Zonation of the I-284 MIS 7–9 diatom record using R

1. Introduction

This document outlines the steps taken to split the Lake Ioannina MIS 7-9 diatom record into diatom assemblage zones, and it contains the R code used to do so. The reason for splitting the record into zones is to delineate sections of the record that share a similar diatom assemblage, helping to identify major changes in the record and making it easier to describe (the results can be written up by zone).

Creating the zones involves two steps:

1. cluster analysis
2. determination of the number of zones

Clustering algorithm choice

A *constrained* cluster analysis is required in order to group samples together that are stratigraphically adjacent. The specific cluster analysis chosen is the **Constrained Incremental Sum of Squares (CONISS)** method, as outlined by Grimm (1987).

First, a matrix of dissimilarities between samples is computed. The CONISS algorithm then computes a statistic known as the *sum of squares** between each pair of adjacent samples (each sample can be considered a cluster of just one sample at this stage). The pair with the smallest sum of squares is joined into a cluster, and then the sum of squares is recalculated for all samples with these newly joined samples receiving one sum of squares value for their cluster. The clusters with the smallest sum of squares is joined and the sum of squares recalculated. This process continues, clustering samples into successively larger groups (it is therefore an agglomerative technique).

*The sum of squares is the squared difference between the value of a taxon in one sample of a cluster divided by the average value of that taxon across all samples in that cluster, which is then summed for each taxon, which is then summed for each sample in the cluster.

2. Import packages and data

```
library(rioja) # for chclust function
library(ggdendro) # for dendro_data function
library(ggplot2)
library(patchwork)
source("scripts/02-manipulate.R") # import data
source("scripts/borders-for-ggplot2.R")
```

A note on importing the data

The data for the Ioannina MIS 7–9 analyses are stored in a single Excel workbook (ioannina.xlxs) in order to contain all the data in a single file and allow details of the data to be stored alongside them. The data are then explored through a series of R scripts. The code was split into these multiple scripts to make it easier to read and understand the code.

The first R script (01-load.R) reads in the data from the Excel workbook and assigns each worksheet to a dataframe in a list entitled **imported**. The second R script (02-manipulate.R) runs 01-load.R and subsequently manipulates the data into various forms (e.g. converts raw count data to percentage abundances). The data are placed in new dataframes within lists to keep similar data together.

The analyses in this document use the **counts** dataframe within the **taxa** list (i.e. `taxa$counts`). This simply contains the raw counts for each diatom taxon by sample number. The corresponding depths to the sample numbers are kept in another dataframe—**depths**, which is left unmanipulated within the **imported** list (i.e. `imported$depths`).

3. Prepare Data

The following steps are taken to prepare the data specifically for the analyses in this document:

1. Remove samples with no diatoms and an outlier, which contains some diatoms but is poorly preserved.

2. Discard rare taxa. These analyses are performed on taxa that are present at ≥4% in at least one sample. This decision was taken based on the following review of the literature.

   Gordon and Birks (1972) suggest that the cluster analysis should only be run on taxa present at ≥5% in at least one sample as the low abundance taxa are of little numerical importance. Grimm (1987) eliminated those present at <3% at every level and noted that eliminating rare taxa has little effect on the analysis. Bennett (1996) also uses only taxa present at >5% in at least one sample and then goes on to assess the effect of decreasing and increasing the threshold for taxa inclusion. In an example using the CONISS algorithm, Bennett (1996) identified 6 zones with the threshold set at 0-5%, 1%, 2% or 5%, and identified 5 zones with the threshold set at 10% or 20%—a difference of only one zone.

   Eliminating the rare taxa doesn't seem to make much difference, however Birks (1986) explains that although rare taxa could be of ecological importance, the counts of rare taxa are associated with a high relative error so are poorly estimated numerically unless very large counts are made.

   Obviously the work cited so far is quite old so I have looked around to see if there has been any progression on this. Birks (2012:357) states that the basic principles "remain largely unchanged" since they were established in Gordon and Birks (1972) and Birks and Gordon (1985).

3. Convert raw diatom valve counts into relative proportions (out of 1.0).

4. Square-root transform the data.

```r
coniss <- list()

coniss$data <- taxa$counts %>%

  # Replace sample numbers with depths.
  mutate(depth = imported$depths$depth, sample_no = NULL) %>%

  # Remove samples with no diatoms and outlier with poor preservation.
  filter(rowSums(select(., !matches("depth"))) > 0 & depth != 175.62) %>%

  # Isolate taxa present at ≥4%. Note that this uses a user-defined function
  # (abundant_taxa) that is part of the 02-manipulate.R script. It retrieves
  # the variable names of taxa that are present at least at the specified
  # abundance (here 4%) in at least one sample.
  column_to_rownames("depth") %>%
  select(all_of(abundant_taxa(taxa$rel_ab, 4))) %>%

  # Convert raw counts to proportion based on only these abundant taxa.
  decostand(method = "total", na.rm = "TRUE") %>%

  # Square-root transform data.
  sqrt()
```

4. Calculate distance matrix

The squared Euclidean distance between samples is calculated. This dissimilarity index is the one used by the Tilia program.

```
coniss$dist_matrix <- designdist(coniss$data,
                                 method = "A+B-2*J",
                                 terms = "quadratic")
```

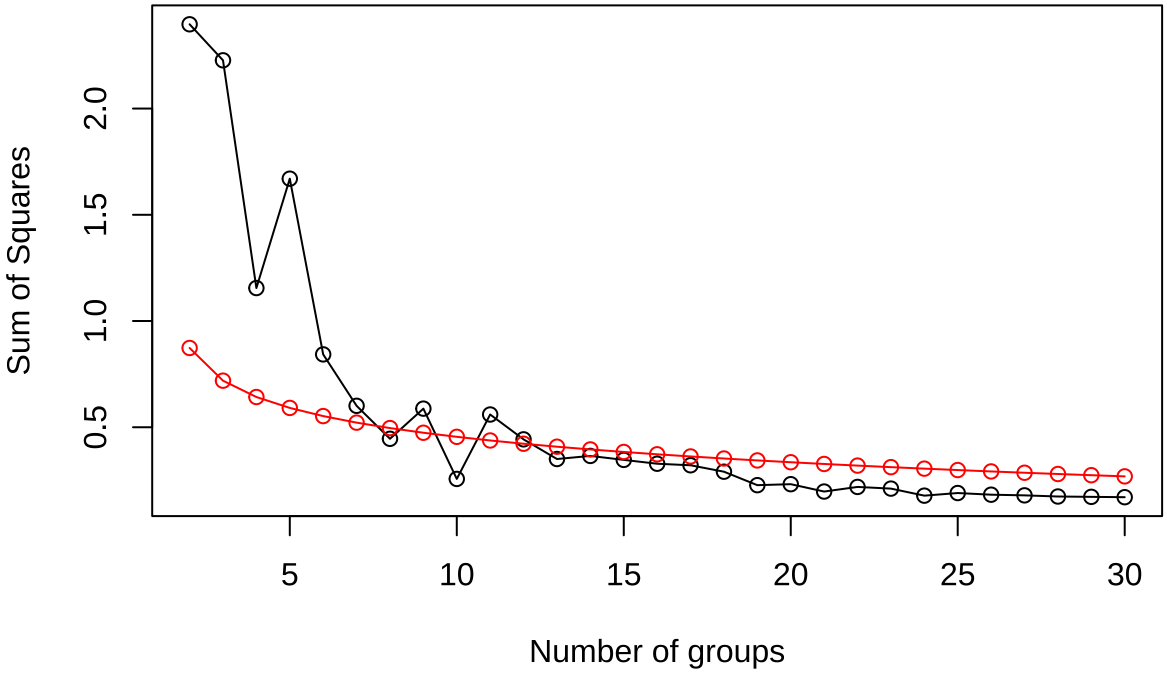
5. Cluster analysis

```
coniss$results <- chclust(coniss$dist_matrix, method = "coniss")
```

6. Determination of the number of zones

A broken stick model (Bennett, 1996) is used to determine the number of significant zones in the record. The plot below shows that after splitting the record into 12 groups (zones), the observed reduction in within-group sum of squares (black line) drops (and remains) below that expected from the broken stick model (red line). Therefore, there are 12 significant zones in this record.

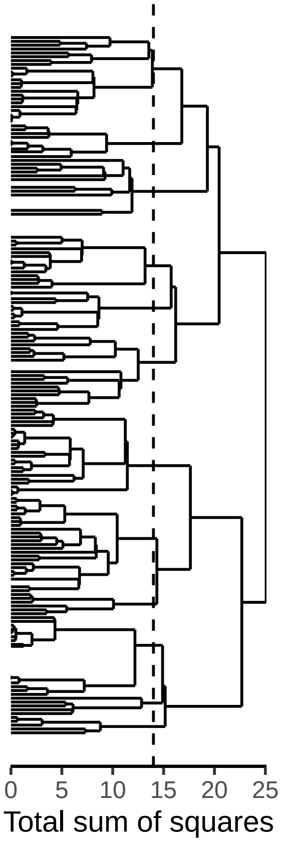
```
bstick(coniss$results, ng = 30)
```



7. Plot dendrogram

```
# Extract data for plotting.
coniss$ddata<- dendro_data(coniss$results, type = "rectangle")
# Modify x values so leaves are plotted by depth in core rather than in
# sequential order.
new_x <- approxfun(coniss$ddata$labels$x,
                   as.numeric(as.character(coniss$ddata$labels$label)))
coniss$ddata$segments$x <- new_x(coniss$ddata$segments$x)
coniss$ddata$segments$xend <- new_x(coniss$ddata$segments$xend)

# Create plot.
ggplot(segment(coniss$ddata)) +
  geom_segment(aes(x = x, y = y, xend = xend, yend = yend)) +
  coord_flip() +
  scale_y_continuous(expand = c(0, 0)) +
  scale_x_reverse(breaks = NULL,
                  labels = NULL) +
  labs(x = "",
       y = "Total sum of squares") +
  theme_bw() +
  theme(aspect.ratio = 3,
        legend.position = "none",
        panel.grid = element_blank(),
        panel.border = theme_border("bottom"),
        panel.background = element_rect(fill = "transparent"),
        plot.background = element_rect(fill = "transparent")) +
  geom_hline(yintercept = 14,
               linetype = 2)
```



References

Bennett, K. D. (1996) Determination of the number of zones in a biostratigraphical sequence. *New Phytologist*, 132, 155–170.

Birks, H. J. B. (1986) Numerical zonation, comparison and correlation of Quaternary pollen-stratigraphical data. In Berglund, B. E. (ed) *Handbook of Palaeoecology and Palaeohydrology*. Chichester: John Wiley & Sons.

Birks, H. J. B. (2012) Analysis of stratigraphical data. In Birks, H. J. B., Lotter, A. F., Juggins, S. & Smol, J. P. (eds) *Tracking Environmental Change using Lake Sediments. Volume 5: Data Handling and Numerical Techniques*. Dordrecht: Springer.

Birks, H. J. B. & Gordon, A. D. (1985) *Numerical methods in Quaternary pollen analysis*. London: Academic Press.

Gordon, A. D. & Birks, H. J. B. (1972) Numerical methods in Quaternary palaeoecology. I. Zonation of pollen diagrams. *New Phytologist*, 71, 961–979.

Grimm, E. C. (1987) CONISS: A Fortran 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. *Computers & Geosciences*, 13, 13–35.

Grimm, E. C. (2011) *Tilia version 1.7.16* [Software]. Illinois State Museum, Springfield. Available online: https://www.tiliait.com.