in more detail the interpretation and significance of the differences between the clusters by means of CoDa techniques such as relative variation biplots (RVB) (Aitchison and Greenacre 2002), multivariate analysis of variance (MANOVA) contrast, R-mode cluster analysis (Martín-Fernández et al. 2023) and geometric mean bar plots (GMBP) (Martín-Fernández et al. 2015). Among these four techniques, the first three involve applying the classical methods of biplot, MANOVA, and R-mode clustering to the log-ratio coordinates of CoDa (Sect. 3). In contrast, the GMBP is a technique specifically designed for CoDa. When analysing data with two or more groups, a geometric-mean bar plot can be used to visually compare their centres. For each group, we first calculate the ratio of the overall geometric mean to the group-specific geometric mean. Then, these ratios are displayed in a bar plot using a logarithmic scale. If the group's centre matches the overall centre, the ratio for each part will be 1, corresponding to a value of zero on the log scale. However, when the group's centre differs from the overall centre, the ratio deviates from 1, producing a positive or negative logarithmic value. Larger bars, whether positive or negative, reflect greater differences between the group and overall means.

3 Application to Fossil Assemblages from Sediment Cores

In general, fossil assemblage data arise from counting of specimens from samples. It is quite intuitive to consider that the Euclidean distance is strongly determined by the total of assemblages rather than the relative ratios between the components. Since in the analysis of fossil assemblage compositions the total of the vector is not informative, this problem may be in part circumvented by using angular distance measures to define the similarity (or distance) between assemblages, such as the squared chord distance. It can be noted, however, that the CONISS algorithm cannot be properly applied on a distance matrix built with squared chord distance (Palarea-Albaladejo et al. 2012; Di Donato et al. 2009, 2019).

The analysis of relative abundance (percentage) data does not suffer of a "size of samples" effect. However, it is important to define an appropriate distance measure among compositions expressed in terms of percentages. Percentage (closed) data only bring relative information; thus, they require a specific statistical framework, such as that represented by the CoDa analysis (Aitchison 1986). In recent years, CoDa tools have been proposed to also include the total, if it is of interest, in the analysis (Pawlowsky-Glahn et al. 2015). However, these issues are beyond the scope of this article. Some statistical techniques (e.g., PCA) can be performed following CoDa properties of the Aitchison geometry (Pawlowsky-Glahn et al. 2015) on the the centred log-ratio (clr) scores defined as $\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_D}{g(\mathbf{x})}\right)$, for a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$, where $g(\cdot)$ is the geometric mean of $\mathbf{x}$. Based on the definition of a compositional inner product between to compositional vectors $< \mathbf{x}, \mathbf{y} >_A = < \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) >_E$, the Aitchison distance is derived as $d_A(\mathbf{x}, \mathbf{y}) = d_E(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}))$, where the subindex "A" and "E" mean Aitchison and Euclidean, respectively. These elements allow us to create an orthonormal log-ratio (olr) basis in the sample space of CoDa (the simplex S^D). An olr-basis can be created using a data-driven method such as principal balances (Martín-Fernández et al. 2018) or R-mode cluster analysis (Martín-Fernández et al. 2023). Alternatively, the knowledge of the researcher can be used to improve the interpretation of the models when creating the olr-basis by a sequential binary partition (SBP) process (Egozcue and Pawlowsky-Glahn 2005). The olr-coordinates of the general form can be defined as

$$\mathrm{olr}(\mathbf{x})_k = \sqrt{\frac{n_k \cdot d_k}{n_k + d_k}} \ln \frac{(x_{i_1} \cdots x_{i_{n_k}})^{1/n_k}}{(x_{j_1} \cdots x_{j_{d_k}})^{1/d_k}}, \; k = 1, \ldots, D-1, \tag{1}$$

where n_k and d_k are the number of parts in the numerator $(x_{i_1}, \ldots, x_{i_{n_k}})$ and in the denominator $(x_{j_1}, \ldots, x_{j_{d_k}})$, respectively. Note that the olr-coordinate in Eq. (1) is a "balance" between the average of two sets of parts, which generalises the expression of the clr-scores, balancing one part of the composition against the average of the others. Consequently, any multivariate analysis can be performed on the olr-coordinates of the compositions (Pawlowsky-Glahn et al. 2015). Due to the challenges associated with clr-scores in terms of subcompositional coherence and the degeneration of the covariance matrix, it is preferable to use olr-coordinates in analyses that require probabilistic models (Pawlowsky-Glahn et al. 2015), such as the simulation of artificial data

(Sect. 5). On the other hand, the Aitchison distance (Aitchison et al. 2000) is equal to the Euclidean distance between both the clr-scores or olr-coordinates (Palarea-Albaladejo et al. 2012). Being Euclidean, this distance allows the adoption of algorithms based on calculating the sum of squares. It is important to note that, because the distances are invariant under change of basis, the divisive algorithms based on SSB and SSW are invariant as well. As a consequence, there is no difference in the results if olr-coordinates are obtained by means of principal balances, by means of R-mode cluster analysis or any other particular SBP.

Log-ratio formulae can only be applied to strictly positive data. Thus, before calculating the balances, a zero values substitution is needed. In our case, this substitution was done by following the approach of Palarea-Albaladejo and Martín-Fernández (2015). In order to reduce the number of zero values to be substituted, amalgamation of parts into informal taxonomical groups can be considered.

4 Application Examples

To evaluate the performance of algorithms taken into account, we considered a simple experiment based on an artificial dataset and two application examples with literature datasets consisting of planktonic foraminiferal assemblages. An interesting aspect of palaeontology is the transition between different data groups. The bivariate artificial dataset consisted of four stratigraphically constrained non-overlapping groups, each with 60 observations. For each group, we considered a bivariate normal distribution with a diagonal covariance matrix with variances equal to 0.01. Although made up of real numbers, the dataset was conceived as representing a set of olr-coordinates derived from a three-part compositional dataset. In this regard, it is worth recalling that the normal probability function of a random composition in S^D is given by $f_{\mathbf{x}}(\mathbf{x}) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^{(D-1)}}} \exp[-\frac{1}{2}(\mathrm{olr}(\mathbf{x})-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{(-1)}(\mathrm{olr}(\mathbf{x})-\boldsymbol{\mu})]$, where μ and Σ denote the expected value and the covariance matrix of the vector of coefficients $\mathrm{olr}(\mathbf{x})$, respectively (Mateu-Figueras et al. 2003). Intervals were numbered from top to bottom. In between the groups II/I and IV/III we inserted 12 and 6 observations, respectively, to simulate a "gradual" and a "faster" transition. The transition from group III to II was considered as an abrupt shift. Artificial compositions were numbered from 1 (top) to 252 (bottom). To better assess the behaviour of the algorithms considered, we replicated the generation of the dataset and the analysis 30 times.

The planktonic foraminiferal assemblages were obtained from two cores retrieved in the Mediterranean Sea. The record of the Core TEA-C6 (Gulf of Taranto, Ionian Sea–Central Mediterranean Sea) covers the last 15 ka, while that of Core GNS84-C106 (Gulf of Salerno, Tyrrhenian Sea, Western Mediterranean Sea) covers the last 33 ka. Compositional zones for cores TEA-C6 and GNS84-C106 were formerly obtained by means of CONISS applied to the clr-scores of the CoDa. For methods and details on the age models of these cores, the reader is referred to Di Donato et al. (2009) and Di Donato et al. (2019). The two datasets considered here consist of 11 clr-scores and, respectively, 228 and 144 samples.

5 Results

5.1 Artificial Dataset Experiment

An example of the artificial datasets generated, shown in terms of a three-part composition and the corresponding olr-coordinates, is presented in Fig. 1. In all the runs performed with simulated data, both the CH_{nc} index and Mojena's stopping rule correctly identified the four predefined groups. For this example, CONCS and DCONSIL produced slightly higher cophenetic correlation (CC) values, while the Goodman–Kruskal (GK) coefficient values were fairly similar across methods (Table 3). For these datasets constructed with four different intervals, the focus was on how the algorithms defined the transitions between intervals and the CH_{nc}, H_{nc}, and ASW values obtained for a four-cluster partition. Across all runs, CONCS, DCONSIL, and CONISS consistently identified the abrupt boundary between intervals II/III (observation 128/129). However, for transitions I/II (gradual) and III/IV (rapid), the algorithms showed some variation. Divisive algorithms (CONCS and DCONSIL) reliably detected boundaries in the middle of transitions: between observations 65/66 for I/II and 190/191 for III/IV. In contrast, CONISS showed more variability. For transition I/II, CONISS identified the boundary at the beginning of the transition (observation 68/69) in 18 out of 30 cases, while in 10 cases it placed the boundary at the top (observation 60/61) and in one case slightly earlier (observation 61/62). In only one case did it coincide with the boundary detected by CONCS and DCONSIL. A similar pattern was observed for transition III/IV. CONISS placed the boundary at the beginning of the transition (observations 191/192 or 192/193) in 14 out of 30 runs or at the end of the transition (observation 188/189) in 10 runs. Summarizing, the divisive algorithms consistently identify interval boundaries at the middle of transitions, which aligns with stratigraphic practices (e.g., in isotopic stratigraphy, Marine Isotopic Stage terminations are placed mid-transition). In contrast, CONISS tends to place boundaries either near the base or the top of transitions, with some variability due to randomness. A notable distinction between CONCS and DCONSIL is evident in the dendrograms shown in Fig. 2a. In CONCS, transitions are grouped into subclusters, whereas in DCONSIL, observations within transitions are separated sequentially.

For the ASW metric, CONCS and DCONSIL consistently produced positive silhouette values for all observations. CONISS, on the other hand, often produced negative silhouette values for observations around transition I/II (e.g., 60/61 or 68/69) (Table 1). This discrepancy resulted in overall higher ASW values for CONCS and DCONSIL compared to CONISS in a four-interval zonation (Table 2). Similarly, slightly better performance for CONCS and DCONSIL was suggested by the CH_{nc} and H_{nc} indices, which indicate better separation of groups 1 and 2 as defined by the divisive algorithms.

The results obtained with ACONSIL are more problematic. As shown in Fig. 2b, both abrupt and gradual transitions were inconsistently identified across different runs. For example, the transition II/III was detected at observation 128/129 in 16 out of 30 runs, but in the remaining runs, it was identified in different positions or, in some cases, not recognised at all in the four-interval partition. Similar problems were observed for

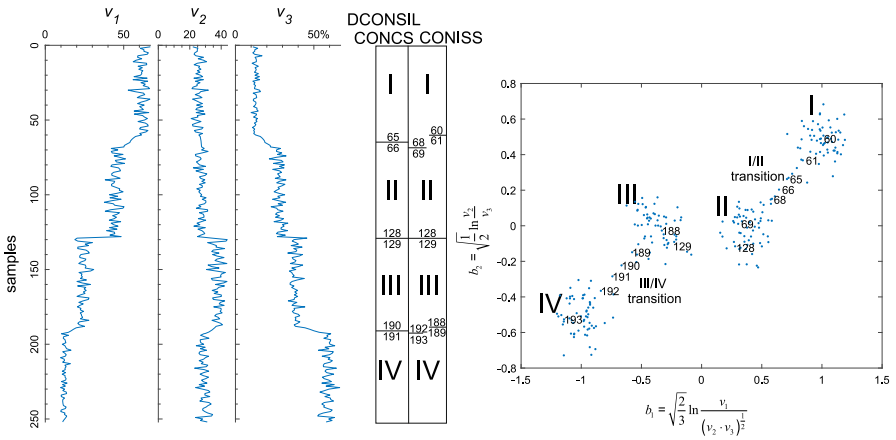

Fig. 1 Diagrams (left) of a three-part composition of the artificial dataset and scatter plot (right) of corresponding olr-coordinates (balances). Roman numerals indicate the intervals from top to bottom. Samples corresponding to boundaries between zones are highlighted

the other transitions, highlighting instability and high sensitivity to random variation. Furthermore, the dendrograms generated by ACONSIL showed a tendency to form chains of successively aggregated observations. Overall, these results suggest that the lack of stability in ACONSIL makes it unsuitable for achieving consistent zonation schemes. Comparisons of the two methods based on ASW showed that DCONSIL significantly outperformed ACONSIL, consistent with, but extending, the findings of Roux (2018).

5.2 Planktonic Foraminiferal Assemblages

Figure 3 shows dendrograms obtained for the Core TEA-C6 with CONCS and DCONSIL where different options for the node levels are used (Sect. 2). As pointed out above, by adopting SSW as a criterion for the nodes the resulting dendrogram does not include inversions except by the CONISS (Fig. 3g). On the other hand, nodes using SSB (Fig. 3a) and ASW (Fig. 3f) include inversions, with the latter being the worst case. Figures 4 and 5 show the zonation obtained with the agglomerative and divisive algorithms on the foraminiferal assemblage data of the cores TEA-C6 and GNS84-C106. Due to the weaknesses identified in the artificial case study, we have not considered the algorithm ACONSIL in the results. According to the CH_{nc} index both successions can be divided into two main intervals. The Mojena stopping rule suggests two main intervals for the Core TEA-C6 datasets, and three main intervals for the Core GNS84-C106 dataset. In Figs. 4 and 5, two main intervals are considered, with further subdivisions into subzones.

The zonation schemes appear well outlined, yet, for the Core TEA-C6, there is a difference in the main partition, since Compositional Zone 1 (CZ1) obtained with CONISS and DCONSIL coincides with the base of the Holocene, while the zone CONCS-CZ1 includes only the last 11 ka. Instead, in CONCS the base of the Holocene would be recognised at subinterval level as a further partition of Zone CONCS-CZ2a.

In this example CONCS and CONISS show similar behaviour to that observed in the artificial data experiment. In fact, with CONCS, the division between the two main intervals falls in the middle of the transition between late glacial and fully Holocene assemblages that are established around 10 ka BP. Moreover, the two halves of the transition could be recognised as subintervals. CONISS, instead cuts the two main intervals at 11.65 ka BP, at the beginning of the transition. The behaviour is even more evident as the zonation schemes are compared (Fig. 6(down)) with scores of the two first components of the clr-biplot computed from planktonic foraminiferal assemblages (Fig. 8(up)). All the algorithms give evidence to a Middle Holocene (Northgrippian stage) interval characterised by a relative increase in *Neogloboquadrina incompta* and *Globorotalia inflata* (CONCS and CONISS CZ1b subzones) and to their strong decrease at around 6 cal ka BP. Subdivision into subzones obtained with DCONSIL is a bit more complicated. In fact, the division of the zone DCONSIL-CZ1 also yields a single element interval (DCONSIL CZ1-1b), while a Middle Holocene division could be recognised as a further division of Subzone CZ1-1a.

As in the previous example, for Core GNS84-C106 two main intervals were considered (Fig. 5). CONCS and DCONSIL provided the same zonation scheme, while CONISS generated a slightly different one. In fact, Zones CONCS-CZ1 and

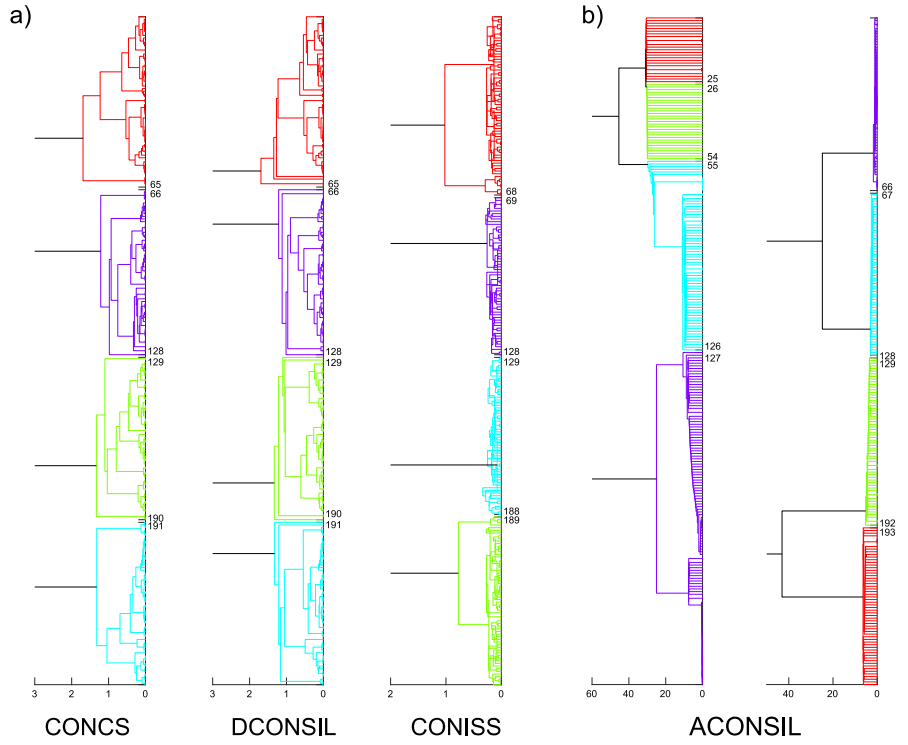

Fig. 2 **a** Dendrograms obtained from artificial dataset experiment with CONCS, DCONSIL and CONISS (run 3). **b** Two examples of dendrograms obtained from artificial dataset experiment with ACONSIL (runs 3 and 10). Tick marks indicate the observations corresponding to the boundaries between intervals

Table 1 Silhouette values for observations located around the transitions between intervals (CONISS results are reported from both the outputs of the runs 3 and 10 of the artificial dataset experiment)

Observation	CONCS/DCONSIL	CONISS (run 3)	CONISS (run 10)
Transition I/II			
60	0.961	0.946	0.936
61	0.842	0.842	−0.781
62	0.704	0.718	−0.587
63	0.525	0.557	−0.339
64	0.331	0.382	−0.069
65	0.236	0.296	0.060
66	0.396	−0.359	0.578
67	0.707	−0.689	0.792
68	0.735	−0.718	0.811
69	0.935	0.941	0.915
Transition II/III			
127	0.896	0.909	0.881
128	0.926	0.925	0.912
129	0.718	0.700	0.802
130	0.816	0.810	0.946
Transition III/IV			
188	0.946	0.949	0.940
189	0.742	−0.734	0.799
190	0.270	−0.247	0.410
191	0.410	0.441	−0.319
192	0.817	0.820	−0.801
193	0.946	0.935	0.919

Table 2 Calinski–Harabasz index, Hartigan index, and Average Silhouette Width for a four-interval zonation of the artificial dataset experiment

Metric	CONCS/DCONSIL	CONISS (run 3)	CONISS (run 10)
Calinski–Harabasz index	2,612.78	2,216.75	2,425.84
Hartigan index	3.45	3.29	3.38
Average Silhouette Width	0.897	0.877	0.881

DCONSIL-CZ1 correspond to the Holocene, while the CONISS-CZ1 encompasses the Late Glacial and the Holocene. As in the previous examples, with CONCS (and in this case also with DCONISS) intervals are defined so that the limit falls in the middle of the transition between 15 ka PB and 10 ka BP, during which full glacial assemblages are replaced by post-glacial ones (Figs. 5 and 6(up)). With CONISS, instead, the boundary is located at the beginning of the transition from the Last Glacial period to the Late Glacial. As in the Ionian Sea record, the Middle Holocene compositional

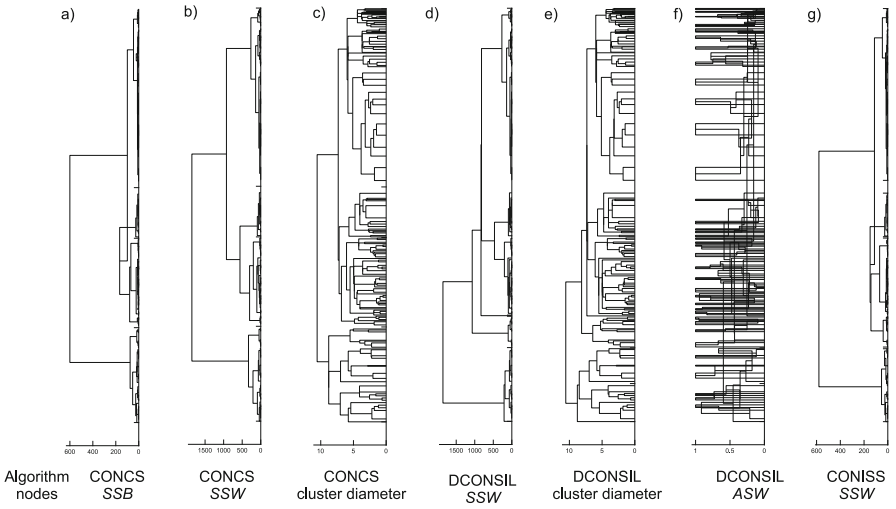

Fig. 3 Dendrograms obtained for the Core TEA-C6 with CONCS and DCONSIL and the different options described in Sect. refsec:Methods

change characterised by a marked decrease in *N. incompta* and, to a lesser extent, *G. inflata* recorded around 5−5.5 cal ka BP, has been detected at the subinterval level (CZ1a/b, regardless of the algorithm adopted).

In evaluating algorithmic performance for the two examples based on real data, for Core Tea-C6, higher GK and CC values are obtained with DCONSIL, while for Core GNS84-C106, higher CC and GK values are obtained with both CONCS and DCONSIL (Table 3). As regards CH_{nc} and H_{nc} indexes, CONCS always provides, as expected, the highest values for a 2-interval zonation (Fig. 7) and it can obviously be expected that further partition of each group would produce the highest values of these indexes computed on the two subgroups. In terms of overall evaluation, in the case of the Core TEA-C6 dataset, it can be noted that with an increasing number of subintervals up to the six subzones considered, CONISS produces very slightly higher values than CONCS, while lower values are obtained with DCONSIL. In the case of the Core GNS84-C106 dataset, the algorithms generate quite similar values. In particular, CONISS gives slightly lower CH_{n_c} and H_{n_c} values for the 2-interval main partition but higher for a three-interval zonation. In the evaluation of the ASW index, DCONSIL obviously provides the highest value for the main partition, although not much higher than CONISS and CONCS. In the case of the Core Tea-C6, for a three-interval zonation, the DCONSIL provided higher ASW values than other algorithms, in conjunction with the identification of the quite short Subzone CONSIL-CZ1c. Conversely, for a five- or six-subzone zonation, values are lower.

Regarding the expensiveness of the algorithms, the elapsed times required for the analyses are reported in Table 3c. For our datasets, the analyses were carried out in less than a second with both CONCS (both version 1 and 2) and DCONSIL, while a higher elapsed time was required by ACONSIL.

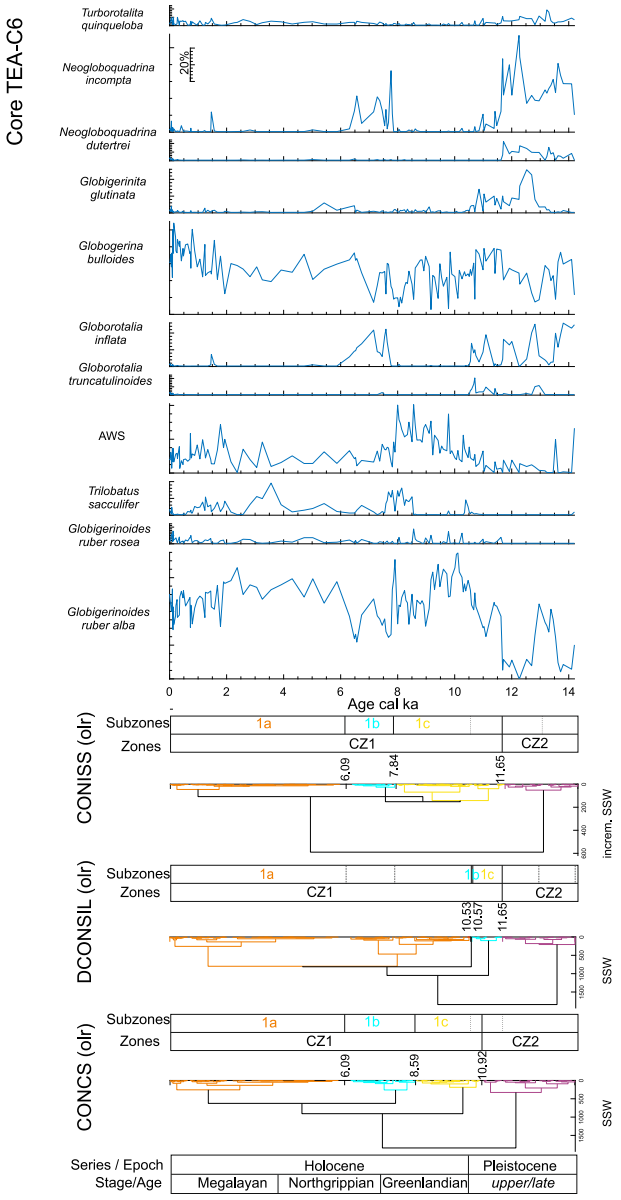

Fig. 4 Zonations provided by CONCS, DCONSIL and CONISS for the late Pleistocene to Holocene planktonic foraminiferal assemblages of the Core TEA-C6 (Ionian Sea). Numbers indicate the age in ka BP of the boundaries between zones. Percentage abundance of the taxa is shown on the top. AWS indicates an amalgamated warm water group

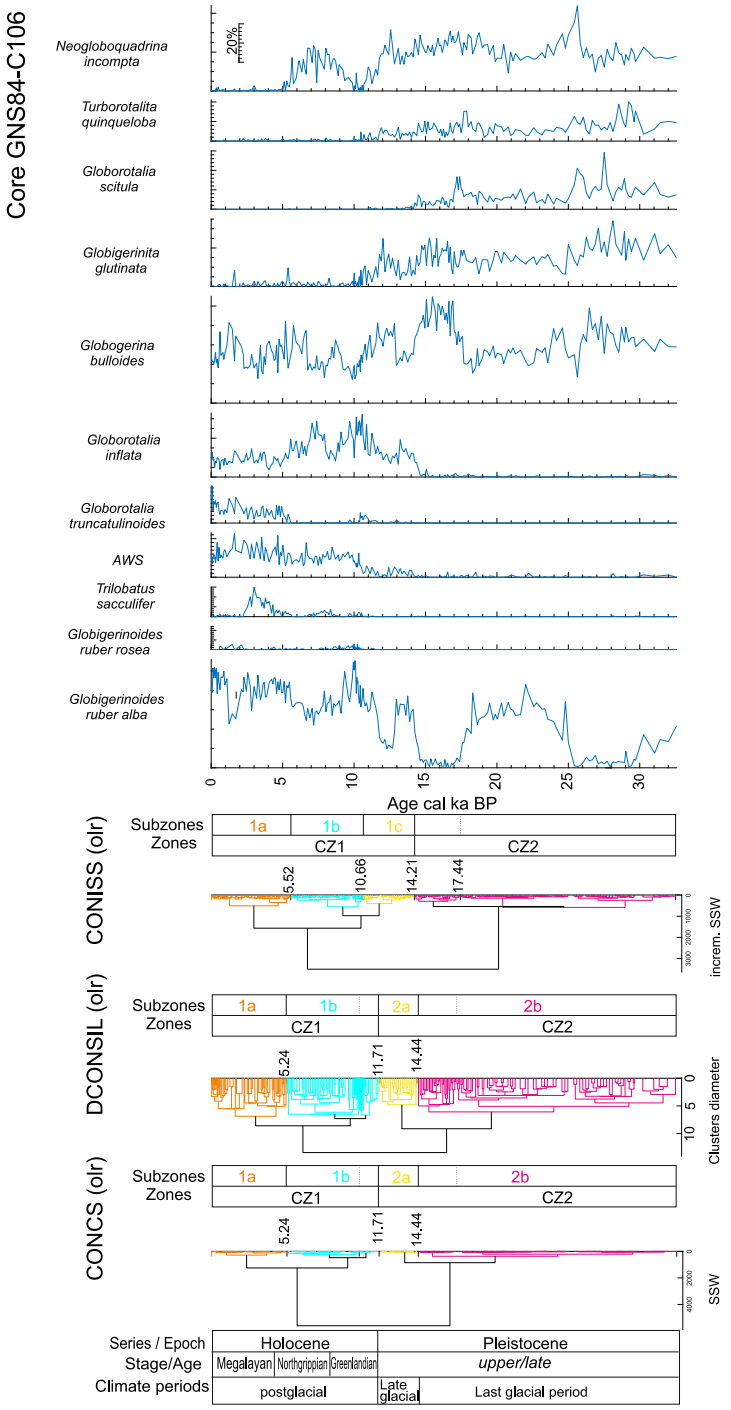

Fig. 5 Zonations provided by CONCS, DCONSIL and CONISS for the late Pleistocene to Holocene planktonic foraminiferal assemblages of the Core GNS84-C106 (Tyrrhenian Sea)

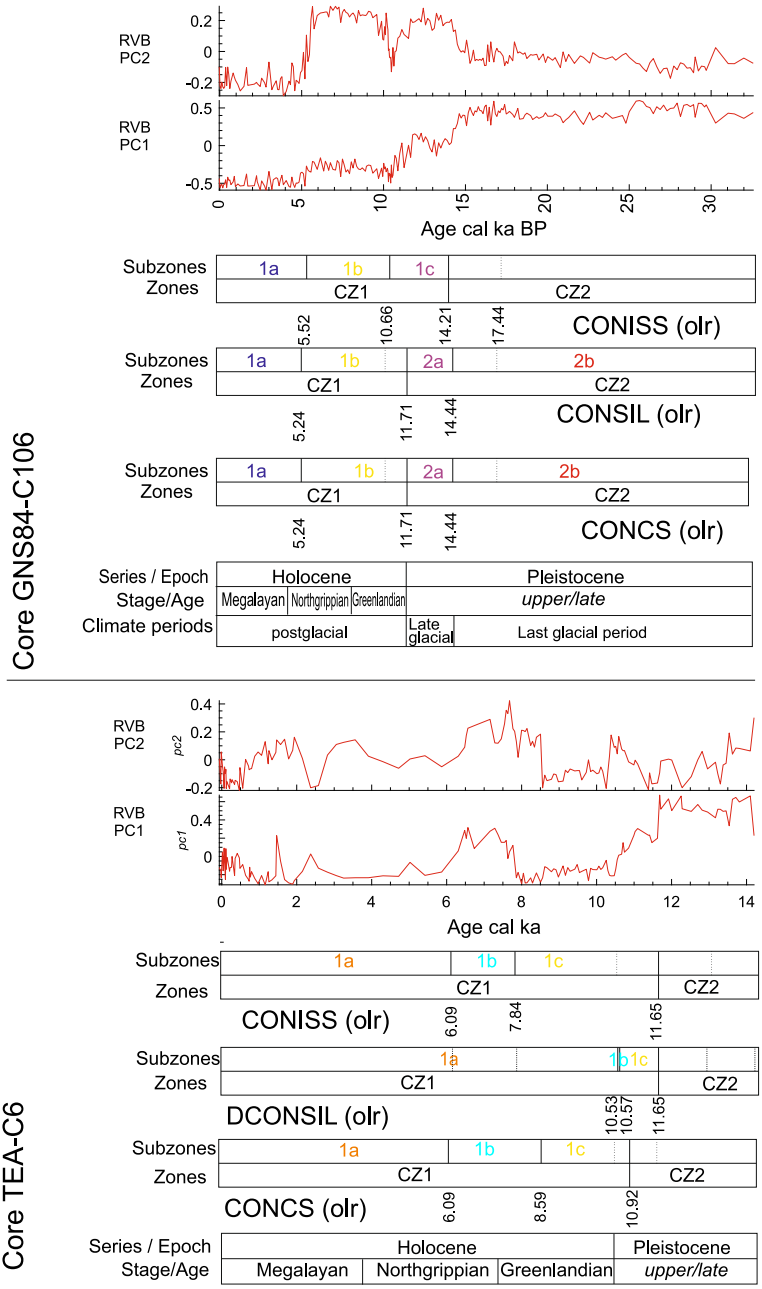

Fig. 6 Comparison between the CONCS, DCONSIL and CONISS zonation schemes and relative variation form biplots first two principal components scores of the planktonic foraminiferal assemblages of the Cores GNS84-C106 (Up) and TEA-C6 (Down)

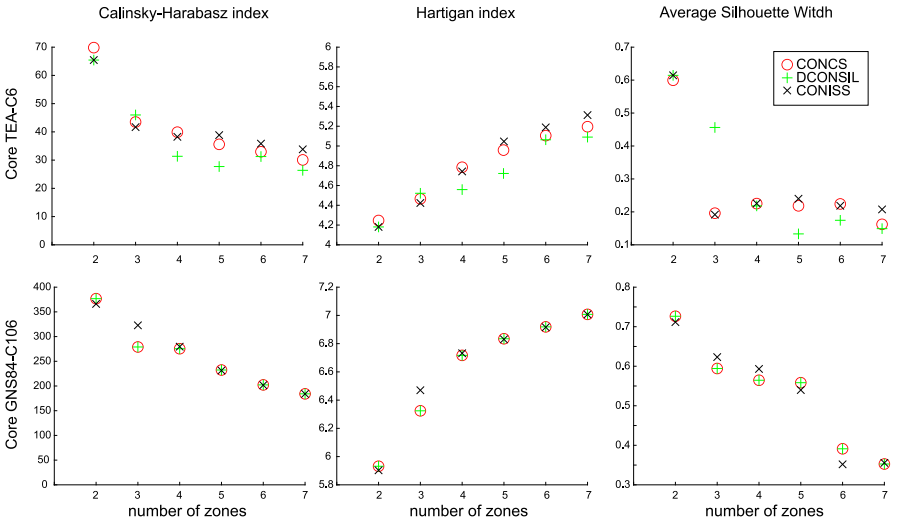

Fig. 7 Summary diagrams of the indices adopted to evaluate the performance of the algorithms for interval numbers varying from 2 to 7

Table 3 Goodman–Kruskal's coefficient (a), Cophenetic Correlation (b), and Elapsed time (c) values obtained for the three datasets with the clustering algorithms

Method	Artificial dataset	GNS84-C106	TEA-C6
(a) Goodman–Kruskal's coefficient			
CONCS	0.757	0.774	0.560
DCONSIL	0.755	0.775	0.630
CONISS	0.755	0.760	0.547
(b) Cophenetic Correlation			
CONCS	0.798	0.833	0.623
DCONSIL	0.797	0.834	0.770
CONISS	0.770	0.795	0.721
(c) Elapsed time (s)			
CONCS ver 1	0.669	0.922	0.867
CONCS ver 2	0.361	0.323	0.190
ACONSIL	115.158	76.106	12.282
DCONSIL	0.544	0.404	0.177
CONISS	0.265	0.274	0.204

Elapsed time (s) with MATLAB running with an Intel(R) Core(TM) i7-1065G7 CPU @ 1.50 GHz processor

5.3 Insight into Meaning and Interpretation of Zones

Figures 8 and 9 show the clr-biplots and GMBP computed on planktonic foraminiferal assemblages of cores Tea-C6 and GNS84-C106, with the division into the main zones and subzones considered. GMBP refers to the CONCS zonation. For Core GNS84-

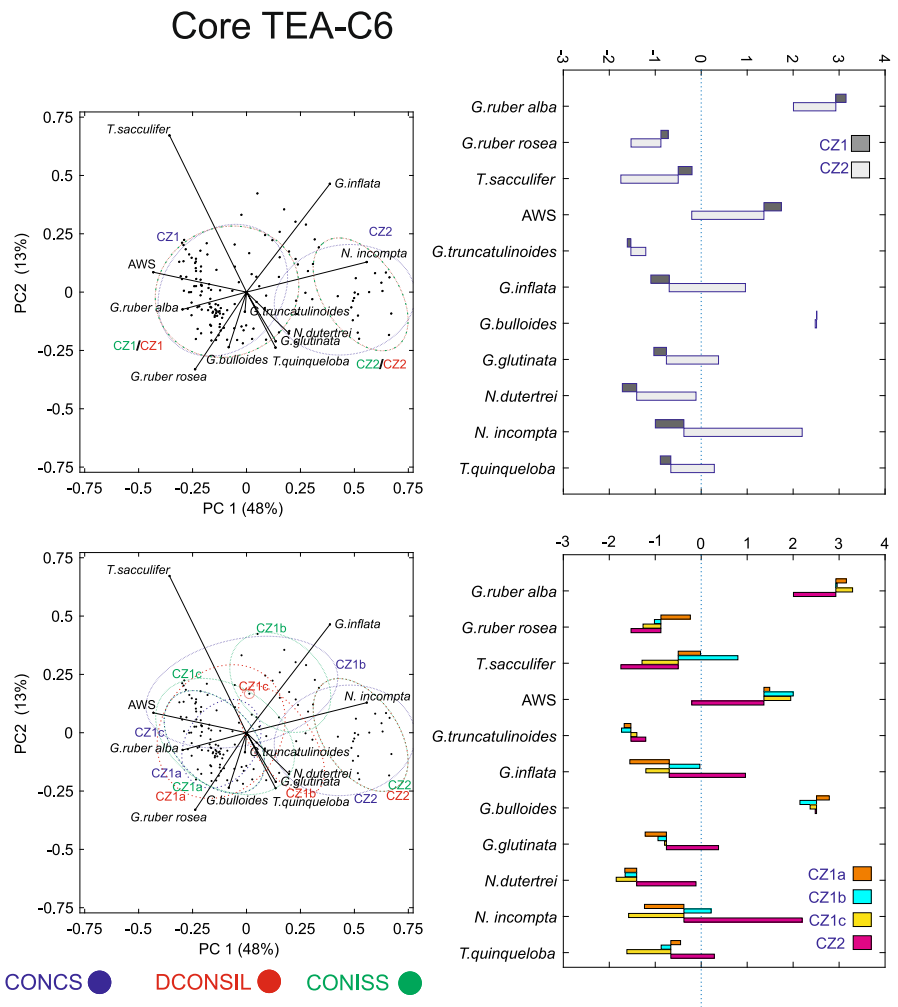

Fig. 8 Clr-biplots (left) and Geometric Mean Bar Plots (right) for Core TEA-C6. 2-sigma confidence ellipses are given for each zonation interval. The biplot at the top of the figure shows the division into two main zones, those at the bottom the division into four subzones

C106, the confidence ellipses of the subzones appear, though not completely, fairly well separated. It can also be noted that the spread from CZ1c to CZ1a subzones is oriented as the link between the rays of *N. incompta* and *G. truncatulinoides* (Fig. 8), related to their decreasing logratios recorded from the Northgrippian to the Megalayan interval of the core. With reference to the main division obtained for Core Tea-C6, the zones appear to be well separated, with a strong contrast between taxa more abundant in the Late Glacial and in the Holocene, respectively. It can also be noted that *G. bulloides* shows low variability across the intervals. With reference to the subzones, some effects, not unexpected for the constrained clustering, can be reported. In fact, a MANOVA performed on olr-coordinates of the four subzones considered, found a significant

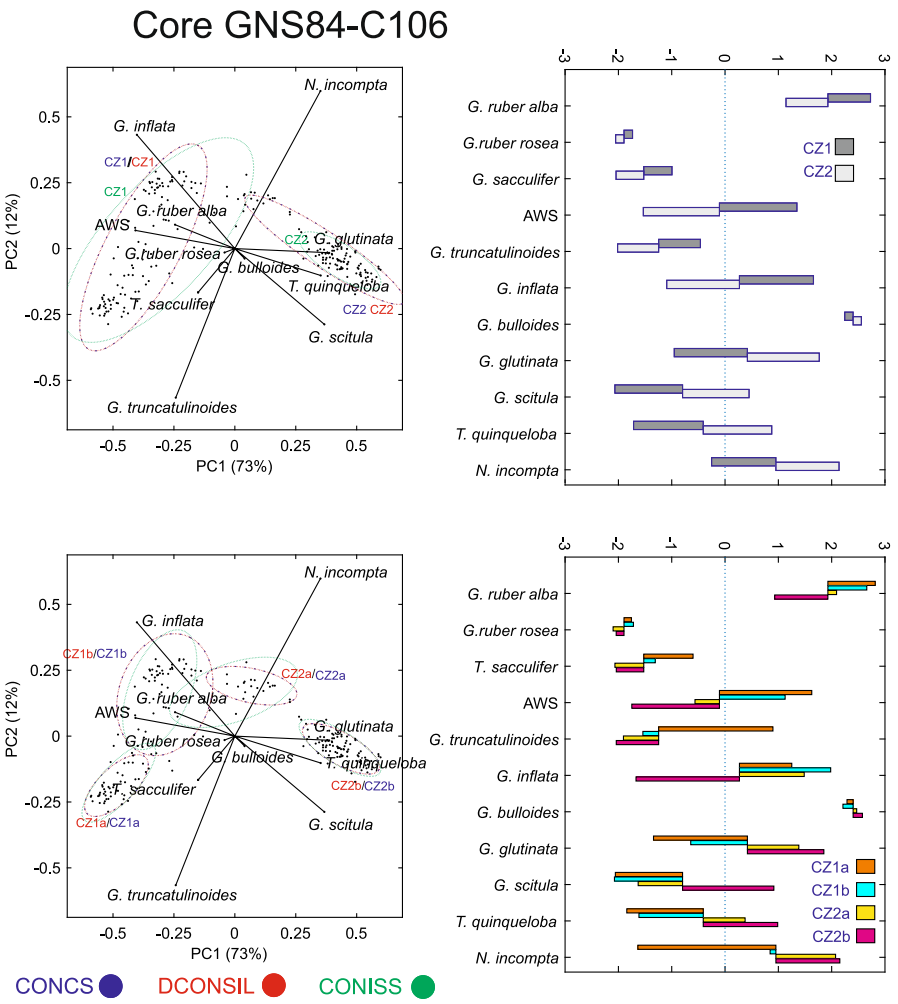

Fig. 9 Clr-biplots (left) and Geometric Mean Bar Plots (right) for Core GNS84-C106. 2-sigma confidence ellipses are given for each zonation interval. The biplots at the top of the figure show the division into two main zones, those at the bottom the division into four subzones

difference between mean vectors (p-value < 0.001). However, for Core TEA-C6, the confidence ellipses of subzones CONCS-CZ1a and CZ1c, overlap strongly, related to two intervals of the core, not immediately following each other, characterised by quite similar assemblages. The similarity between these subzones is also highlighted by the GMBP (Fig. 8). This case suggests considering pairwise comparisons between consecutive intervals as post hoc tests.

6 Concluding Remarks

Divisive algorithms are a viable alternative to hierarchical agglomerative algorithms for identifying intervals in a succession, especially in relation to the relatively reduced

number of subdivisions being considered in ecozonations. In particular, it seems logical to consider that while the clustering in the agglomerative process is less constrained at low levels and more constrained at high levels of aggregation, the partitioning of the divisive algorithms is less constrained at high levels, allowing for a better division, in relation to the criteria adopted, of the units of higher rank. More application examples, on real or artificial datasets, may help to deepen aspects of the behaviour of the divisive algorithms considered. However, in the application examples considered in this article, divisive algorithms provided valid results in terms of stability and in the evaluation of the indexes considered. As for comparing (and possibly, choosing) between the two divisive algorithms, it is up to the sensitivity of the analyst. Reasonably, CONCS privileges separation and homogeneity of groups, while DCONSIL should favour overall good classification of the observations. In the case of compositions, as in the case of planktonic foraminiferal assemblages, the framework on the CoDa methodology provides the tools for rigorous and appropriate application of algorithms in which concepts related to distance and between- and within-groups sum of squares measures are considered.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This research was supported by the Ministerio de Ciencia e Innovación under the projects “CODA-GENERA” (Ref. PID2021-123833OB-I00) and “CONBACO” (Ref. PID2021-125380OB-I00); and by the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project "COSDA" (Ref. 2021SGR01197).

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman and Hall Ltd, London. (Reprinted in 2003 by Blackburn Press, p 416)

Aitchison J, Greenacre M (2002) Biplots of compositional data. J R Stat Soc Ser C (Applied Stat) 51:375–392

Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32(3):271–275

Allen J, Brandt U, Brauer A, Hubberten HW, Huntley B, Keller J, Zolitschka B (1999) Rapid environmental changes in southern Europe during the last glacial period. Nature 400:740–743

Batool F, Hennig C (2021) Clustering with the average silhouette width. Comput Stat Data Anal 158:107190

Calinski T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3(1):1–27

Capotondi L, Borsetti A, Morigi C (1999) Foraminiferal ecozones, a high resolution proxy for the late quaternary biochronology in the central mediterranean sea. Mar Geol 153:253–274
Di Donato V, Esposito P, Russo-Ermolli E, Scarano A, Cheddadi R (2008) Coupled atmospheric and marine palaeoclimatic reconstruction for the last 35 kyr in the sele plain-gulf of salerno area (southern italy). Quatern Int 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, Insinga DD, 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. Glob Planet Change 172:278–297
Edwards A, Cavalli-Sforza LL (1965) A method for cluster analysis. Biometrics 21(2):362–375
Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Math Geol 37:795–828
Gill D (1970) Application of a statistical zonation method to reservoir evaluation and digitized-log analysis. AAPG Bull 54(5):719–729
Goodman L, Kruskal W (1954) Measures of association for cross-validations, part 1. J Am Stat Assoc 49:732–764
Gordon AD, Birks H (1972) Numerical methods in quaternary palaeoecology I. Zonation of pollen diagrams. New Phytol 71(5):961–979
Grimm E (1987) Coniss: a Fortran 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. Comput Geosci 13:13–35
Hartigan J (1975) Clustering algorithms. Wiley, New York
Hennig C, Meila M, Murtagh F, Rocci R (eds) (2015) Handbook of cluster analysis. Chapman and Hall/CRC, Boca Raton
Kaufman L, Rousseeuw P (1990) Finding groups in data: an introduction to cluster analysis. Wiley, Hoboken
Killick R, Fearnhead P, Eckley I (2012) Optimal detection of changepoints with a linear computational cost. J Am Stat Assoc 107(500):1590–1598
Martín-Fernández JA, Daunis-i Estadella J, Mateu-Figueras G (2015) On the interpretation of differences between groups for compositional data. SORT 39(2):231–252
Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosona-Delgado R (2018) Advances in principal balances for compositional data. Math Geosci 50(3):273–298
Martín-Fernández JA, Di Donato V, Pawlowsky-Glahn V, Egozcue JJ (2023) Insights in r-mode hierarchical clustering for compositional data. Math Geosci. https://doi.org/10.1007/s11004-023-10115-4
Mateu-Figueras G, Pawlowsky-Glahn V, Barceló-Vidal C (2003) Distributions on the simplex. In Thió-Henestrosa S, Martín-Fernández JA (eds) Proceedings of the 1st international workshop on compositional data analysis, Girona, Spain, p 17
Mojena R (1977) Hierarchical grouping methods and stopping rules: an evaluation. Comput J 20(4):359–363
Palarea-Albaladejo J, Martín-Fernández JA (2015) zcompositions - r package for multivariate imputation of nondetects and zeros in compositional data sets. Chemom Intell Lab Syst 143:85–96
Palarea-Albaladejo J, Martín-Fernández JA, Soto JA (2012) Dealing with distances and transformations for fuzzy c-means clustering of compositional data. J Classif 29(2):144–169
Pawlowsky-Glahn V, Egozcue JJ, Lovell D (2015) Tools for compositional data with a total. Stat Model 15(2):175–190
Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modelling and analysis of compositional data. Wiley, Chichester
Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Comput Appl Math 20:53–65
Roux M (2018) A comparative study of divisive and agglomerative hierarchical clustering algorithms. J Classif 35:345–366
Scott AJ, Symons MJ (1971) On the Edwards and Cavalli-Sforza method of cluster analysis. Biometrics 27(1):217–219
Siani G, Paterne M, Colin C (2010) Late glacial to holocene planktic foraminifera bioevents and climatic record in the south adriatic sea. J Quat Sci 25(5):808–821
Sprovieri R, Di Stefano E, Incarbona A, Gargano ME (2003) A high resolution record of the last deglaciation in the sicily channel based on foraminifera and calcareous nannofossil quantitative distribution. Palaeogeogr Palaeoclimatol Palaeoecol 202:119–142

Ward JHJ (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244