PCA-NMDS

Santos G

Tabla de contenidos

1 Contexto de proyecto (PCA-NMDS)

El objetivo de este bloque es aplicar un Análisis de Componentes Principales (PCA) a variables morfométricas para resumir la variación multivariada en pocos ejes interpretables. El PCA ayudará a reducir la dimensionalidad, identificar correlaciones entre rasgos y construir índices compuestos que puedan usarse en modelos posteriores o informes técnicos.

El NMDS se aplicará sobre una matriz de abundancia de especies, con el propósito de evaluar la similitud entre sitios de muestreo y explorar posibles gradientes ecológicos subyacentes, complementando el enfoque morfométrico del PCA con una perspectiva basada en la composición comunitaria.

2 Carga de librerías y dataset

```
# Librerías
library(tidyverse)
library(janitor)
library(knitr)
library(vegan)
```

```r
library(ggrepel)
library(MVN)
library(psych)
library(ggcorrplot)
library(palmerpenguins)

# Cargar dataset
df_raw <- penguins %>% as_tibble()
```

3 Preparación de datos y verificación de supuestos

```r
# --- Preparación de datos ---
df_pca <- df_raw %>%
  select(species, island, bill_length_mm, bill_depth_mm,
         flipper_length_mm, body_mass_g) %>%
  drop_na()
```

```r
# Test de normalidad multivariada (Mardia)
mardia_test <- MVN::mvn(
  data = df_pca[, 3:6],
  mvn_test = "mardia",
  univariate_test = "AD",
  descriptives = FALSE,
  tidy = TRUE
)
norm_tbl <- mardia_test$multivariate_normality %>%
  dplyr::select(Test, Statistic, p.value, MVN) %>%
  dplyr::mutate(across(where(is.numeric), round, 3))

knitr::kable(norm_tbl)
```

Tabla 1: Normalidad multivariada (test de Mardia)

Test	Statistic	p.value	MVN
Mardia Skewness	130.931	<0.001	□ Not normal
Mardia Kurtosis	-2.499	0.012	□ Not normal

La evaluación de la normalidad multivariada (test de Mardia) indicó que los datos morfométricos de los pingüinos no siguen una distribución normal multivariada (ver Tabla 1). Tanto el componente de asimetría (skewness) como el de curtosis fueron significativos ($p < 0.05$), lo que sugiere desviaciones respecto a la simetría y al aplanamiento esperados bajo normalidad. Este resultado es común en variables biológicas que presentan heterogeneidad entre especies o efectos de tamaño corporal.

```r
# Esfericidad (Bartlett) y adecuación muestral (KMO)
bart <- psych::cortest.bartlett(cor(df_pca[, 3:6]), n = nrow(df_pca))
kmo  <- psych::KMO(cor(df_pca[, 3:6]))
```

```r
sphere_tbl <- tibble(
  Test = c("Bartlett's test of sphericity", "Kaiser-Meyer-Olkin (KMO)"),
  Statistic = c(round(bart$chisq, 3), NA),
  df = c(bart$df, NA),
  p_value = c(ifelse(bart$p.value < 0.001, "<0.001",
          round(bart$p.value, 3)), NA),
  Measure = c(NA, round(kmo$MSA, 3))
)

knitr::kable(sphere_tbl)
```

Tabla 2: Pruebas de esfericidad y adecuación muestral

Test	Statistic	df	p_value	Measure
Bartlett's test of sphericity	838.079	6	<0.001	NA
Kaiser-Meyer-Olkin (KMO)	NA	NA	NA	0.687

Las pruebas de esfericidad y adecuación muestral indicaron que los datos son apropiados para un análisis de componentes principales (ver Tabla 2), ya que el test de Bartlett resultó altamente significativo (χ^2 = 838.08, p < 0.001), rechazando la hipótesis nula de que la matriz de correlaciones sea una identidad, lo que confirma la existencia de correlaciones suficientes entre las variables. Por su parte, el índice KMO fue de 0.687, un valor considerado aceptable (por encima del umbral mínimo de 0.6), indicando que el tamaño de muestra y la estructura de correlaciones son adecuados para aplicar un PCA.

```r
# Matriz de correlaciones (ggcorrplot)

cor_mat <- cor(df_pca[, 3:6], method = "spearman",
               use = "pairwise.complete.obs")
ggcorrplot(
  cor_mat,
  hc.order = TRUE,
  type = "lower",
  lab = TRUE,
  lab_size = 3,
  method = "square",
  colors = c("#6D9EC1", "white", "#E46726"),
  title = "Matriz de correlaciones (Spearman)",
  ggtheme = ggplot2::theme_minimal()
)
```

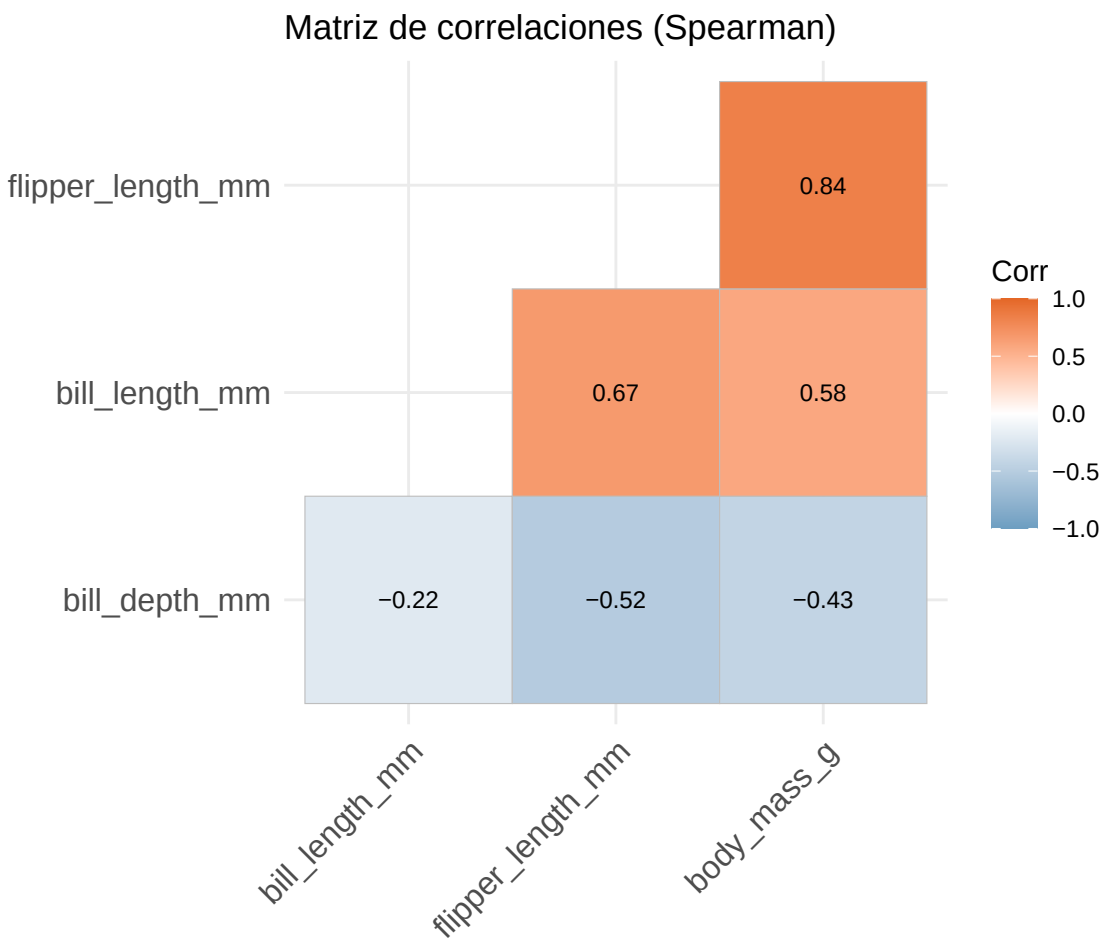

Figura 1: Matriz de correlaciones entre variables morfométricas (Spearman)

La matriz de correlaciones (Spearman) entre las variables morfométricas revela patrones claros de asociación entre las dimensiones corporales de los pingüinos. Las correlaciones más fuertes se observan entre longitud del aleta y masa corporal ($\rho = 0.84$), indicando que los individuos con aletas más largas tienden a ser más pesados. Asimismo, la longitud del pico se asocia positivamente con ambas variables ($\rho \approx 0.58$ a 0.67), lo que sugiere una coherencia morfológica general: los individuos de mayor tamaño presentan picos y aletas más desarrollados. Por el contrario, el ancho del pico muestra correlaciones negativas con las demás medidas ($\rho \approx$ -0.43 a -0.52), lo que indica que las especies con picos más amchos tienden a tener aletas más cortas y menor masa corporal (ver Figura 1).

4 Ejecutar PCA con `prcomp()` y extraer resultados

```
pca_fit <- prcomp(df_pca %>% select(where(is.numeric)), scale. = TRUE,
                  center = TRUE)
# Varianza explicada
pca_var <- pca_fit$sdev^2
pca_var_prop <- pca_var / sum(pca_var)

pca_summary <- tibble(
  PC = paste0("PC", seq_along(pca_var)),
```

```r
  sdev = round(pca_fit$sdev, 3),
  variance = round(pca_var, 3),
  prop.var = round(pca_var_prop, 3),
  cum.var = round(cumsum(pca_var_prop), 3)
)

knitr::kable(pca_summary)
```

Tabla 3: Desviaciones, varianza y proporción por componente

PC	sdev	variance	prop.var	cum.var
PC1	1.659	2.754	0.688	0.688
PC2	0.879	0.773	0.193	0.882
PC3	0.604	0.365	0.091	0.973
PC4	0.329	0.108	0.027	1.000

5 Gráficos: screeplot y biplot (scores + loadings)

```r
# Screeplot
scree_df <- pca_summary
ggplot(scree_df, aes(x = as.numeric(gsub("PC","",PC)), y = prop.var)) +
  geom_col() +
  geom_line(aes(y = cum.var), color = "blue") +
  geom_point(aes(y = cum.var), color = "blue") +
  labs(x = "Componente principal", y = "Proporción de varianza explicada",
       title = "Screeplot PCA") +
  theme_minimal()
```

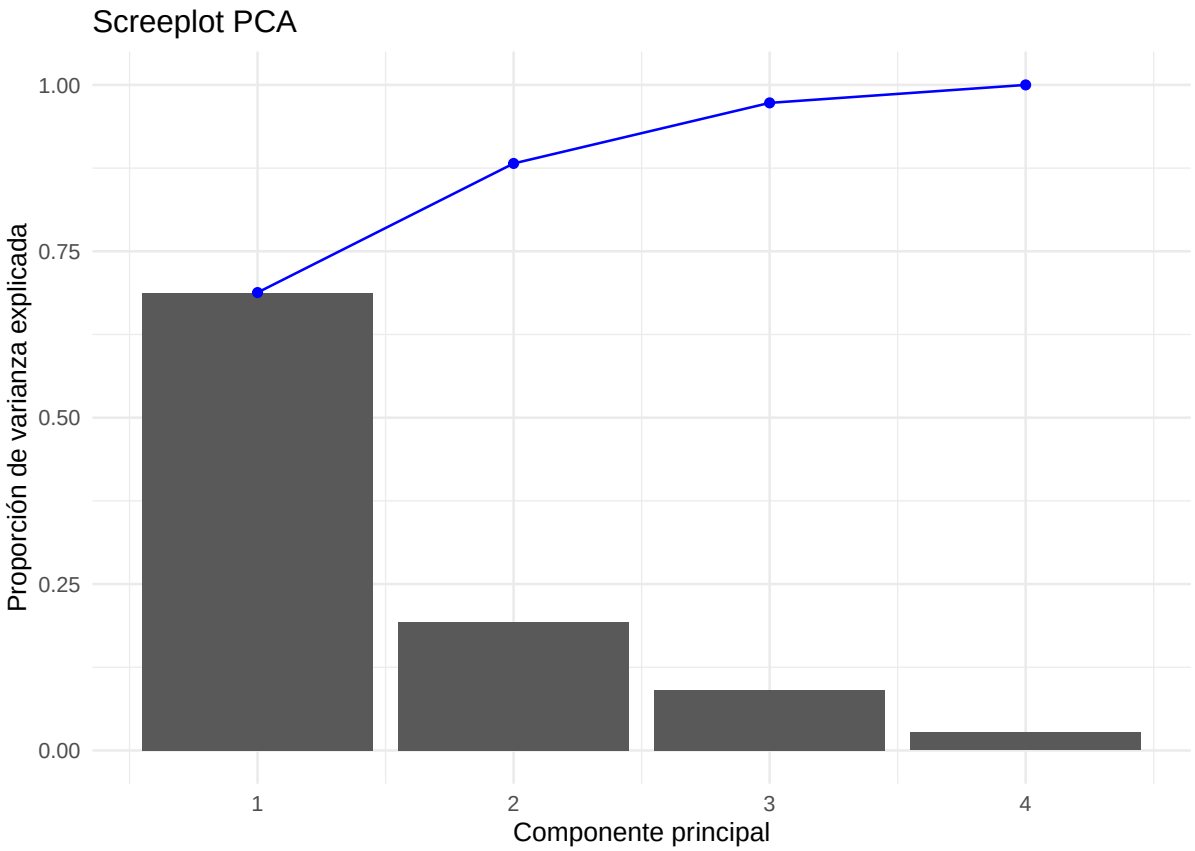

Figura 2: Screeplot PCA:proporción de varianza explicada por componente

El análisis de componentes principales muestra que las dos primeras componentes (PC1 y PC2) explican conjuntamente alrededor del 88.2 % de la varianza total en las variables morfométricas, lo cual es un nivel de representación muy adecuado para la reducción de dimensionalidad (ver Tabla 3 y Figura 2) .

```
# Cargas (loadings) de cada variable en las primeras dos componentes
loadings_tbl <- as_tibble(pca_fit$rotation[, 1:2], rownames = "Variable") %>%
  rename(PC1 = PC1, PC2 = PC2) %>%
  mutate(across(where(is.numeric), round, 3))

knitr::kable(loadings_tbl)
```

Tabla 4: Contribución de las variables a los componentes principales

Variable	PC1	PC2
bill_length_mm	0.455	-0.597
bill_depth_mm	-0.400	-0.798
flipper_length_mm	0.576	-0.002
body_mass_g	0.548	-0.084

La primera componente (PC1), que explica el 68.8 % de la varianza total, refleja un gradiente general de tamaño corporal. Las variables con mayores pesos positivos son *flipper_length_mm* (0.576) y *body_mass_g* (0.548), seguidas por *bill_length_mm* (0.455). En contraste, *bill_depth_mm* presenta un peso negativo moderado (-0.400). Esto indica que los individuos con mayores valores en PC1 tienden

a tener alas más largas, mayor masa corporal y picos más largos pero menos profundos, es decir, una morfología asociada a un tamaño corporal globalmente mayor (ver Tabla 4).

La segunda componente (PC2), que aporta un 19.3 % adicional de la varianza, representa principalmente variaciones en la forma del pico. Aquí destacan los pesos negativos elevados de *bill_depth_mm* (-0.798) y *bill_length_mm* (-0.597), lo que sugiere un eje de contraste entre especies con picos largos y estrechos frente a aquellas con picos cortos y robustos. Las otras variables (*flipper_length_mm* y *body_mass_g*) tienen pesos cercanos a cero, mostrando poca influencia sobre esta dimensión (ver Tabla 4).

```r
# Biplot de variables
biplot_data <- as_tibble(pca_fit$x[, 1:2]) %>%
  mutate(species = df_pca$species)

# vectores de las variables (loadings)
loadings <- as.data.frame(pca_fit$rotation[, 1:2])

ggplot(biplot_data, aes(PC1, PC2, color = species)) +
  geom_point(alpha = 0.7, size = 2) +
  geom_segment(data = loadings,
               aes(x = 0, y = 0, xend = PC1 * 3, yend = PC2 * 3),
               arrow = arrow(length = unit(0.25, "cm")),
               color = "black") +
  geom_text_repel(data = loadings,
                  aes(x = PC1 * 3.2, y = PC2 * 3.2, label = rownames(loadings)),
                  color = "black", size = 3.5) +
  labs(title = "Biplot PCA: especies y variables morfométricas",
       x = "Componente 1",
       y = "Componente 2") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

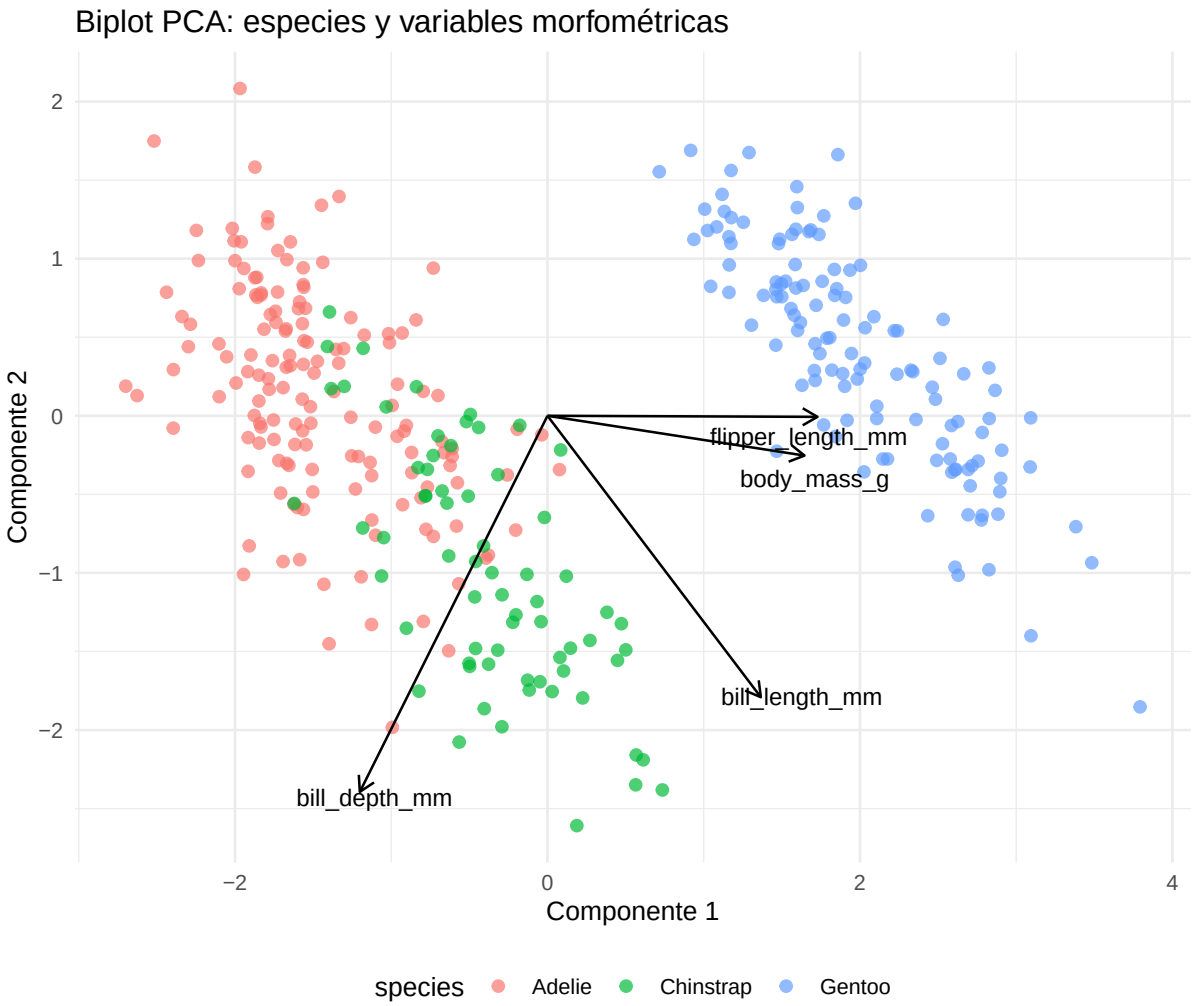

Figura 3: Biplot PCA: relación entre variables y componentes principales

El biplot de componentes principales muestra simultáneamente la posición de los individuos de cada especie en el espacio definido por los dos primeros componentes y la dirección de las variables morfométricas que contribuyen a dicha variación.

El primer componente (PC1) está fuertemente asociado con el tamaño corporal total, determinado principalmente por las variables *flipper_length_mm* y *body_mass_g*, que presentan vectores largos y orientados en la misma dirección. Las especies con valores altos en este eje, como *Gentoo*, se agrupan hacia el extremo positivo del PC1, indicando individuos más grandes y pesados. En contraste, *Adelie* se concentra en el extremo negativo, reflejando individuos de menor tamaño (ver Figura 3).

El segundo componente (PC2) captura variaciones en la morfología del pico, principalmente asociadas a las variables *bill_length_mm* y *bill_depth_mm*. En este eje, *Chinstrap* tiende a ocupar posiciones intermedias o positivas, caracterizadas por picos más largos y delgados, mientras que *Adelie* mantiene picos más cortos y profundos, situándose en la dirección opuesta (ver Figura 3).

En síntesis, el análisis de componentes principales permite visualizar de manera clara la diferenciación morfológica entre las especies de pingüinos, destacando que el tamaño corporal y la forma del pico son los principales ejes de variación. Estas diferencias reflejan adaptaciones específicas de cada especie a su entorno y estrategias alimenticias, lo que respalda la utilidad del PCA como herramienta para comprender patrones de variación biológica y relaciones morfométricas entre grupos cercanos.

6 Preparación de datos — transformación Hellinger (NMDS)

```r
data(varespec)
data(varechem)

# Hellinger transforma abundancias para métodos basados en distancia euclidiana,
varespec_hel <- decostand(varespec, method = "hellinger")
```

7 Ejecutar NMDS con `metaMDS()` (Bray-Curtis por defecto)

```r
# Ajuste del NMDS (sin mostrar mensajes de ejecución)
set.seed(42)
suppressMessages({
  capture.output({
    nmds <- metaMDS(
      varespec_hel,
      distance = "bray",
      k = 2,
      trymax = 100,
      autotransform = FALSE
    )
  })
})
```

```r
# Curva de ajuste (stressplot)
vegan::stressplot(nmds)
```

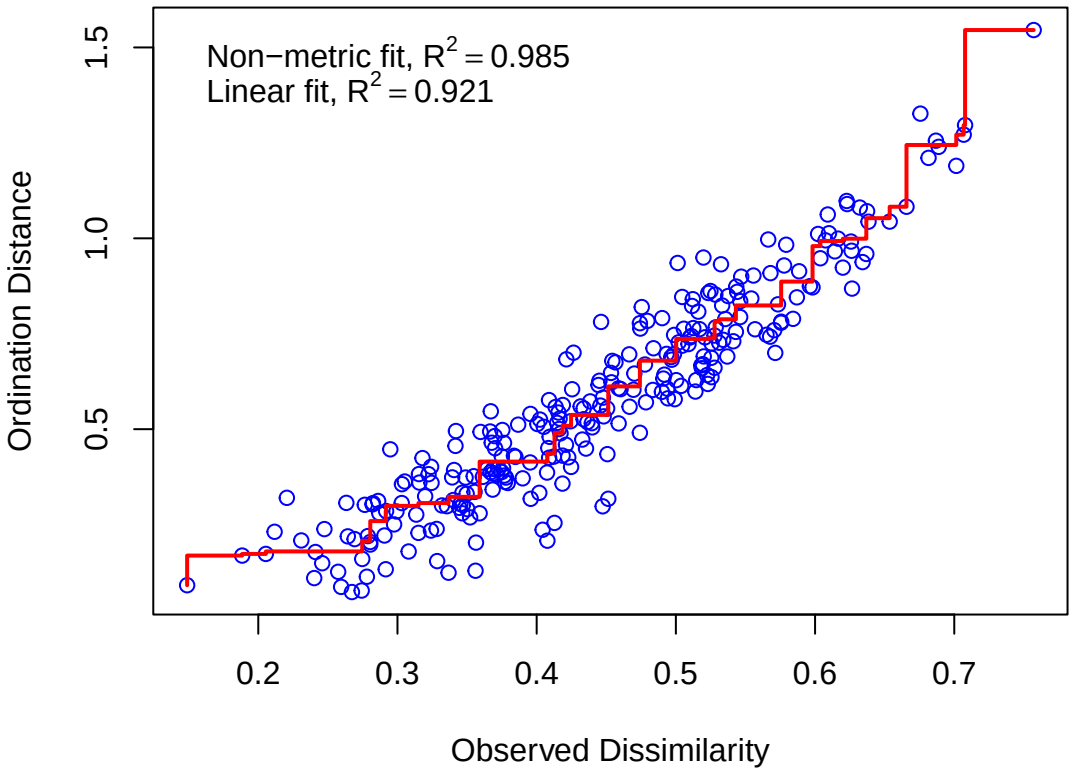

Figura 4: Curva de ajuste NMDS (stressplot) para evaluar la calidad del ordenamiento

El análisis de escala multidimensional no métrica (NMDS) alcanzó un valor de *stress* de 0.12, lo cual indica una representación bidimensional adecuada según los criterios de Kruskal (valores < 0.2 son aceptables, y < 0.1 excelentes). Los indicadores de ajuste complementarios refuerzan la calidad del ordenamiento: Non-metric fit $R^2 = 0.985$: sugiere que el 98.5 % de la variación en las distancias rankeadas se preserva en el espacio de dos dimensiones. Linear fit $R^2 = 0.921$: indica una alta correspondencia lineal entre las distancias originales y las del NMDS (ver Figura 4).

8 Visualización NMDS: sitios + especies + envfit

```r
# Ajustar variables ambientales (envfit)
ef <- envfit(nmds, varechem, permutations = 999)

# Extraer resultados del envfit
ef_results <- as.data.frame(ef$vectors$arrows)
ef_r2 <- ef$vectors$r
ef_p <- ef$vectors$pvals

# Crear tabla resumen
tabla_envfit <- ef_results %>%
```

```
  rownames_to_column("Variable") %>%
  mutate(
    R2 = round(ef_r2, 3),
    p_value = signif(ef_p, 3),
    NMDS1 = round(NMDS1, 3),
    NMDS2 = round(NMDS2, 3)
  ) %>%
  arrange(desc(R2))

knitr::kable(
  tabla_envfit,
  align = "lcccc"
)
```

Tabla 5: Resultados del ajuste de variables ambientales (envfit) sobre el espacio NMDS

Variable	NMDS1	NMDS2	R2	p_value
Mn	0.999	0.043	0.468	0.001
Humdepth	0.994	0.110	0.466	0.003
Al	-0.981	0.196	0.460	0.002
Fe	-0.970	0.244	0.414	0.003
Ca	0.959	0.284	0.308	0.020
Mg	0.998	0.065	0.223	0.077
P	0.882	0.471	0.217	0.084
pH	-0.875	0.484	0.215	0.069
K	0.954	0.300	0.195	0.104
Baresoil	0.889	-0.458	0.184	0.123
Zn	0.954	-0.300	0.149	0.178
N	-0.024	-1.000	0.091	0.339
Mo	-0.606	-0.796	0.076	0.429
S	0.853	0.522	0.040	0.658

Los resultados del ajuste de variables ambientales sobre el espacio de ordenamiento NMDS (ver Tabla 5) permiten cuantificar la influencia de cada factor edáfico sobre la composición de especies. Los valores de R^2 indican la fuerza de la relación entre cada variable y la estructura de la comunidad, mientras que los valores de p reflejan su significancia estadística.

Las variables Mn, Humdepth, Al y Fe muestran los valores de R^2 más altos (≥ 0.41, $p < 0.01$), lo que evidencia que son los principales controladores de la variación observada. Mn y Humdepth se orientan hacia el cuadrante derecho del diagrama, representando sitios con suelos más fértiles, mayor contenido de materia orgánica y disponibilidad de nutrientes. En contraste, Al y Fe apuntan hacia el cuadrante izquierdo, definiendo un gradiente de acidez y metalización, asociado a suelos menos fértiles y más restrictivos para el crecimiento vegetal.

Un segundo grupo de variables, con valores intermedios de R^2 (entre 0.20 y 0.30), incluye Ca, Mg, P y pH. Estas variables refuerzan el eje principal del ordenamiento: mientras Ca, Mg y P se agrupan hacia la derecha, representando la fertilidad edáfica, el pH (con sentido opuesto) marca el extremo ácido de dicho gradiente.

Finalmente, variables como K, Baresoil, Zn, N, Mo y S presentan valores bajos de R^2 (< 0.20) y sin significancia estadística ($p > 0.05$), por lo que su contribución a la organización global de las

comunidades es menor. No obstante, podrían reflejar variaciones locales o efectos secundarios asociados a la exposición, perturbación o heterogeneidad microambiental.

En conjunto, los resultados del *envfit* confirman que el ordenamiento NMDS está estructurado principalmente por un gradiente de fertilidad–acidez del suelo, donde la composición vegetal responde a la disponibilidad de nutrientes y a las condiciones edáficas que limitan o favorecen el desarrollo de las especies.

```r
# Extraer coordenadas de sitios
sites_scores <- as_tibble(scores(nmds, display = "sites")) %>%
  mutate(site = rownames(varespec))

# Extraer coordenadas de especies
species_scores <- as_tibble(scores(nmds, display = "species")) %>%
  rownames_to_column("species")

# Ajustar variables ambientales (envfit)
ef <- envfit(nmds, varechem, permutations = 999)

# Mantener nombres correctos de variables ambientales
ef_arrows <- as.data.frame(ef$vectors$arrows)
ef_arrows$var <- rownames(ef_arrows)
ef_arrows <- ef_arrows %>%
  as_tibble() %>%
  rename(NMDS1 = NMDS1, NMDS2 = NMDS2)

# Gráfico NMDS con vectores ambientales
ggplot(sites_scores, aes(x = NMDS1, y = NMDS2)) +
  geom_point(size = 3, alpha = 0.8) +
  geom_text_repel(aes(label = site), size = 3, max.overlaps = 15) +
  geom_segment(
    data = ef_arrows,
    aes(x = 0, y = 0, xend = NMDS1, yend = NMDS2),
    arrow = arrow(length = unit(0.25, "cm")),
    color = "black"
  ) +
  geom_text_repel(
    data = ef_arrows,
    aes(x = NMDS1, y = NMDS2, label = var),
    size = 3,
    color = "black"
  ) +
  labs(
    title = paste0("NMDS (stress = ", round(nmds$stress, 3), ")"),
    x = "NMDS1",
    y = "NMDS2"
  ) +
  theme_minimal()
```

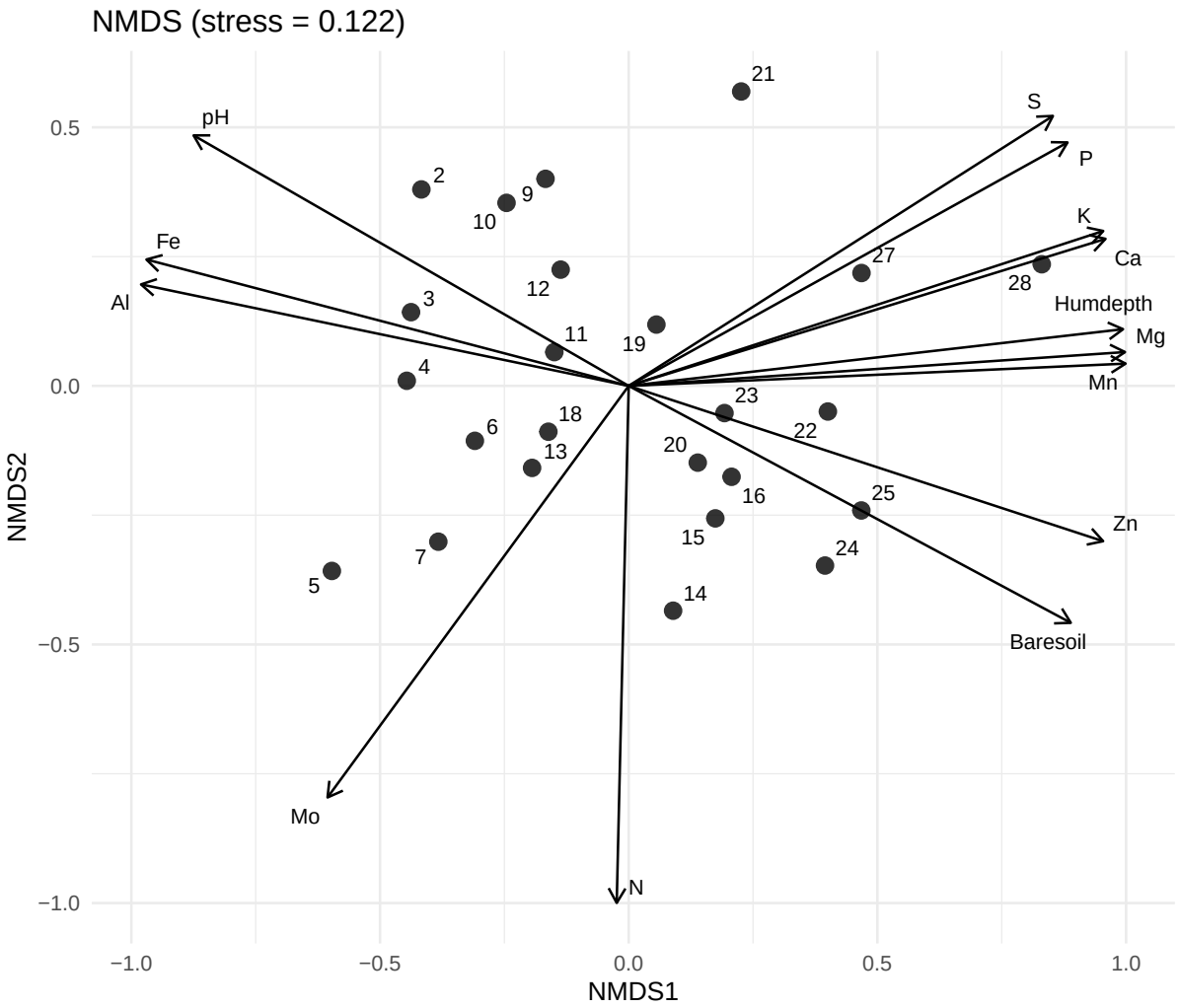

Figura 5: Ordenamiento NMDS basado en distancias de Bray–Curtis, con ajuste de variables ambientales mediante envfit

El biplot del NMDS representa la estructura de similitud entre los sitios en función de la composición de especies (varespec), incorporando además los gradientes ambientales medidos en cada sitio (varechem). Los vectores ambientales indican la dirección y magnitud de las variables que mejor explican la variación en la comunidad biológica: cuanto más largos y definidos, mayor es su poder explicativo sobre la distribución de las especies (ver Figura 5).

El nitrógeno (N) muestra un vector orientado hacia la parte inferior del eje NMDS2, lo que sugiere que los sitios en esa dirección presentan concentraciones elevadas de este nutriente y comunidades asociadas a ambientes más enriquecidos.

Las variables S (azufre), P (fósforo), K (potasio) y Ca (calcio) apuntan hacia el cuadrante superior derecho del diagrama, indicando que los sitios en esta región se asocian a suelos más fértiles y con mayor disponibilidad de nutrientes, donde predominan especies adaptadas a condiciones más productivas.

Por otro lado, pH, Fe (hierro) y Al (aluminio) se orientan hacia el cuadrante superior izquierdo, definiendo un gradiente de acidez y metalización. Los sitios en esta dirección tienden a presentar suelos más ácidos y menos fértiles, lo que puede restringir el crecimiento vegetal y favorecer especies tolerantes a condiciones edáficas limitantes.

Las variables Baresoil (suelo desnudo) y Zn (zinc) apuntan hacia el cuadrante inferior derecho, lo que podría reflejar sitios más abiertos o degradados, con menor cobertura vegetal y condiciones más extremas o expuestas.

Además, la proximidad entre los vectores de S, P, K y Ca indica que estas variables co-varían positivamente, definiendo un eje de fertilidad edáfica. En cambio, la orientación opuesta de pH, Fe y Al revela un gradiente inverso de acidez y metalización, contrapuesto a la fertilidad. Por su parte, el nitrógeno (N) mantiene una posición relativamente independiente, lo que sugiere que su variación no está directamente alineada con los principales gradientes ambientales.

La posición de los sitios en el espacio NMDS refuerza estas asociaciones:

- Los sitios cercanos entre sí (por ejemplo, 1, 3, 9, 10 y 12) presentan una composición florística similar, vinculada con suelos ácidos y altos contenidos de Fe y Al.
- En contraste, los sitios 14, 15, 24 y 25 se agrupan hacia el extremo inferior derecho, asociados a altos valores de Zn y mayor proporción de suelo desnudo, característicos de hábitats más empobrecidos o erosionados.
- Algunos sitios más dispersos, como 21 o 27, podrían representar zonas de transición ecológica, donde confluyen gradientes intermedios de fertilidad y acidez.

En conjunto, el ordenamiento NMDS muestra una estructura clara donde los patrones de composición de especies se explican principalmente por gradientes edáficos de fertilidad, acidez y contenido metálico, evidenciando la fuerte influencia del ambiente físico sobre la organización ecológica de las comunidades.

9 Conclusiones

El análisis de componentes principales permitió reducir cuatro variables morfométricas a dos ejes biológicamente interpretables, PC1 (68.8 %): gradiente de tamaño corporal general.y PC2 (19.3 %): gradiente de forma del pico. Estos resultados confirman que las diferencias entre *Adelie*, *Chinstrap* y *Gentoo* responden principalmente a contrastes en masa corporal, longitud de aleta y morfología del pico, variables estrechamente vinculadas con la ecología trófica y las estrategias adaptativas de cada especie.

En síntesis, el NMDS evidenció que la estructura de las comunidades vegetales está fuertemente condicionada por gradientes de fertilidad (definido por concentraciones de S, P, K y Ca) y acidez del suelo (explicado por pH, Fe y Al), y constituye un método flexible y poderoso para explorar patrones ecológicos complejos, incluso en escenarios con datos no normales o altamente heterogéneos.