Diatom Community Shifts Across a Late-Pleistocene to Holocene Climate Transition

Aryan Bhojani

Table of contents

0.1 Project overview

This project explores how diatom community composition changes through time in a Gulf of California sediment core, and how those changes relate to a major climate transition spanning the end of the Pleistocene and the start of the Holocene. I treat each depth sample as a time-indexed snapshot of phytoplankton community structure.

Main goals: 1. Clean and standardize diatom count data into comparable relative abundances. 2. Visualize how taxa vary through time and across climate epochs. 3. Relate diatom abundance patterns to reconstructed sea surface temperature (SST). 4. Summarize multivariate community structure using PCA and compare it to a supervised separation direction from logistic regression.

0.2 Setup

```
rm(list = ls())

library(tidyverse)
library(patchwork)
library(ggridges)
library(corrplot)
library(colorspace)

library(hexbin)
library(RColorBrewer)

set.seed(1)
```

0.3 Data description

Diatom counts were recorded at evenly spaced depths in a sediment core. Depth corresponds to time before present (older at greater depth). The dataset contains:

- Depth (cm)
- Age (thousands of years before present, KyrBP)
- Multiple diatom taxa counts
- Num.counted: total phytoplankton counted in that sample (not only diatoms)

I convert counts to relative abundances by dividing each taxon count by Num.counted, so values are comparable across samples.

0.4 Load data

```
diatoms_raw <- read_csv("data/barron-diatoms.csv")
glimpse(diatoms_raw)
```

```
Rows: 230
Columns: 11
$ Depth       <dbl> 0.00, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45~
$ Age         <dbl> 1.33, 1.37, 1.42, 1.46, 1.51, 1.55, 1.59, 1.64, 1.68, 1.72~
$ A_curv      <dbl> 5, 8, 8, 11, 11, 4, 5, 7, 14, 6, 8, 5, 9, 5, 13, 6, 9, 14,~
$ A_octon     <dbl> 2, 2, 6, 1, 1, 9, 3, 4, 7, 3, 4, 3, 3, 4, 2, 2, 2, 3, 5, 4~
$ ActinSpp    <dbl> 32, 31, 33, 21, 38, 30, 21, 36, 40, 37, 20, 35, 30, 27, 36~
$ A_nodul     <dbl> 14, 16, 18, 1, 3, 10, 4, 19, 4, 6, 11, 3, 3, 6, 8, 4, 3, 1~
$ CoscinSpp   <dbl> 21, 20, 29, 12, 18, 16, 12, 19, 24, 26, 11, 15, 18, 19, 15~
$ CyclotSpp   <dbl> 22, 16, 7, 28, 24, 14, 16, 13, 11, 20, 20, 22, 27, 35, 21,~
$ Rop_tess    <dbl> 1, 7, 1, 25, 3, 16, 40, 3, 3, 3, 27, 12, 7, 4, 6, 1, 4, 4,~
$ StephanSpp  <dbl> 1, 2, 1, 3, NA, NA, NA, 1, 1, 1, 1, 3, 1, NA, NA, NA, 1, N~
$ Num.counted <dbl> 201, 200, 200, 200, 300, 203, 200, 200, 200, 200, 200, 201~
```

0.5 Data quality: missing values

In this dataset, missing values represent taxa that were not observed in a sample (recorded as NA). I quantify missingness by taxon, then replace NA with 0.

```r
missing_frac <- diatoms_raw |>
summarize(
across(
A_curv:StephanSpp,
~ mean(is.na(.x)),
.names = "{.col}_missing_frac"
)
)

missing_frac
```

```
# A tibble: 1 x 8
  A_curv_missing_frac A_octon_missing_frac ActinSpp_missing_frac
                <dbl>                <dbl>                 <dbl>
1              0.0391               0.0435                     0
# i 5 more variables: A_nodul_missing_frac <dbl>, CoscinSpp_missing_frac <dbl>,
#   CyclotSpp_missing_frac <dbl>, Rop_tess_missing_frac <dbl>,
#   StephanSpp_missing_frac <dbl>
```

0.5.1 Replace NA with 0 for taxa counts

```r
diatoms1 <- diatoms_raw |>
mutate(across(A_curv:StephanSpp, ~ replace_na(.x, 0)))

diatoms_mod1_examplerows <- diatoms1 |>
slice(94, 95)

diatoms_mod1_examplerows
```

```
# A tibble: 2 x 11
  Depth   Age A_curv A_octon ActinSpp A_nodul CoscinSpp CyclotSpp Rop_tess
  <dbl> <dbl>  <dbl>   <dbl>    <dbl>   <dbl>     <dbl>     <dbl>    <dbl>
1  5.6   6.24      9      13       17       0        22        14       25
2  5.65  6.29      8       6       19       0        15        20       32
# i 2 more variables: StephanSpp <dbl>, Num.counted <dbl>
```

0.6 Transform counts to relative abundances

Two steps:

1. Convert diatom counts to proportions by dividing by Num.counted.
2. Flip Age to negative so that "older" is more negative (useful for plotting timelines left-to-right toward the present).

```
diatoms2 <- diatoms1 |>
mutate(
Age = -1 * Age,
across(A_curv:StephanSpp, ~ .x / Num.counted)
)

diatoms2 |> head()
```

```
# A tibble: 6 x 11
  Depth   Age A_curv A_octon ActinSpp A_nodul CoscinSpp CyclotSpp Rop_tess
  <dbl> <dbl>  <dbl>   <dbl>    <dbl>   <dbl>     <dbl>     <dbl>    <dbl>
1  0    -1.33 0.0249 0.00995    0.159  0.0697    0.104     0.109   0.00498
2  0.05 -1.37 0.04   0.01       0.155  0.08      0.1       0.08    0.035
3  0.1  -1.42 0.04   0.03       0.165  0.09      0.145     0.035   0.005
4  0.15 -1.46 0.055  0.005      0.105  0.005     0.06      0.14    0.125
5  0.2  -1.51 0.0367 0.00333    0.127  0.01      0.06      0.08    0.01
6  0.25 -1.55 0.0197 0.0443     0.148  0.0493    0.0788    0.0690  0.0788
# i 2 more variables: StephanSpp <dbl>, Num.counted <dbl>
```

0.7 Time coverage and sampling resolution

0.7.1 Time span (in years)

```
oldest_years_bp <- max(diatoms_raw$Age, na.rm = TRUE) * 1000
newest_years_bp <- min(diatoms_raw$Age, na.rm = TRUE) * 1000

tibble(
newest_years_bp = newest_years_bp,
oldest_years_bp = oldest_years_bp,
span_years = oldest_years_bp - newest_years_bp
)
```

```
# A tibble: 1 x 3
  newest_years_bp oldest_years_bp span_years
            <dbl>           <dbl>      <dbl>
1            1330           15190      13860
```

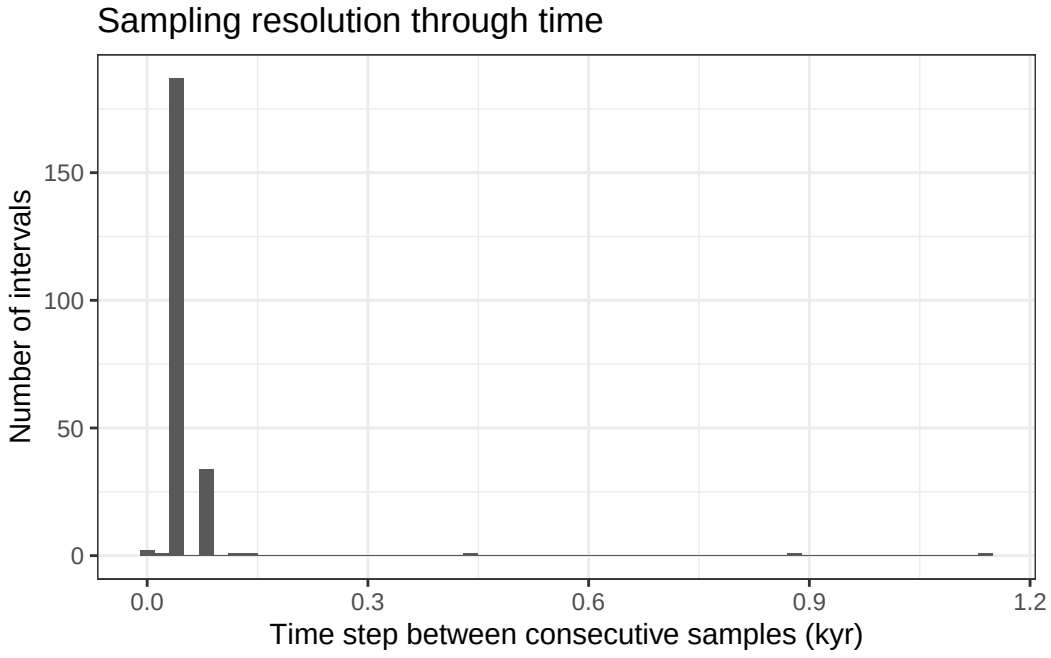
0.7.2 Sampling gaps

I compute time steps between consecutive samples and visualize the gap distribution.

```
diatoms3 <- diatoms2 |>
arrange(Age) |>
mutate(Gap = c(NA, diff(Age)))

gap_histogram <- diatoms3 |>
ggplot(aes(x = Gap)) +
geom_histogram(binwidth = 0.02) +
labs(
x = "Time step between consecutive samples (kyr)",
y = "Number of intervals",
title = "Sampling resolution through time"
) +
theme_bw()

gap_histogram
```



0.8 Univariate summaries: which taxa dominate?

I summarize mean and standard deviation of each taxon's relative abundance.

```
diatom_summary <- diatoms3 |>
summarize(
across(
A_curv:StephanSpp,
list(
mean = ~ mean(.x, na.rm = TRUE),
sd   = ~ sd(.x,   na.rm = TRUE)
),
.names = "{.col}_{.fn}"
)
)

diatom_summary
```

```
# A tibble: 1 x 16
  A_curv_mean A_curv_sd A_octon_mean A_octon_sd ActinSpp_mean ActinSpp_sd
        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <dbl>
1      0.0290    0.0186       0.0183     0.0165         0.136      0.0538
# i 10 more variables: A_nodul_mean <dbl>, A_nodul_sd <dbl>,
#   CoscinSpp_mean <dbl>, CoscinSpp_sd <dbl>, CyclotSpp_mean <dbl>,
#   CyclotSpp_sd <dbl>, Rop_tess_mean <dbl>, Rop_tess_sd <dbl>,
#   StephanSpp_mean <dbl>, StephanSpp_sd <dbl>
```

0.8.1 Tidy comparison table (mean vs sd)

```
diatom_summary2 <- diatom_summary |>
pivot_longer(
cols = everything(),
names_pattern = "(.+)_(mean|sd)$",
names_to = c("taxon", "statistic"),
values_to = "value"
) |>
pivot_wider(names_from = statistic, values_from = value) |>
arrange(desc(mean))

diatom_summary2
```

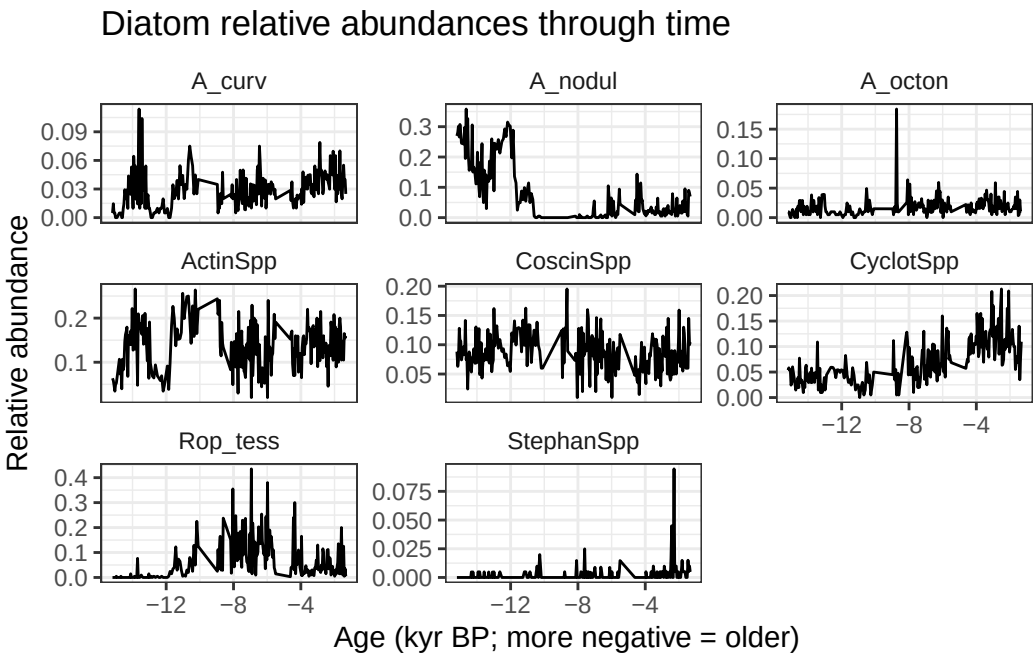
```
# A tibble: 8 x 3
  taxon         mean      sd
  <chr>        <dbl>   <dbl>
1 ActinSpp   0.136   0.0538
2 CoscinSpp  0.0859  0.0318
3 A_nodul    0.0729  0.0927
4 CyclotSpp  0.0704  0.0424
5 Rop_tess   0.0604  0.0761
6 A_curv     0.0290  0.0186
7 A_octon    0.0183  0.0165
8 StephanSpp 0.00245 0.00772
```

0.9 Time series: relative abundance trajectories

I visualize each taxon's relative abundance through time.

```
ts_plot <- diatoms3 |>
pivot_longer(
cols = A_curv:StephanSpp,
names_to = "taxon",
values_to = "rel_abund"
) |>
ggplot(aes(x = Age, y = rel_abund)) +
geom_line(linewidth = 0.5) +
facet_wrap(~ taxon, scales = "free_y") +
labs(
x = "Age (kyr BP; more negative = older)",
y = "Relative abundance",
title = "Diatom relative abundances through time"
) +
theme_bw() +
theme(strip.background = element_blank())

ts_plot
```



0.10 Climate context: SST and key intervals

I use SST reconstructions to contextualize biological change across two well-known intervals:

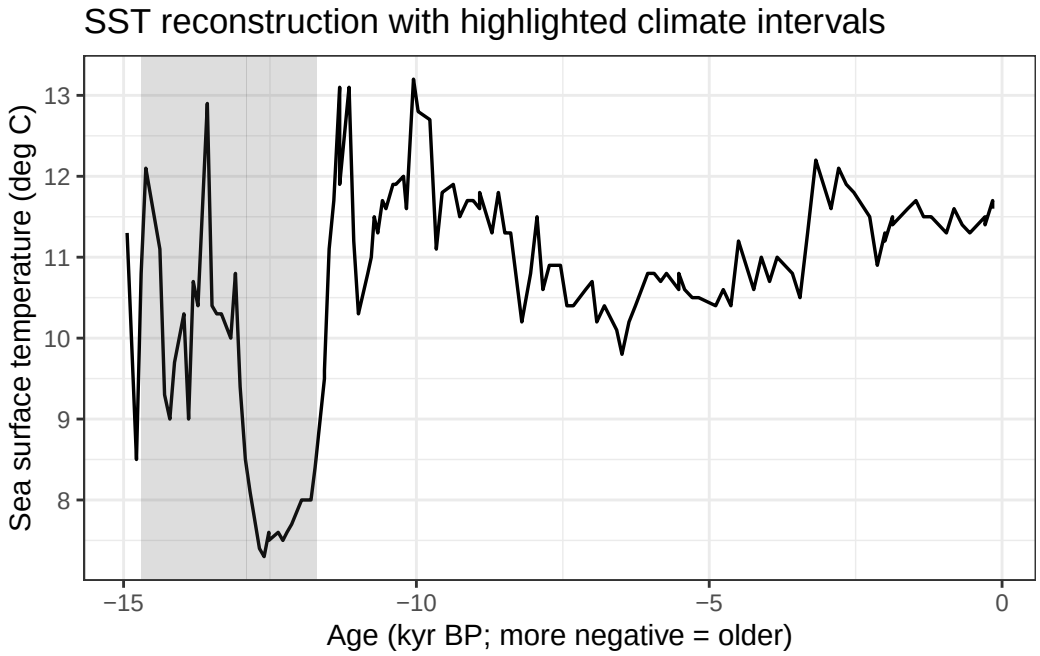
- Late Glacial Interstadial: 14.7 to 12.9 kyr BP
- Younger Dryas: 12.9 to 11.7 kyr BP

```
seatemps <- read_csv("data/barron-sst.csv") |>
mutate(Age = -1 * Age)

sst_plot <- seatemps |>
ggplot(aes(x = Age, y = SST)) +
geom_line(linewidth = 0.6) +
annotate(
"rect",
xmin = -14.7, xmax = -12.9,
ymin = -Inf, ymax = Inf,
alpha = 0.2
) +
annotate(
"rect",
xmin = -12.9, xmax = -11.7,
```

```
ymin = -Inf, ymax = Inf,
alpha = 0.2
) +
labs(
x = "Age (kyr BP; more negative = older)",
y = "Sea surface temperature (deg C)",
title = "SST reconstruction with highlighted climate intervals"
) +
theme_bw()

sst_plot
```



0.11 Epoch comparison: before vs after 11.7 kyr BP

I split samples into two epochs:

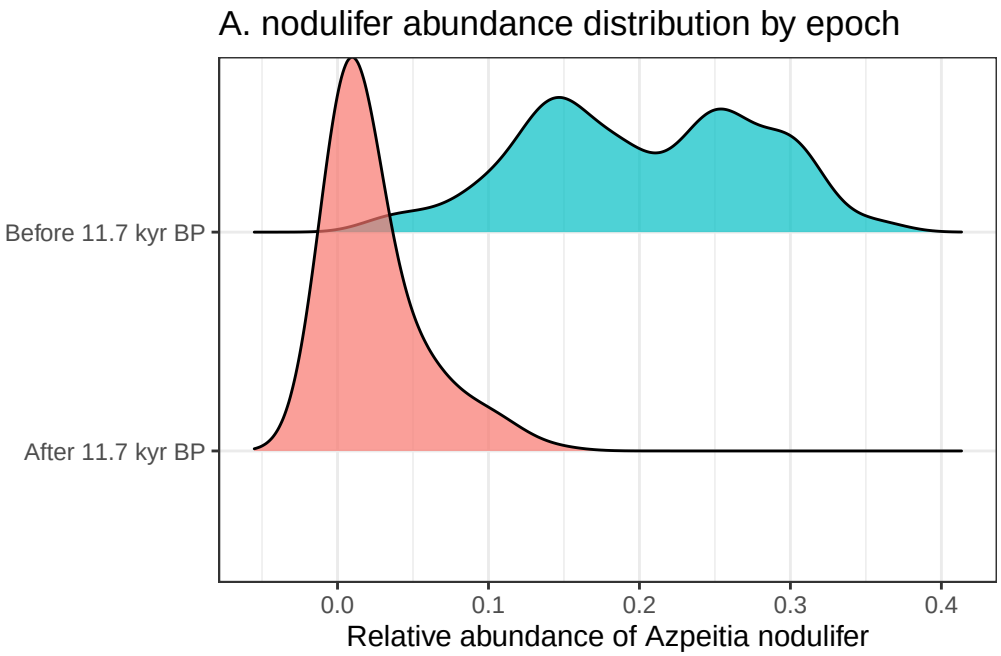
- Before or equal to 11.7 kyr BP (older; late Pleistocene / transition)
- After 11.7 kyr BP (younger; Holocene)

Then I compare the distribution of Azpeitia nodulifer relative abundance.

```
diatoms4 <- diatoms3 |>
mutate(
epoch = if_else(
Age <= -11.7,
"Before 11.7 kyr BP",
"After 11.7 kyr BP"
)
)

ridge_plot <- diatoms4 |>
ggplot(aes(x = A_nodul, y = epoch, fill = epoch)) +
geom_density_ridges(alpha = 0.7) +
labs(
x = "Relative abundance of Azpeitia nodulifer",
y = "",
title = "A. nodulifer abundance distribution by epoch"
) +
theme_bw() +
theme(legend.position = "none")

ridge_plot
```



0.12 Linking diatoms to SST via LOESS interpolation

SST is not observed at exactly the same ages as the diatom samples. I fit a LOESS curve and predict SST at diatom ages, then visualize SST vs abundance relationships.

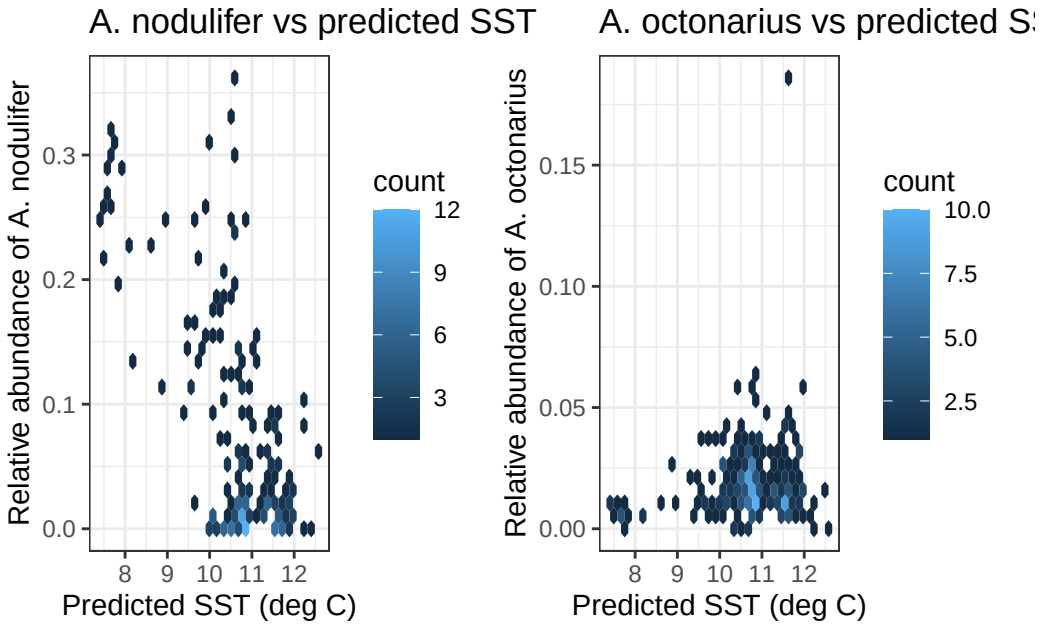
```
loess_fit <- loess(SST ~ Age, data = seatemps, span = 0.1)

diatoms3 <- diatoms3 |>
mutate(sst_pred = predict(loess_fit, newdata = Age))

hex1 <- diatoms3 |>
ggplot(aes(x = sst_pred, y = A_nodul)) +
geom_hex() +
labs(
x = "Predicted SST (deg C)",
y = "Relative abundance of A. nodulifer",
title = "A. nodulifer vs predicted SST"
) +
theme_bw()

hex2 <- diatoms3 |>
ggplot(aes(x = sst_pred, y = A_octon)) +
geom_hex() +
labs(
x = "Predicted SST (deg C)",
y = "Relative abundance of A. octonarius",
title = "A. octonarius vs predicted SST"
) +
theme_bw()

hex1 + hex2
```



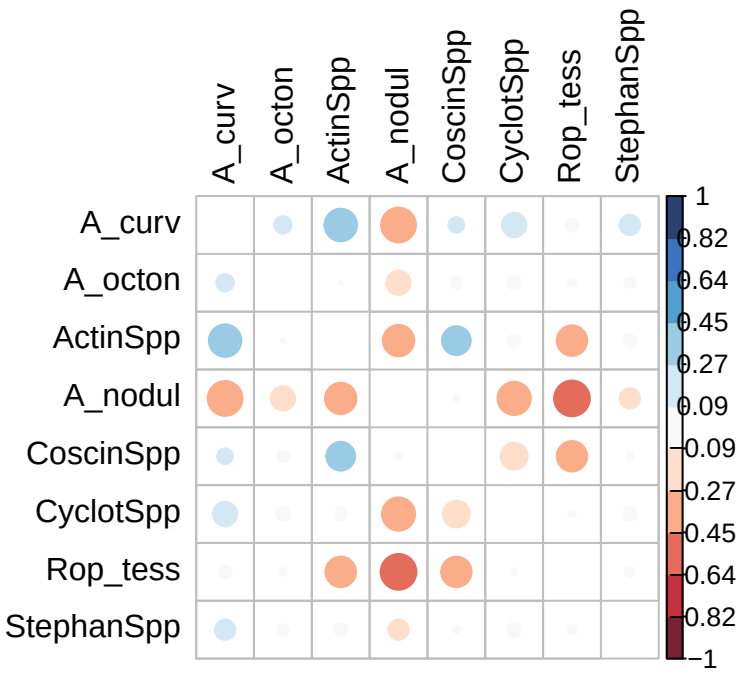
0.13 Community-level structure

0.13.1 Correlation structure across taxa

```
corr_mx <- diatoms3 |>
select(A_curv:StephanSpp) |>
cor(use = "pairwise.complete.obs")

base_cols <- brewer.pal(11, "RdBu")
plot_cols <- lighten(base_cols, amount = 0.1)

corrplot(
corr_mx,
method = "circle",
type   = "full",
diag   = FALSE,
col    = plot_cols,
tl.col = "black"
)
```



0.13.2 Standardize taxa for multivariate models

```
diatoms5 <- diatoms4 |>
mutate(
across(A_curv:StephanSpp, ~ as.numeric(scale(.x))),
pleistocene = if_else(epoch == "Before 11.7 kyr BP", 1L, 0L)
)

scale_check <- diatoms5 |>
summarize(
across(
A_curv:StephanSpp,
list(
mean = ~ mean(.x, na.rm = TRUE),
sd   = ~ sd(.x,   na.rm = TRUE)
)
)
)

scale_check
```

```
# A tibble: 1 x 16
  A_curv_mean A_curv_sd A_octon_mean A_octon_sd ActinSpp_mean ActinSpp_sd
        <dbl>     <dbl>        <dbl>      <dbl>         <dbl>       <dbl>
1   -5.20e-16         1     7.00e-16          1      2.05e-15           1
# i 10 more variables: A_nodul_mean <dbl>, A_nodul_sd <dbl>,
#   CoscinSpp_mean <dbl>, CoscinSpp_sd <dbl>, CyclotSpp_mean <dbl>,
#   CyclotSpp_sd <dbl>, Rop_tess_mean <dbl>, Rop_tess_sd <dbl>,
#   StephanSpp_mean <dbl>, StephanSpp_sd <dbl>
```

Interpretation note: means are extremely close to 0 and SDs extremely close to 1. Tiny deviations are expected from floating-point arithmetic.

0.14 Logistic regression: supervised separation of epochs

I fit a logistic regression where the response is epoch membership and predictors are standardized taxa.

```
diatoms_glm <- diatoms5 |>
select(pleistocene, A_curv:StephanSpp)

logistic_result <- glm(
pleistocene ~ .,
data = diatoms_glm,
family = binomial
)

summary(logistic_result)
```

```
Call:
glm(formula = pleistocene ~ ., family = binomial, data = diatoms_glm)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2118.21   55066.43  -0.038    0.969
A_curv         73.21    2122.08   0.034    0.972
A_octon      -294.38    7781.45  -0.038    0.970
ActinSpp     -518.94   13669.55  -0.038    0.970
A_nodul       856.31   22011.47   0.039    0.969
CoscinSpp    -204.64    5371.27  -0.038    0.970
CyclotSpp    -245.70    7222.34  -0.034    0.973
```

```
Rop_tess     -3414.94   88653.84  -0.039    0.969
StephanSpp     243.96    6352.19   0.038    0.969

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2.7007e+02  on 229  degrees of freedom
Residual deviance: 5.2656e-06  on 221  degrees of freedom
AIC: 18

Number of Fisher Scoring iterations: 25
```

```
lo_fit <- fitted(logistic_result)

tibble(
  fitted_min = min(lo_fit),
  fitted_q25 = quantile(lo_fit, 0.25),
  fitted_med = median(lo_fit),
  fitted_q75 = quantile(lo_fit, 0.75),
  fitted_max = max(lo_fit)
)
```

```
# A tibble: 1 x 5
  fitted_min fitted_q25 fitted_med fitted_q75 fitted_max
       <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
1   2.22e-16   2.22e-16   2.22e-16          1          1
```

```
pred_class <- as.integer(lo_fit >= 0.5)

confusion <- table(
  truth = diatoms_glm$pleistocene,
  pred  = pred_class
)

confusion
```

```
     pred
truth   0   1
    0 167   0
    1   0  63
```

```r
mean(pred_class == diatoms_glm$pleistocene)
```

```
[1] 1
```

0.14.1 Extract the normalized regression direction

```r
reg_direction <- coef(logistic_result)[-1]
reg_direction <- reg_direction / sqrt(sum(reg_direction^2))

length(reg_direction)
```

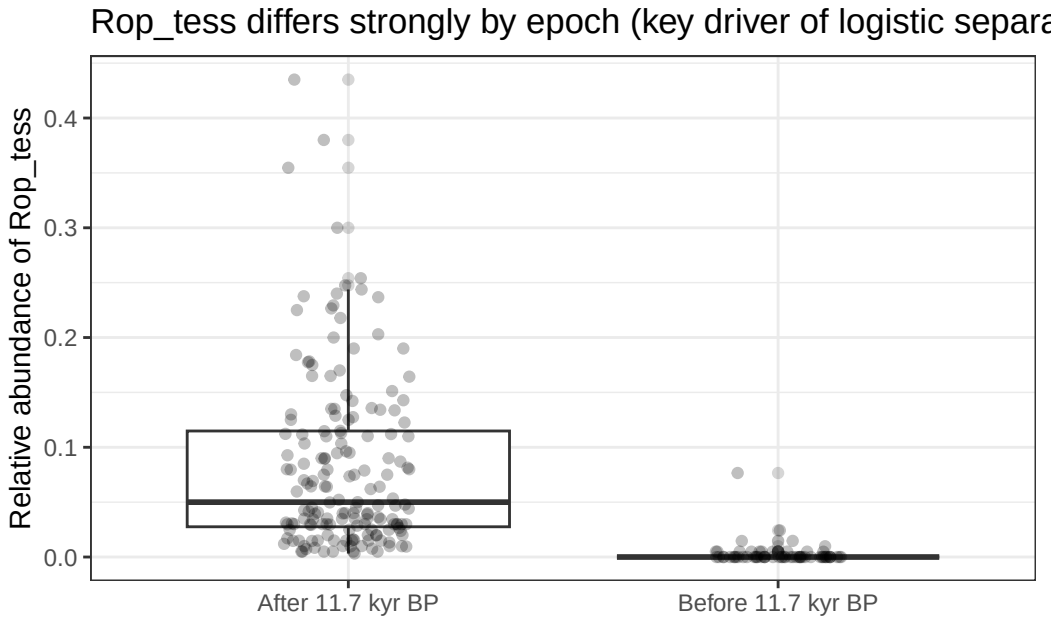
```
[1] 8
```

```r
reg_direction
```

```
     A_curv     A_octon    ActinSpp     A_nodul   CoscinSpp   CyclotSpp
 0.02036786 -0.08190556 -0.14438346  0.23824985 -0.05693539 -0.06835936
   Rop_tess  StephanSpp
-0.95012904  0.06787650
```

```r
rop_epoch_plot <- diatoms4 |>
  ggplot(aes(x = epoch, y = Rop_tess)) +
  geom_boxplot(outlier.alpha = 0.2) +
  geom_jitter(width = 0.15, alpha = 0.25) +
  labs(
    x = "",
    y = "Relative abundance of Rop_tess",
    title = "Rop_tess differs strongly by epoch (key driver of logistic separation)"
  ) +
  theme_bw() +
  theme(legend.position = "none")

rop_epoch_plot
```



0.15 PCA: unsupervised community variation

I run PCA on standardized taxa (already scaled, so I do not re-center or re-scale inside prcomp).

```
diatoms_pca <- diatoms5 |>
select(A_curv:StephanSpp)

pca_result <- prcomp(diatoms_pca, center = FALSE, scale. = FALSE)

pc1_direction <- pca_result$rotation[, 1]
pc2_direction <- pca_result$rotation[, 2]

sqrt(sum(pc1_direction^2))
```

```
[1] 1
```

```
sqrt(sum(pc2_direction^2))
```

```
[1] 1
```

```
pc1_direction
```

```
     A_curv     A_octon    ActinSpp     A_nodul   CoscinSpp   CyclotSpp
 0.52137754  0.19451989  0.37381456 -0.61156303  0.04119872  0.34572601
   Rop_tess  StephanSpp
 0.11678604  0.20424991
```

0.16 Compare PCA and logistic directions

I compare absolute loadings from:

- logistic regression direction (epoch-separating)
- PC1 direction (dominant overall variance)

```
reg_n_pc1 <- tibble(
taxon    = names(reg_direction),
logistic = abs(reg_direction),
pc1      = abs(pc1_direction)
) |>
pivot_longer(
cols = logistic:pc1,
names_to = "direction",
values_to = "loadings"
)

reg_n_pc1
```

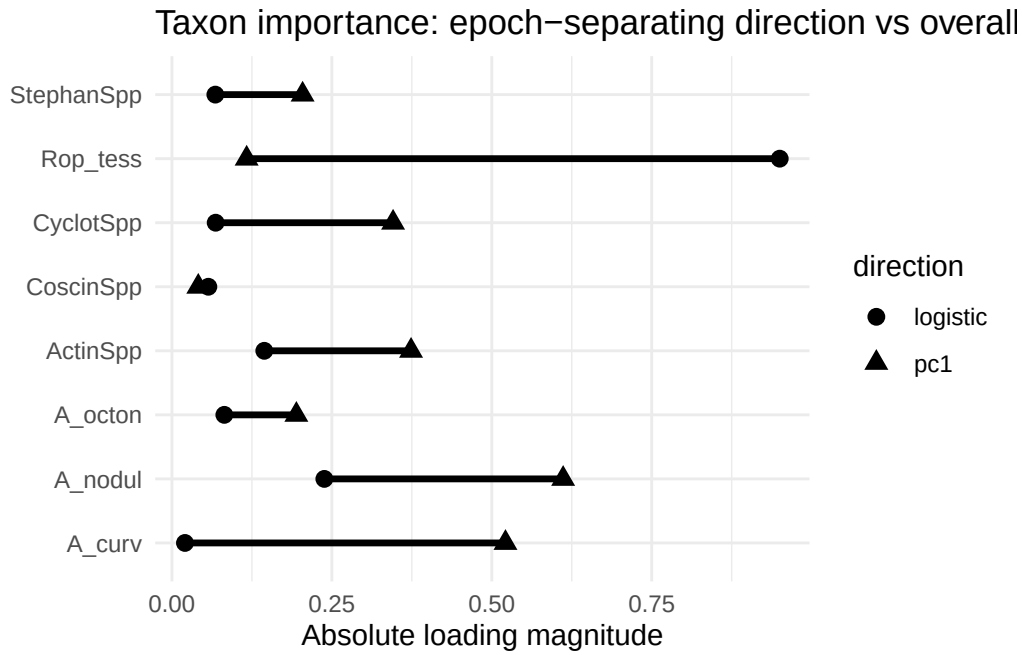
```
# A tibble: 16 x 3
   taxon      direction loadings
   <chr>      <chr>        <dbl>
 1 A_curv     logistic    0.0204
 2 A_curv     pc1         0.521
 3 A_octon    logistic    0.0819
 4 A_octon    pc1         0.195
 5 ActinSpp   logistic    0.144
 6 ActinSpp   pc1         0.374
 7 A_nodul    logistic    0.238
 8 A_nodul    pc1         0.612
 9 CoscinSpp  logistic    0.0569
10 CoscinSpp  pc1         0.0412
```

```
11 CyclotSpp  logistic    0.0684
12 CyclotSpp  pc1         0.346
13 Rop_tess   logistic    0.950
14 Rop_tess   pc1         0.117
15 StephanSpp logistic    0.0679
16 StephanSpp pc1         0.204
```

0.16.1 Dumbbell-style plot

```
dumbbell_plot <- reg_n_pc1 |>
ggplot(aes(x = loadings, y = taxon, shape = direction)) +
geom_line(aes(group = taxon), linewidth = 1.2) +
geom_point(size = 2.8) +
labs(
x = "Absolute loading magnitude",
y = "",
title = "Taxon importance: epoch-separating direction vs overall-variance direction"
) +
theme_minimal()

dumbbell_plot
```



0.17 Project samples onto both directions

```
diatoms5_matrix <- diatoms5 |>
select(A_curv:StephanSpp) |>
as.matrix()

pc1_score <- as.vector(diatoms5_matrix %*% pc1_direction)
logistic_score <- as.vector(diatoms5_matrix %*% reg_direction)

score_tables <- tibble(
ID = 1:nrow(diatoms5),
pc1_score = pc1_score,
logistic_score = logistic_score,
pleistocene = diatoms5$pleistocene
)

score_tables |>
summarize(
var_pc1 = var(pc1_score),
var_logistic = var(logistic_score)
)
```

```
# A tibble: 1 x 2
  var_pc1 var_logistic
    <dbl>        <dbl>
1    2.04         1.12
```

PC1 should typically have larger variance than a random direction because it is defined to maximize variance among unit-length directions.

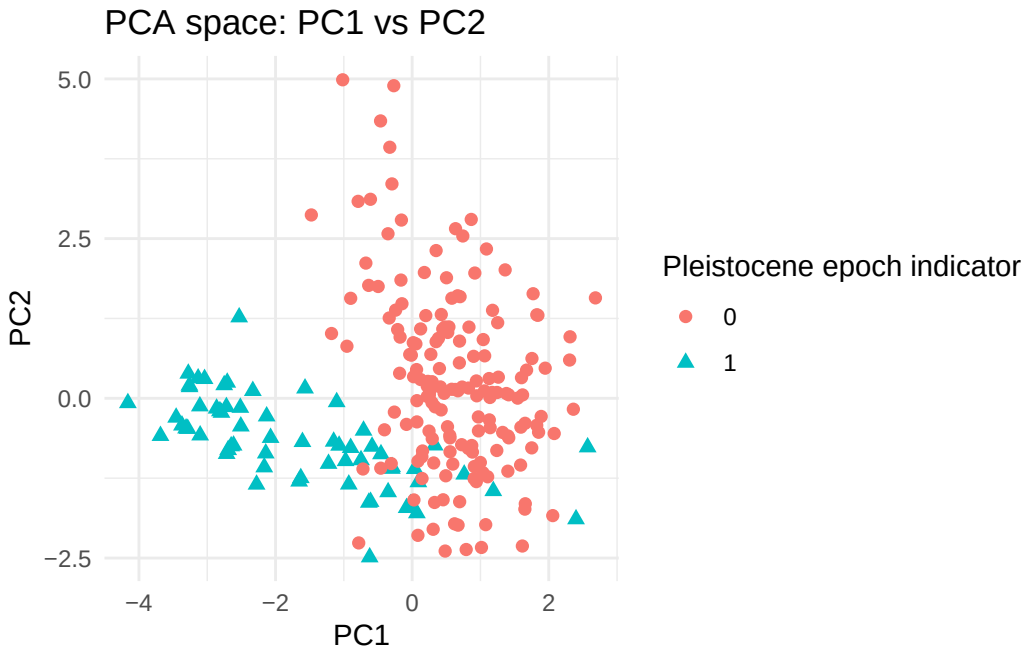
0.17.1 PC1 vs PC2 scatter (redundant coding)

```
pc_scores <- as_tibble(pca_result$x[, 1:2])

score_tables <- score_tables |>
mutate(
PC1 = pc_scores$PC1,
PC2 = pc_scores$PC2
)
```

```
pc_scatter <- score_tables |>
ggplot(aes(
x = PC1, y = PC2,
color = factor(pleistocene),
shape = factor(pleistocene)
)) +
geom_point(size = 2) +
labs(
color = "Pleistocene epoch indicator",
shape = "Pleistocene epoch indicator",
title = "PCA space: PC1 vs PC2"
) +
theme_minimal()

pc_scatter
```

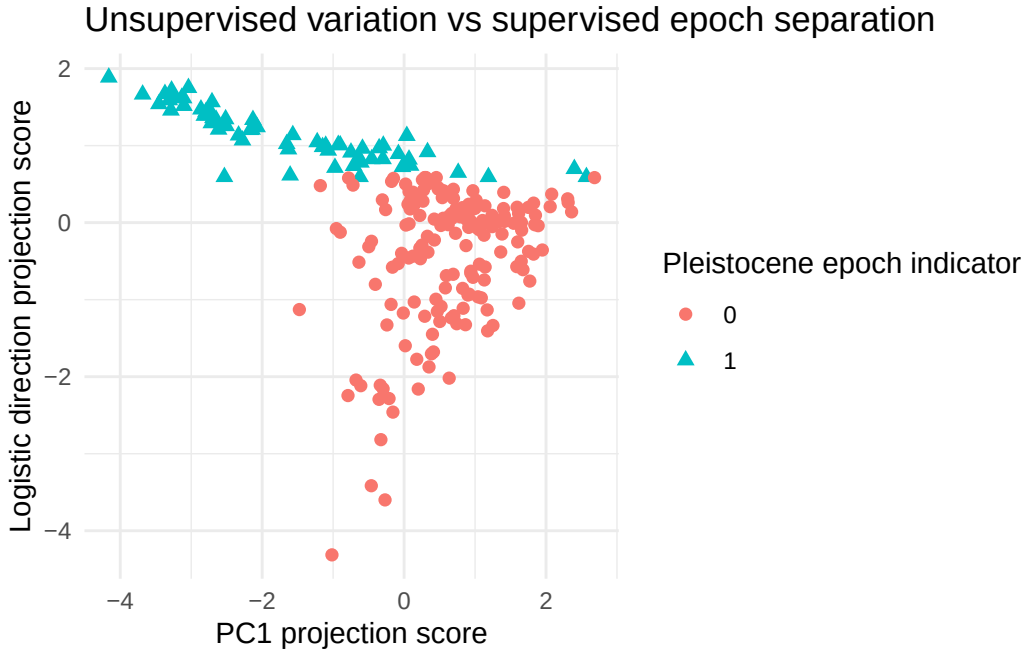


0.17.2 PC1 vs logistic score scatter

```
pc1_log_scatter <- score_tables |>
ggplot(aes(
```

```
x = pc1_score, y = logistic_score,
color = factor(pleistocene),
shape = factor(pleistocene)
)) +
geom_point(size = 2) +
labs(
color = "Pleistocene epoch indicator",
shape = "Pleistocene epoch indicator",
x = "PC1 projection score",
y = "Logistic direction projection score",
title = "Unsupervised variation vs supervised epoch separation"
) +
theme_minimal()

pc1_log_scatter
```



0.18 Results and discussion

0.18.1 What changed through time?

Across the time series plots, several taxa show pronounced shifts in relative abundance through the record. Converting counts to relative abundance was essential because Num.counted varies

across samples, so raw counts are not directly comparable.

0.18.2 Epoch-level differences (before vs after 11.7 kyr BP)

The ridgeline plot suggests Azpeitia nodulifer tends to have higher and more variable relative abundance in the older portion of the record (before 11.7 kyr BP), with the distribution shifting toward lower values after 11.7 kyr BP. This pattern is consistent with a community-level response to changing ocean conditions at the Pleistocene-Holocene transition.

0.18.3 Relationship with SST

Using LOESS to predict SST at diatom sample ages enables direct comparison between temperature and relative abundance. The hexbin plots provide a density-aware view of how abundances occupy different temperature ranges. In practice, this helps distinguish whether a taxon is:

- broadly present across temperatures (weak relationship), or
- concentrated in certain temperature bands (stronger relationship)

0.18.4 Community structure (PCA vs logistic regression)

PCA summarizes overall community variance without using epoch labels. Logistic regression explicitly seeks a direction that separates epochs. Comparing their loadings helps answer:

- Are the taxa that drive overall variability the same taxa that best distinguish epochs?

If the same taxa have large loadings in both, then the dominant community axis aligns with the epoch shift. If not, then the epoch difference is a secondary pattern relative to other sources of variability. In this dataset, the logistic regression achieves near-perfect epoch separation (residual deviance essentially 0), so I interpret the logistic coefficient vector as a supervised discriminant direction rather than relying on its p-values for inference. The normalized direction is also dominated by Rop_tess, indicating this taxon is a key contributor to epoch separation.

0.18.5 Why separation looks different across plots

In PC1-PC2 space, separation may be oblique and imperfect because PCs are not optimized for class separation. In PC1 vs logistic-score space, separation is typically cleaner along the logistic axis because that direction is trained to discriminate the binary epoch label.

0.19 Limitations and next steps

Limitations:

- Epoch labeling is a hard threshold at 11.7 kyr BP; transitions can be gradual.
- LOESS-based SST interpolation depends on the chosen span.
- Relative abundances are compositional-like measurements; more advanced methods (e.g., centered log-ratio transforms) could be explored.

Next steps:

- Sensitivity analysis for LOESS span and epoch boundary choice.
- Add confidence intervals for smoothed trends and effect sizes.
- Explore alternative multivariate methods (e.g., LDA, PLS-DA, compositional PCA).