

Análisis Multivariados en R

Curso Colaborativo IIAP - UNAMAD

Irwing S. Saldaña

Instituto de Ciencias Antonio Brack

Departamento de Ecoinformática y Biogeografía

Perú, 2021



Blgo. Irving S. Saldaña
Instructor

,
Instituto de Ciencias Antonio Brack, Perú

[Website](#) | [ResearchGate](#) | [Linkedin](#) | [R Latam Blog](#)



Necesidad de representar el mundo en N-dimensiones

[Paradoja de Simpson]



Paradoja de Simpson

Para comprender la importancia de los análisis multivariados comencemos revisando esta base de datos.

```
library(tidyverse)
library(palmerpenguins)
pinguinos<-na.omit(penguins)
```

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Adelie	Torgersen	39.1	18.7	181	3750	male	2007
Adelie	Torgersen	39.5	17.4	186	3800	female	2007
Adelie	Torgersen	40.3	18.0	195	3250	female	2007
Adelie	Torgersen	36.7	19.3	193	3450	female	2007
Adelie	Torgersen	39.3	20.6	190	3650	male	2007
Adelie	Torgersen	38.9	17.8	181	3625	female	2007



Paradoja de Simpson

Código

Gráfico

Código+regresión

Gráfico+regresión

Código+Variable

Gráfico+Variable

```
# Gráfico con ggplot2
library(tidyverse)
pinguinos %>%
  ggplot(aes(x=bill_length_mm,
             y=bill_depth_mm)) +
  geom_point() +
  labs(x="Longitud Pico (mm)",
       y="Profundidad Pico (mm)",
       title="Longitud de pico en función de su profundidad") +
  theme_bw()
```



Introducción a la Estadística Multivariada

[Ordenamientos - Agrupamientos]



El Mundo es Multivariado

- Implica técnicas que involucran tres o más variables a la vez.
- Típicamente no se incluyen como *análisis multivariados* del tipo regresión múltiple, MANOVA.

Métodos interdependientes	Métodos dependientes
Conocido como ordenamiento sin restricciones	Conocido como ordenamiento con restricciones (Canónico)
1 grupo de variables	2 grupos de variables : X e
Permiten describir	Permiten predecir
No buscan definir la relación entre variables dependientes o independientes	Buscan definir la relación entre variables dependientes o independientes
Busca observar patrones de agrupamiento en los datos	Busca definir la causa de los patrones de agrupamiento
No hay contraste de hipótesis.	Hay contraste de hipótesis
Veremos: Clustering, CA, PCA, MDS, NMDS	Veremos: RDA, CCA



Bases de datos en Análisis Multivariados

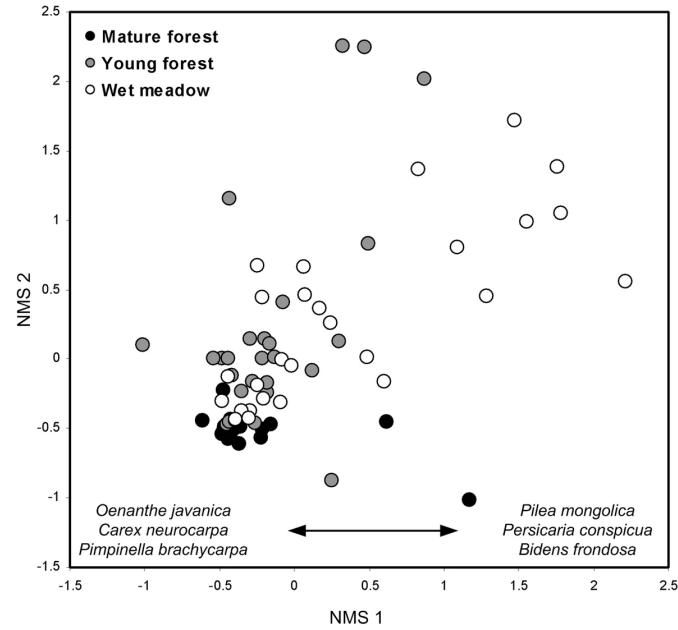
Sitio	A	B	C	D	E	Profundidad	Polucion	Temp	Sedimento
s1	0	2	9	14	2	72	4.8	3.5	S
s2	26	4	13	11	0	75	2.8	2.5	C
s3	0	10	9	8	0	59	5.4	2.7	C
s4	0	0	15	3	0	64	8.2	2.9	S
s5	13	5	3	10	7	61	3.9	3.1	C
s6	31	21	13	16	5	94	2.6	3.5	G
s7	9	6	0	11	2	53	4.6	2.9	S
s8	2	0	0	0	1	61	5.1	3.3	C
s9	17	7	10	14	6	68	3.9	3.4	C
s10	0	5	26	9	0	69	10.0	3.0	S
s11	0	8	8	6	7	57	6.5	3.3	C
s12	14	11	13	15	0	84	3.8	3.1	S



¿Qué es ordenamiento?

- Los métodos de ordenamientos "ordenan" las observaciones "filas" (que pueden representar individuos, sitios, parcelas, etc.), usando la información de múltiples variables "columnas".
- Se utilizan técnicas matemáticas de reducción de dimensiones para lograr proyectar toda (o la mayor parte) la información de una base de datos en tan solo 2 dimensiones (gráficos en plano cartesiano).

Plot_UI	Species	Group	DBH (cm)	Form	Plot size	Distance (Alive or dead)
1.1_1	<i>Eucalyptus obliqua</i>	EUC	93	1 full	L	
1.1_2	<i>Eucalyptus obliqua</i>	EUC	248	3 full	L	
1.1_3	<i>Eucalyptus obliqua</i>	EUC	190	2 full	L	
1.1_3	<i>Eucalyptus regnans</i>	EUC	93	1 full	L	
14.1_1	<i>Eucalyptus regnans</i>	EUC	290	3 full	L	
14.1_2	<i>Eucalyptus regnans</i>	EUC	280	4 full	L	
14.1_2	<i>Eucalyptus regnans</i>	EUC	310	4 full	L	
14.1_2	<i>Eucalyptus regnans</i>	EUC	250	5 full	D	
14.1_3	<i>Eucalyptus regnans</i>	EUC	330	2 full	L	
14.1_3	<i>Eucalyptus regnans</i>	EUC	180	7 full	D	
14.1_3	<i>Eucalyptus regnans</i>	EUC	250	10 full	D	
14.1_3	<i>Eucalyptus regnans</i>	EUC	280	2 full	L	
17.8_1	<i>Eucalyptus regnans</i>	EUC	138	2 extra	28m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	90	2 extra	30m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	220	3 extra	25m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	50	1 extra	26m	L
17.8_1	<i>Eucalyptus obliqua</i>	EUC	85	1 extra	24m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	23	extra	26m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	115	2 extra	24m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	43	1 extra	22m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	15	1 extra	22m	L
17.8_1	<i>Eucalyptus obliqua</i>	EUC	160	2 extra	26m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	80	1 extra	23m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	46	1 extra	25m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	80	1 extra	25m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	25	1 extra	25m	L
17.8_1	<i>Eucalyptus regnans</i>	EUC	330	3 extra	30m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	190	3 full	L	
17.8_2	<i>Eucalyptus obliqua</i>	EUC	54	1 extra	22m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	36	1 extra	22m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	39	1 extra	22m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	75	1 extra	25m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	40	1 extra	25m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	54	1 extra	25m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	400	3 extra	30m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	95	2 extra	26m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	51	1 extra	27m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	23	1 extra	27m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	49	1 extra	28m	L
17.8_2	<i>Eucalyptus obliqua</i>	EUC	110	2 extra	30m	L



Buscando relaciones entre variables

[Variables y la manera de investigar sus relaciones]



Variables

- **Categóricas (Factores):**

- *Binarias* (0,1; presencia-ausencia)
- *Ordinal* (Alto > Medio > Bajo, mes de muestreo,)
- *Nominal* (Sitio A, Sitio B, Sitio C)

- **Numéricas:**

- *Discretas* (Abundancias-Conteos: 1,2,3,4,5)
- *Continuas* (Mediciones: 1.96, 4.56, 3.75)
 - *Ratio* (Cero es absoluto, no negativos. Biomasa, concentraciones, longitudes, pesos)
 - *Intervalo* (Cero no es absoluto. Variables asociadas al tiempo)
 - *Porcentajes* (son especiales, restringidas al rango 0-1)



Análisis Exploratorio Básico

También conocido como exploratory data analysis (EDA). Permite:

- Encontrar patrones en los datos.
- Sugerir estrategias de modelamiento.
- Corregir algún error en la base de datos.

1. Univariado

- Gráficos de barras.
- Gráfico de densidad.
- Histograma de frecuencias.
- Cuantiles o gráfico de boxplot.

2. Bivariado

- Variables **numérica vs numérica**: dispersión de puntos, densidad 2D.
- Variables **numérica vs categórica**: boxplot.
- Variables **categórica vs categórica**: gráfico de barras.



Análisis Exploratorio Básico

Código Gráfico ggpairs()

```
# Carga el archivo "variables ambientales.xlsx"
ambiente <- openxlsx::read.xlsx(file.choose())

# Activa la librería GGally
library(GGally)

# Gráfico
ggpairs(ambiente,
        upper= list(continuous= wrap("cor", method="spearman",
                                      color="#5f00db", size=3.5)),
        lower= list(continuous= wrap("points",
                                      color="#5f00db", alpha=0.5)),
        diag= list(continuous= wrap("densityDiag",
                                    fill="#5f00db", alpha=0.5)))+
theme_test()
```



Libro R for Data Science (Wickham & Grolemund, 2017)

Visita aquí el libro Web, capítulo: Análisis Exploratorio de Datos

Search

Table of contents

[Bienvenida](#)

[1 Introducción](#)

Explorar

[2 Introducción](#)

[3 Visualización de datos](#)

[4 Flujo de trabajo: conocimientos básicos](#)

[5 Transformación de datos](#)

[6 Flujo de trabajo: Scripts](#)

[7 Análisis exploratorio de datos](#)

7 Análisis exploratorio de datos (EDA)

7.1 Introducción

Este capítulo te mostrará cómo usar la visualización y la transformación para explorar tus datos de manera sistemática, una tarea que las personas de Estadística suelen llamar análisis exploratorio de datos, o *EDA* (por sus siglas en inglés **e**xploratory **d**ata **a**nalysis). El *EDA* es un ciclo iterativo en el que:

1. Generas preguntas acerca de tus datos.
2. Buscas respuestas visualizando, transformando y modelando tus datos.

On this page

[7 Análisis exploratorio de datos \(EDA\)](#)

[7.1 Introducción](#)

[7.2 Preguntas](#)

[7.3 Variación](#)

[7.4 Valores faltantes](#)

[7.5 Covariación](#)

[7.6 Patrones y modelos](#)

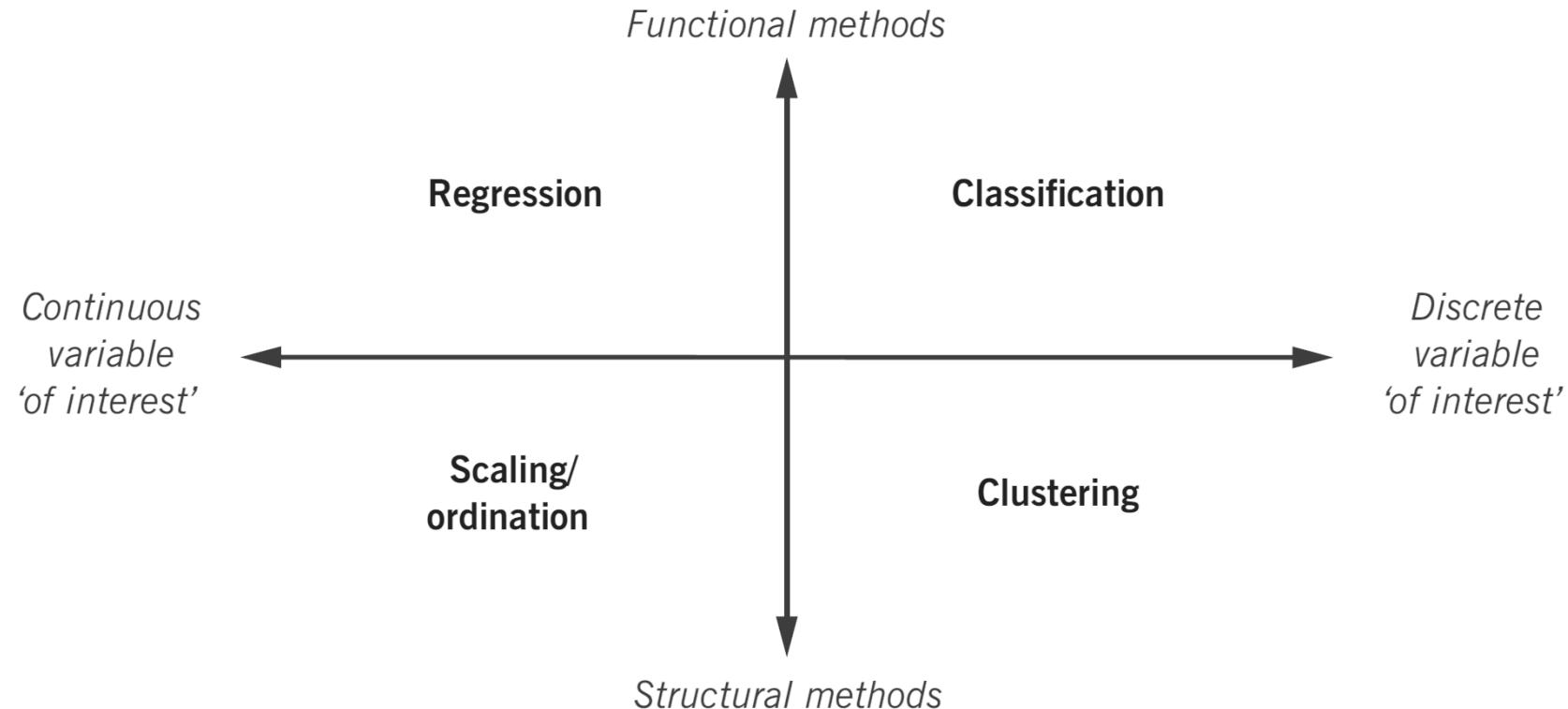
[7.7 Argumentos en ggplot2](#)

Bilo. Irving S. Saldaña | Instituto de Ciencias Antonio Brack, Perú

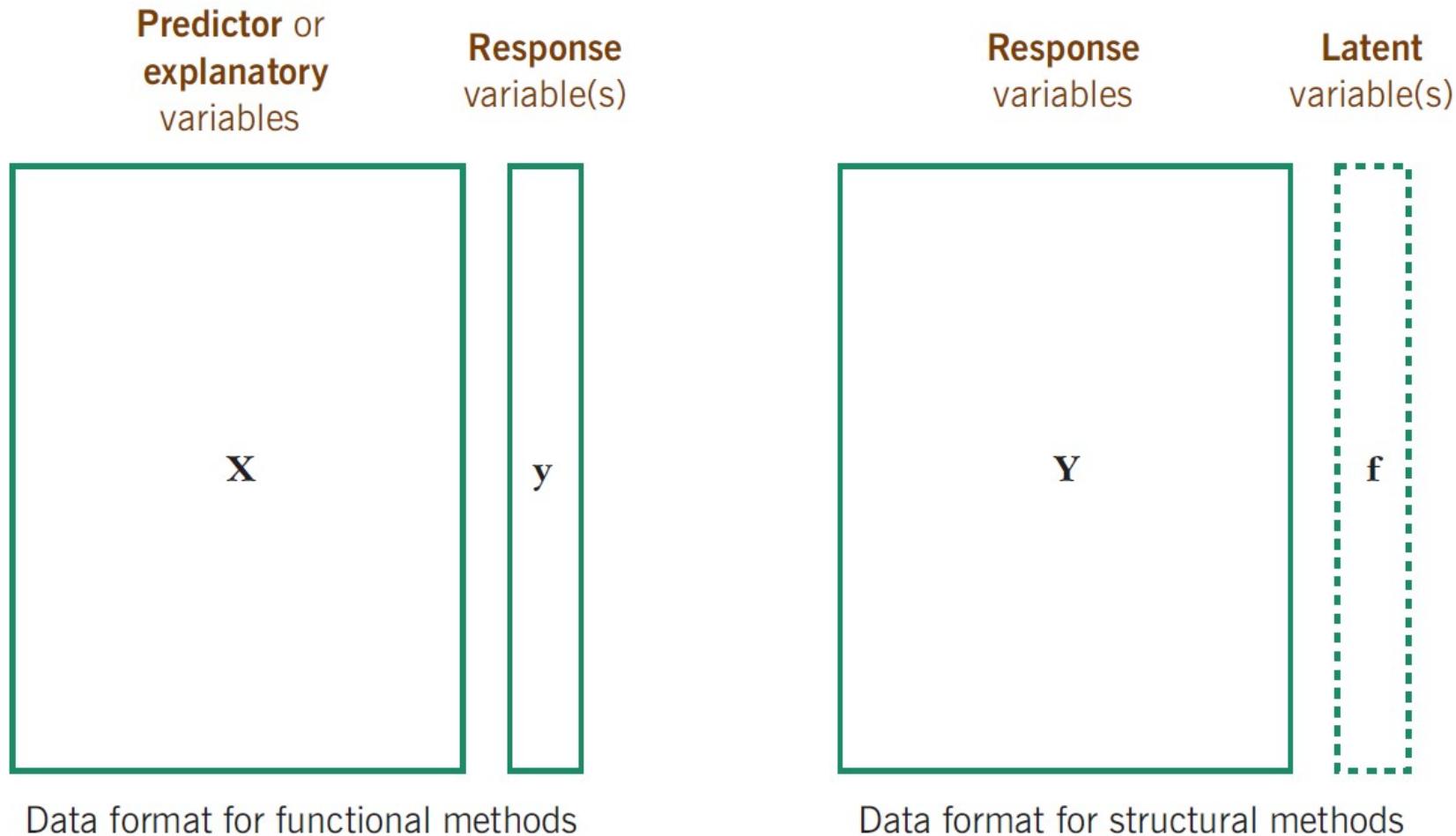
17

Greenacre & Primicerio (2014) propusieron la división de los Análisis Multivariados aplicados a las ciencias ecológicas en:

- **Métodos Estructurales:** buscan la estructura de la base de datos.
- **Métodos Funcionales:** buscan explicar y.
- **Métodos Híbridos:** mezclan características de los anteriores.



La estructura de las matrices de datos está relacionada con el tipo de análisis que se desea realizar



Greenacre, M., & Primicerio, R. (2014). Multivariate analysis of ecological data. Fundacion BBVA.



Transformaciones

1) Logaritmo $\log()$

- Si $x \rightarrow 0$, entonces $\log(x) \rightarrow -\infty$
- Si $x = 0$, entonces $\log(x = 1) = 0$
- Si $x \rightarrow \infty$, entonces $\log(x) \rightarrow \infty$
- Además: $\log(ab) = \log(a) + \log(b)$

2) Transformación de Box-Cox-Chord

$$g(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

3) Raíz cuadrada $\sqrt{()}$

4) Transformación de Hellinger (y_{i+} es $j = 1, \dots, m$)

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$$

5) Transformación de Chi-cuadrado

$$\chi^2_{x,y} = \sqrt{\sum_{j=1}^J \frac{1}{C_j} (x_j - y_j)^2}$$

6) Estandarización (Centrado y Escalado) $scale()$



Midiendo Distancias

[La base de los análisis multivariados es hacerlo bien]



Matrices de Biodiversidad

Antes de ver las matrices de distancias, veamos una matriz de biodiversidad (abundancias, presencia-ausencia):

```
library(vegan)
data("mite")
View(mite)
```

Brachy	PHTH	HPAV	RARD	SSTR	Protopl	MEGR	MPRO	TVIE	HMIN	HMIN2	NPRA	TVEL	ONOV	SUCT	LCIL	Oribatl1	Ceratoz1	PWIL	Galumna1	Stgncrs2
17	5	5	3	2	1	4	2	2	1	4	1	17	4	9	50	3	1	1	8	0
2	7	16	0	6	0	4	2	0	0	1	3	21	27	12	138	6	0	1	3	9
4	3	1	1	2	0	3	0	0	0	6	3	20	17	10	89	3	0	2	1	8
23	7	10	2	2	0	4	0	1	2	10	0	18	47	17	108	10	1	0	1	2
5	8	13	9	0	13	0	0	0	3	14	3	32	43	27	5	1	0	5	2	1
19	7	5	9	3	2	3	0	0	20	16	2	13	38	39	3	5	0	1	1	8
17	3	8	2	3	0	3	0	0	19	3	0	22	27	37	0	4	0	0	1	3
5	4	8	2	1	2	3	0	0	1	4	10	12	25	26	0	6	0	3	1	3



Matrices de Distancias

Los métodos multivariados buscan generar agrupaciones o clasificaciones de las observaciones (filas), basándose en sus distancias.

```
matriz <- vegan:::vegdist(mite, method = "bray")
```

N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	0.000	0.466	0.362	0.408	0.622	0.564	0.536	0.571	0.354	0.516	0.567	0.603	0.691	0.619	0.350	0.527	0.574	0.535	0.519	0.642	0.634	0.686	0.484	0.564
2	0.466	0.000	0.256	0.245	0.572	0.593	0.581	0.563	0.514	0.604	0.591	0.622	0.663	0.609	0.533	0.671	0.693	0.611	0.642	0.580	0.662	0.673	0.742	0.655
3	0.362	0.256	0.000	0.271	0.564	0.544	0.546	0.526	0.411	0.580	0.587	0.649	0.686	0.631	0.448	0.597	0.594	0.605	0.599	0.589	0.624	0.670	0.708	0.556
4	0.408	0.245	0.271	0.000	0.443	0.426	0.473	0.524	0.511	0.429	0.522	0.511	0.551	0.416	0.534	0.645	0.654	0.480	0.574	0.485	0.613	0.563	0.624	0.623
5	0.622	0.572	0.564	0.443	0.000	0.299	0.380	0.354	0.547	0.353	0.393	0.359	0.420	0.389	0.599	0.480	0.453	0.445	0.521	0.389	0.456	0.418	0.636	0.491
6	0.564	0.593	0.544	0.426	0.299	0.000	0.197	0.355	0.578	0.243	0.468	0.303	0.373	0.331	0.650	0.497	0.485	0.327	0.437	0.440	0.344	0.339	0.559	0.529
7	0.536	0.581	0.546	0.473	0.380	0.197	0.000	0.285	0.509	0.207	0.450	0.259	0.428	0.369	0.573	0.375	0.333	0.279	0.371	0.405	0.240	0.305	0.514	0.380
8	0.571	0.563	0.526	0.524	0.354	0.355	0.285	0.000	0.406	0.274	0.474	0.434	0.413	0.458	0.531	0.336	0.380	0.344	0.426	0.406	0.309	0.456	0.517	0.398



Funciones de distancias en R

```
# También se pueden calcular con la función base dist().  
# El Método puede ser uno de los siguiente:  
# "euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski".  
dist(DF, method="Método")
```

```
library(vegan)  
  
# El Método puede ser uno de los siguiente:  
# "manhattan", "euclidean", "canberra", "clark", "bray", "kulczynski", "jaccard",  
# "gower", "altGower", "morisita", "horn", "mountford", "raup", "binomial", "chao",  
# "cao", "mahalanobis", "chisq" "chord".  
vegdist(DF, method="Método")
```



Distancias Euclídeas

Basado en el teorema de Pitágoras

$$A^2 = B^2 + C^2$$

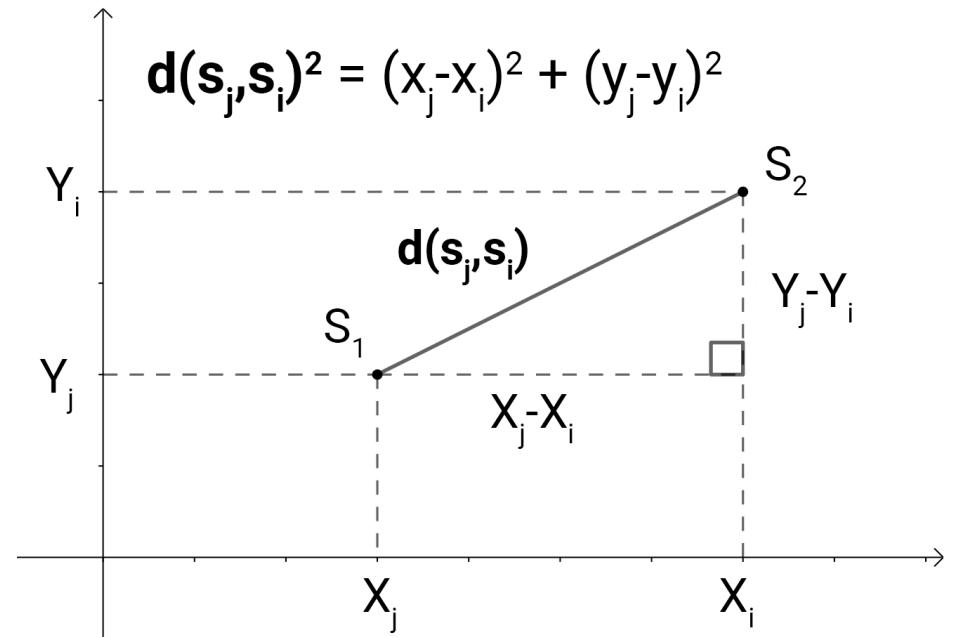
Lo que para efectos de la imagen, la distancia entre las observaciones s_j y s_i , basada en las variables X e Y es:

$$d_{s_j, s_i} = (X_j - X_i)^2 + (Y_j - Y_i)^2$$

$$d_{s_j, s_i} = \sqrt{(X_j - X_i)^2 + (Y_j - Y_i)^2}$$

Para un espacio N-dimensional de N variables:

$$d_{s_j, s_i} = \sqrt{\sum_{n=1}^N (X_{jn} - X_{in})^2}$$



En un espacio N-dimensional d_{s_j, s_i} es la distancia Euclídea.



Si consideramos las filas 20 y 21 de la tabla presentada inicialmente, tendremos:

Sitio	A	B	C	D	E	Profundidad	Polucion	Temp	Sedimento
s20	0	10	14	9	0	73	5.6	3.0	S
s21	8	0	0	4	6	59	4.3	3.4	C

La distancia Euclídea, considerando tres variables: Profundidad, Polución, y Temperatura,

$$d_{s_j, s_i} = \sqrt{\sum_{n=1}^N (X_{jn} - X_{in})^2}$$

$$d_{s_{20}, s_{21}} = \sqrt{(73 - 59)^2 + (5.9 - 4.3)^2 + (3.0 - 3.4)^2}$$

$$d_{s_{20}, s_{21}} = \sqrt{196 + 2.56 + 0.16}$$

$$d_{s_{20}, s_{21}} = \sqrt{198.72}$$

$$d_{s_{20}, s_{21}} = 14.09$$

Este resultado representa el problema de que las variables tengan diferente escala de medición. **Solución:** transformar las variables de análisis.



Pierre Legendre · Eugene D. Gallagher

Ecologically meaningful transformations for ordination of species data

Received: 25 September 2000 / Accepted: 17 March 2001 / Published online: 11 July 2001
© Springer-Verlag 2001

Abstract This paper examines how to obtain species biplots in unconstrained or constrained ordination without resorting to the Euclidean distance [used in principal-component analysis (PCA) and redundancy analysis (RDA)] or the chi-square distance [preserved in correspondence analysis (CA) and canonical correspondence analysis (CCA)] which are not always appropriate for the

strained or constrained ordinations of species abundance data tables and the corresponding biplots or triplots which are extremely useful for ecological interpretation (Fig. 1a, c). Empirical work during the 1970s established that CA was appropriate for such data, while ter Braak (1985) showed that the chi-square distance preserved in CA provided a good approximation for species with uni-

Legendre, P., & Gallagher, E. D. (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129(2), 271-280.



Table 1 Species abundance paradox, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 1 and 2, which have no species in common, is smaller than that between sites 1 and 3 which share species 2 and 3. This example shows that the Euclidean distance is not appropriate for species

community composition data containing zeros. With the other coefficients used here, the distance between sites 1 and 2 is larger than between sites 1 and 3, and the distance between sites 1 and 2 is the same as between sites 2 and 3, or very nearly so

		Species 1	Species 2	Species 3
Species abundance paradox data (three sites, three species)	Site 1	0	1	1
	Site 2	1	0	0
	Site 3	0	4	8
<i>Distance function</i>		$D(\text{site 1, site 2})$	$D(\text{site 1, site 3})$	$D(\text{site 2, site 3})$
$D_{\text{Euclidean}}$		1.7321	7.6158	9.0000
D_{chord}		1.4142	0.3204	1.4142
$D\chi^2_{\text{metric}}$		1.0382	0.0930	1.0352
$D\chi^2_{\text{distance}}$		4.0208	0.3600	4.0092
$D_{\text{species profiles}}$		1.2247	0.2357	1.2472
$D_{\text{Hellinger}}$		1.4142	0.1697	1.4142



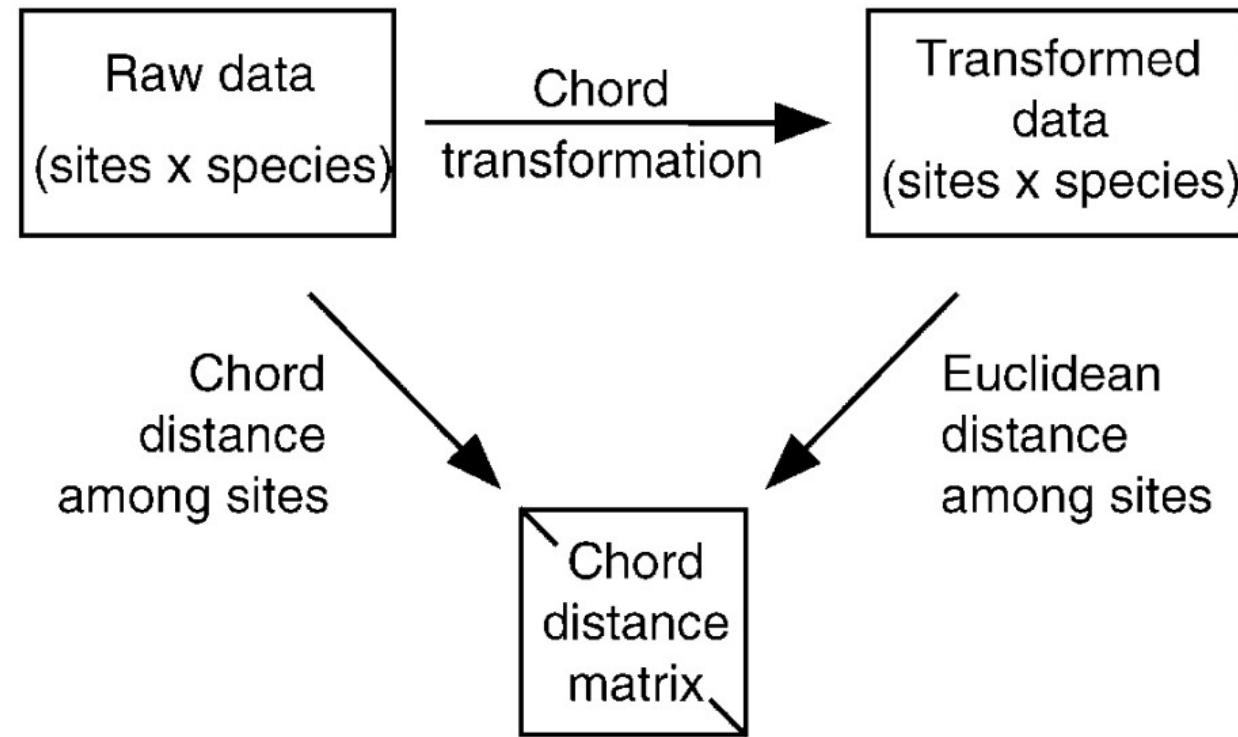
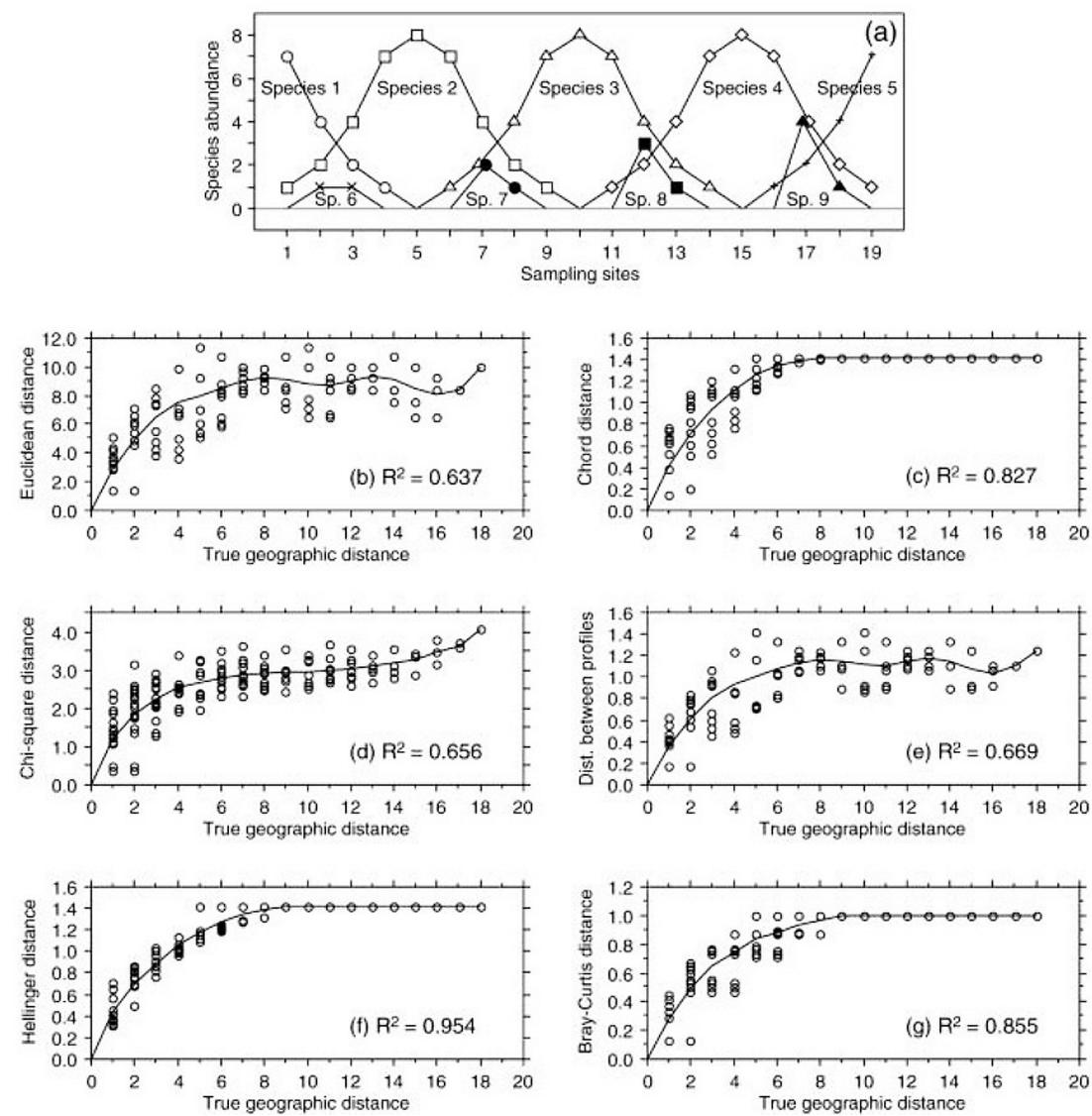


Fig. 2 Illustration of the role of the data transformations as a way of obtaining a given distance function. The example uses the chord distance

Fig. 3a–g Analysis of artificial gradient data. **a** The gradient comprises 19 sites (numbers along abscissa) and nine species (different symbols).

b–g Diastemograms comparing true geographic distances (abscissa) to the computed ecological distances among sites (ordinate). The construction and interpretation of these graphs is described in the text



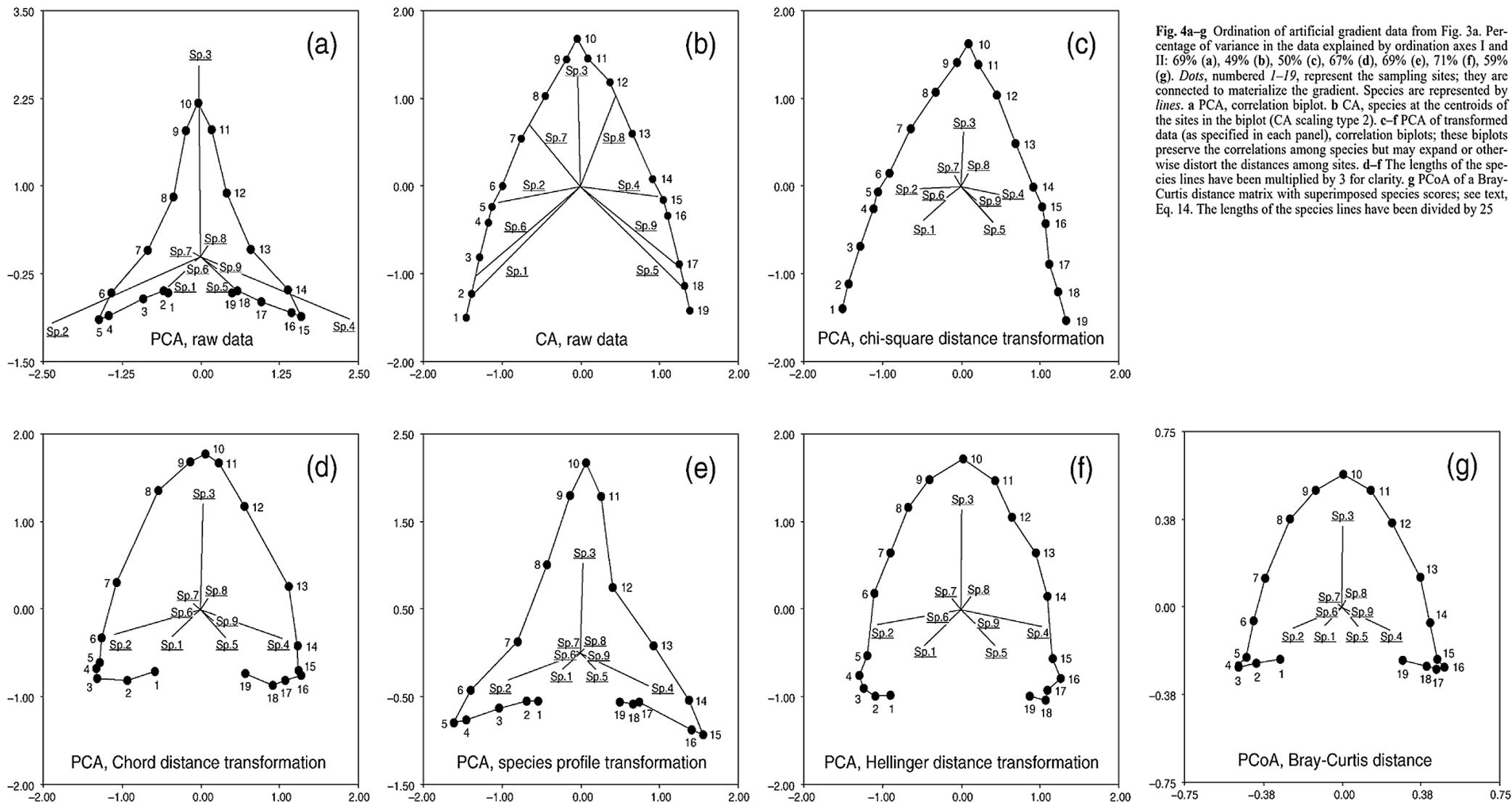


Fig. 4a-g Ordination of artificial gradient data from Fig. 3a. Percentage of variance in the data explained by ordination axes I and II: 69% (a), 49% (b), 50% (c), 67% (d), 69% (e), 71% (f), 59% (g). Dots, numbered 1–19, represent the sampling sites; they are connected to materialize the gradient. Species are represented by lines. a PCA, correlation biplot. b CA, species at the centroids of the sites in the biplot (CA scaling type 2). c–f PCA of transformed data (as specified in each panel), correlation biplots; these biplots preserve the correlations among species but may expand or otherwise distort the distances among sites. d–f The lengths of the species lines have been multiplied by 3 for clarity. g PCoA of a Bray-Curtis distance matrix with superimposed species scores; see text, Eq. 14. The lengths of the species lines have been divided by 25

Medidas de distancias y transformaciones para:

Var. ambientales

Abundancias-PCA

Abundancias-CA

Abundancias-MDS

Binarios-NMDS

Si se desea aplicar un PCA:

1) **Distancia Euclideana, Manhattan, Minkowski.**

- Sobre matriz de datos continuos por naturaleza (mediciones de variables ambientales).
- Es necesario estandarizar (centrar y escalar) las variables para evitar el error por la escala de medición de las variables.



Beta diversity as the variance of community data: dissimilarity coefficients and partitioning

Pierre Legendre^{1*} and Miquel De Cáceres^{2,3}

Abstract

Beta diversity can be measured in different ways. Among these, the total variance of the community data table \mathbf{Y} can be used as an estimate of beta diversity. We show how the total variance of \mathbf{Y} can be calculated either directly or through a dissimilarity matrix obtained using any dissimilarity index deemed appropriate for pairwise comparisons of community composition data. We addressed the question of which index to use by coding 16 indices using 14 properties that are necessary for beta assessment, comparability among data sets, sampling issues and ordination. Our comparison analysis classified the coefficients under study into five types, three of which are appropriate for beta diversity assessment. Our approach links the concept of beta diversity with the analysis of community data by commonly used methods like ordination and ANOVA. Total beta can be partitioned into Species Contributions (SCBD: degree of variation of individual species across the study area) and Local Contributions (LCBD: comparative indicators of the ecological

Legendre, P., & De Cáceres, M. (2013). Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology letters*, 16(8), 951–963.



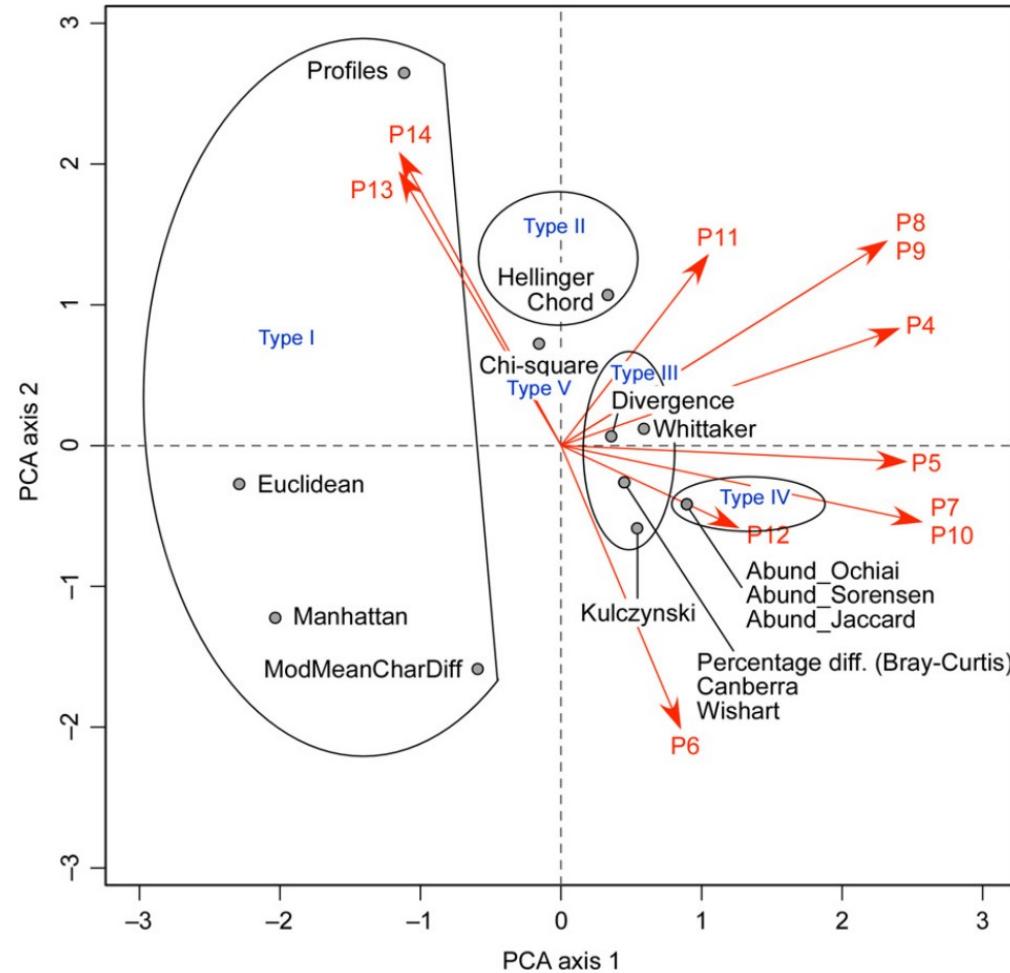


Figure 2 Principal component biplot relating properties P4–P14 (red arrows) to the dissimilarity coefficients.

Tipo I:

- **Euclidean, Manhattan, Modified mean character difference, Species Profile.** son inapropiados para matrices de diversidad (abundancias) no transformadas.

Tipo II:

- Distancias de **Hellinger y Chord** son adecuadas para matrices de diversidad (abundancias) transformadas con Hellinger y Chord respectivamente.

Tipo III:

- Distancias **Canberra, Whittaker, Divergence, Bray-Curtis, Kulczyński**, son adecuadas para matrices de diversidad (abundancias)...

...Si hay valores muy altos entre las abundancias, es mejor logaritmizarlas para evitar darle peso excesivo.

Tipo IV:

- Distancias **cuantitativas** basadas en abundancias: **Jaccard, Sørensen, Ochiai**.

Tipo V:

- Distancia de **Chi-cuadrado** es inapropiado para matrices de diversidad (abundancias) cuando hay presencia de especies raras, en caso se desee realizar un CCA (análisis de correspondencia canónica). Para CA, no influencia de manera importante esta propiedad.



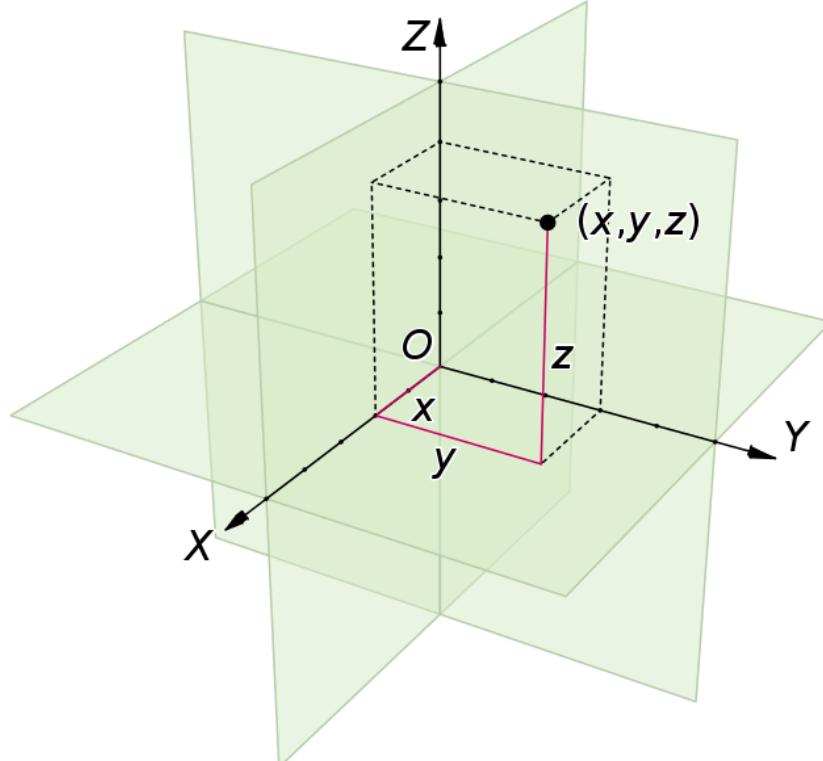
Espacio Euclídeo

1. En el espacio euclidiano:

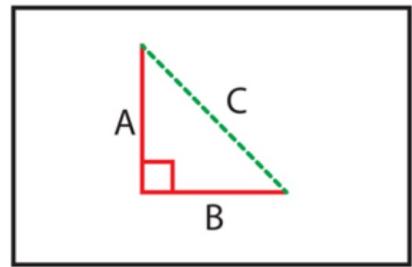
- Sistema de coordenadas cartesiano que respeta y satisface los axiomas de Euclides.
- La distancia entre dos puntos se mide mediante una generalización del teorema de Pitágoras.

2. En el espacio no euclidiano:

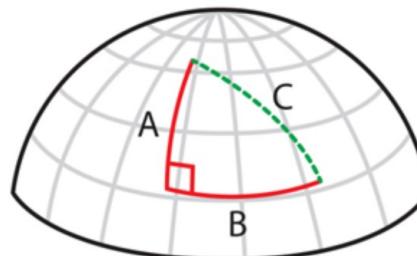
- La distancia medida en estos espacios se mide mediante geometría esférica, hiperbólica, entre otras.
- Estas distancias no son comparables con lo observado en un espacio euclídeo.



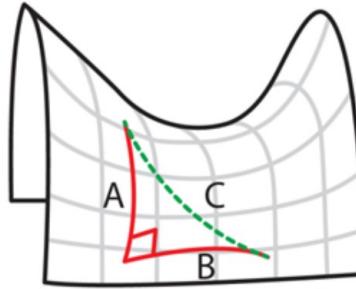
Espacio Euclídeo y No Euclídeo



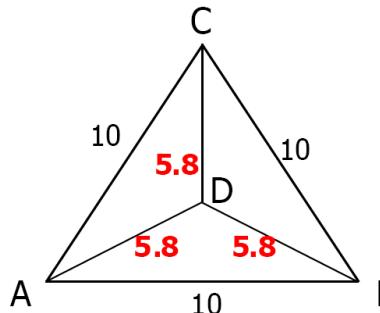
Euclidean
 $a^2 + b^2 = c^2$



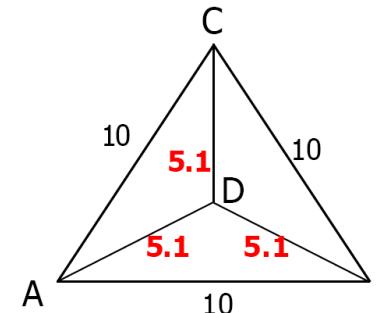
Spherical
 $\cos a \times \cos b = \cos c$



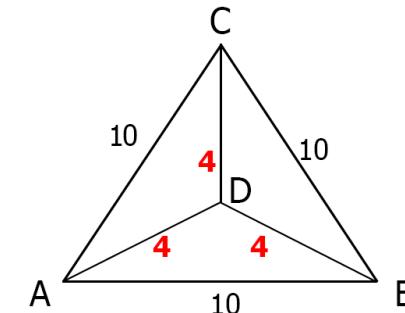
Hyperbolic
 $\cosh a \times \cosh b = \cosh c$



Euclidean
Metric



Non-Euclidean
Metric



Non-Euclidean
Non-Metric

[Link Fig 1](#) | [Link Fig 2](#)

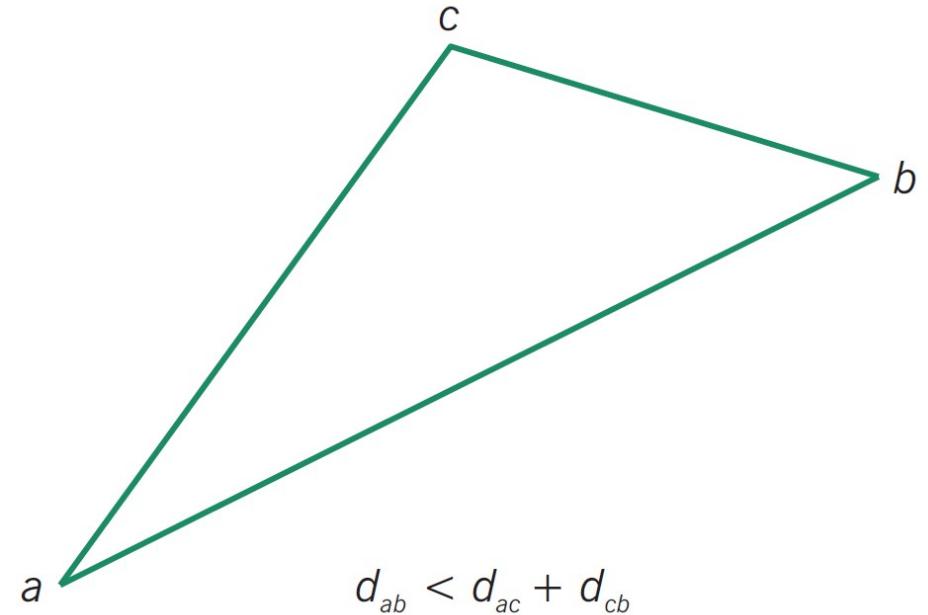


Métrico y No Métrico

Un concepto más general que "Euclídeo y no Euclídeo", es el concepto de distancias métricas y no. Las distancias Euclidianas son métricas.

Los axiomas matemáticos de los espacios métricos son (d_{ab} es la distancia de "a" a "b"):

- $d_{ab} = d_{ba}$
- $d_{ab} \geq 0$ y es solamente 0 si $a = b$
- $d_{ab} \leq d_{ac} + d_{cb}$ (Triangle inequality)



Desarrollemos las secciones:

1) Métodos de Transformación

2) Métodos de Distancias

20 : 00



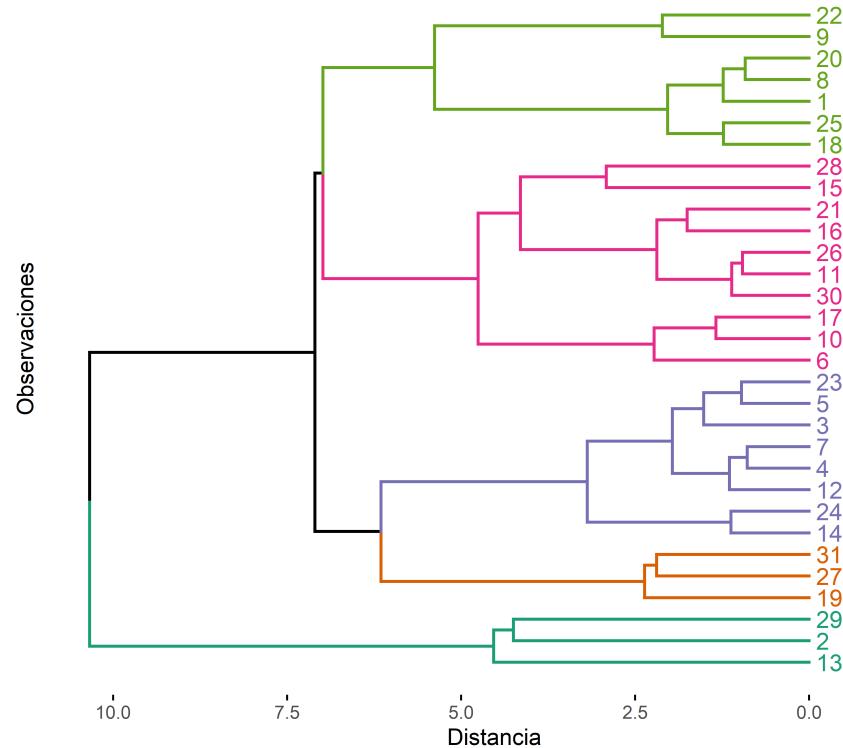
Métodos de Ordenamiento sin Restricciones

[Clustering, CA, PCA, MDS, NMDS]

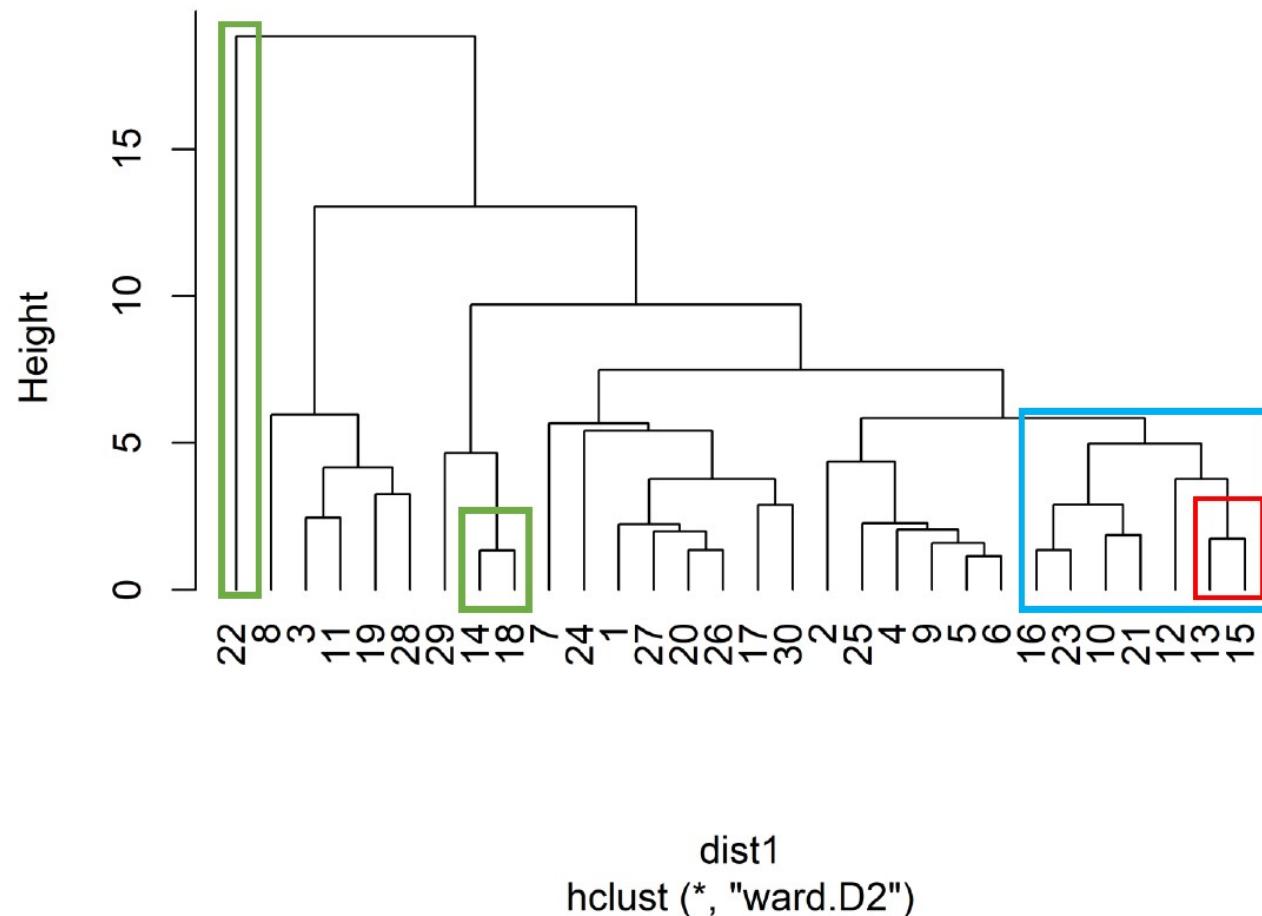


Agrupamiento Jerárquico (Hierarchical Clustering)

- Una aplicación de las matrices de distancias es crear árboles de agrupamientos.
- Este no es un método estadístico (no hay contraste de hipótesis).
- Es una herramienta o técnica para identificar grupos o estructuras en los datos.



Cluster Dendrogram



Las observaciones (filas en la table original) más similares se encuentran juntas.

Estas se encuentran inmersas en agrupaciones más grandes

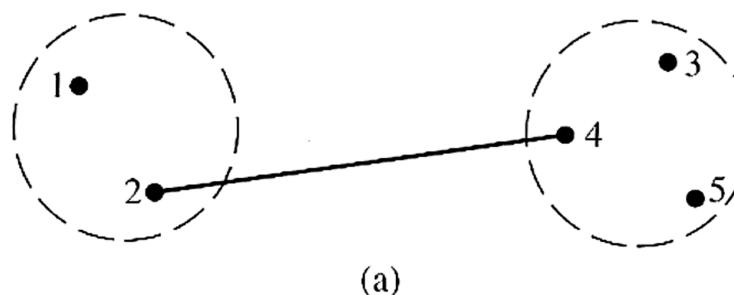
A mayor distancia entre dos observaciones en el gráfico, más diferentes son.



Métodos de Aglomeración

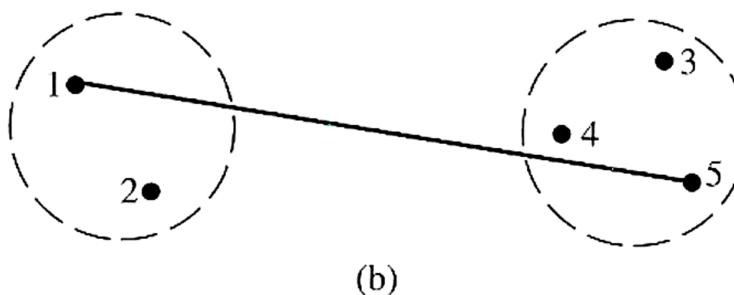
- Single linkage (a)
- Complete linkage (b)
- Average linkage o UPGMA (c)
- Método de Ward (Least Squares)

[Link Figura](#)

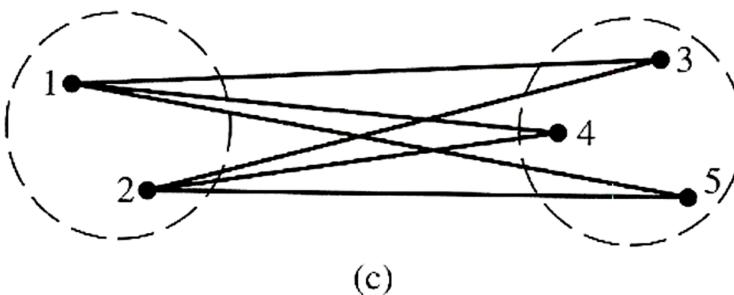


Cluster distance

$$d_{24}$$



$$d_{15}$$



$$\frac{d_{13} + d_{14} + d_{15} + d_{23} + d_{24} + d_{25}}{6}$$



Desarrollemos la sección:

3) Agrupamiento Jerárquico

60 : 00



Análisis de Correspondencia (CA)

- Es una extensión del PCA adecuado para relacionar los casos de dos variables categóricas (matriz de contingencia).
- Genera una decomposición de eigenvectores igual que en el PCA proyectando el resultado en un plano métrico euclidiano.
- La diferencia es que la distancia generada es la distancia de Chi-Cuadrado en lugar de la Euclíadiana.



Desarrollemos la sección:

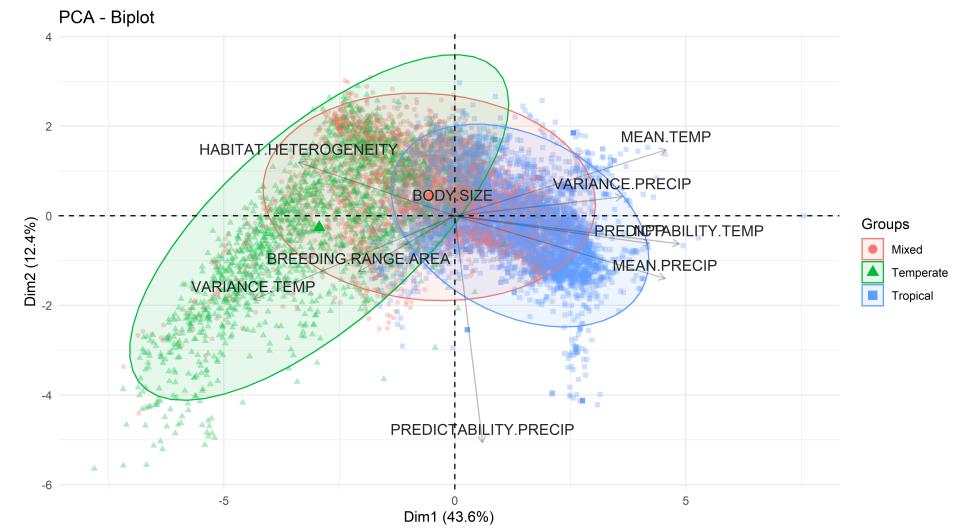
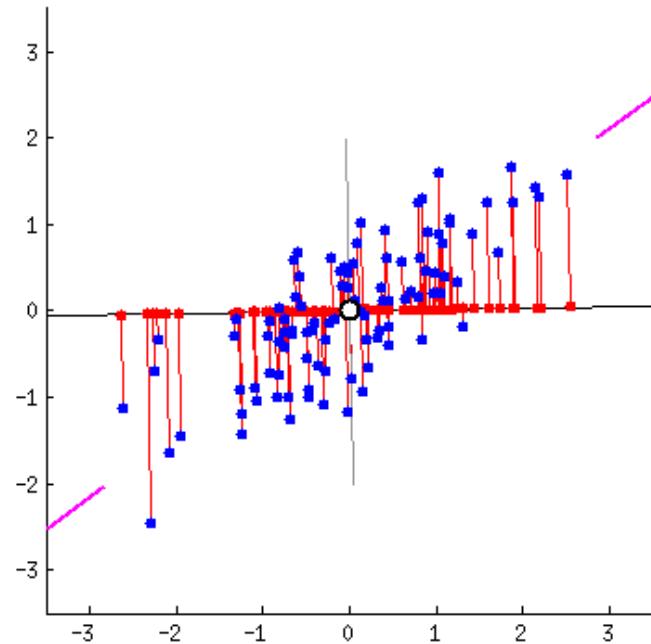
4) Análisis de Correspondencia (CA)

25 : 00



Análisis de Componentes Principales (PCA)

- El PCA es un método de reducción de dimensionalidad.
- Transforma un gran número de variables a un menor conjunto de variables no correlacionadas (ortogonales) llamadas Componentes Principales (PC).



Análisis de Componentes Principales (PCA)

- El PCA es un método de reducción de dimensionalidad.
- Transforma un gran número de variables a un menor conjunto de variables no correlacionadas (ortogonales) llamadas Componentes Principales (PC).

1. Estandarizar la base de datos

($\mu = 0, \sigma^2 = 1$).

```
scale(datos)
```

2. Cálculo de las matrices de covarianzas.
3. Cálculo de los eigenvectores y eigenvalores.
4. Reordenamiento de los eigenvalores y eigenvectores
5. Reproyectar los datos en los ejes principales.



Análisis de Componentes Principales (PCA)

- El PCA es un método de reducción de dimensionalidad.
- Transforma un gran número de variables a un menor conjunto de variables no correlacionadas (ortogonales) llamadas Componentes Principales (PC).

1. Estandarizar la base de datos ($\mu = 0, \sigma^2 = 1$).
2. **Cálculo de las matrices de covarianzas.**
3. Cálculo de los eigenvectores y eigenvalores.
4. Reordenamiento de los eigenvalores y eigenvectores
5. Reproyectar los datos en los ejes principales.

$$\text{corr}(\mathbf{X}) = \begin{bmatrix} 1 & \frac{\mathbb{E}[(X_1 - \mu_1)(X_2 - \mu_2)]}{\sigma(X_1)\sigma(X_2)} & \dots & \frac{\mathbb{E}[(X_1 - \mu_1)(X_n - \mu_n)]}{\sigma(X_1)\sigma(X_n)} \\ \frac{\mathbb{E}[(X_2 - \mu_2)(X_1 - \mu_1)]}{\sigma(X_2)\sigma(X_1)} & 1 & \dots & \frac{\mathbb{E}[(X_2 - \mu_2)(X_n - \mu_n)]}{\sigma(X_2)\sigma(X_n)} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\mathbb{E}[(X_n - \mu_n)(X_1 - \mu_1)]}{\sigma(X_n)\sigma(X_1)} & \frac{\mathbb{E}[(X_n - \mu_n)(X_2 - \mu_2)]}{\sigma(X_n)\sigma(X_2)} & \dots & 1 \end{bmatrix}.$$

$$\begin{array}{ccc} & x & y & z \\ x & \left[\begin{array}{cc} \text{var}(x) & \text{cov}(x, y) \\ \text{cov}(x, y) & \text{var}(y) \end{array} \right] & \left[\begin{array}{ccc} \text{var}(x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(x, y) & \text{var}(y) & \text{cov}(y, z) \\ \text{cov}(x, z) & \text{cov}(y, z) & \text{var}(z) \end{array} \right] \\ y & & \\ z & & \end{array}$$



Análisis de Componentes Principales (PCA)

- El PCA es un método de reducción de dimensionalidad.
- Transforma un gran número de variables a un menor conjunto de variables no correlacionadas (ortogonales) llamadas Componentes Principales (PC).

1. Estandarizar la base de datos
 $(\mu = 0, \sigma^2 = 1)$.

```
eigen(cov(scale(DF)))
```

2. Cálculo de las matrices de covarianzas.
3. **Cálculo de los eigenvectores y eigenvalores.**
4. Reordenamiento de los eigenvalores y eigenvectores
5. Reproyectar los datos en los ejes principales.

```
## eigen() decomposition
## $values
## [1] 1.4237444 0.9917002 0.5845554
##
## $vectors
## [,1]      [,2]      [,3]
## [1,] -0.1396642  0.99004058 0.01770777
## [2,] -0.6993188 -0.11128145 0.70609465
## [3,]  0.7010329  0.08623276 0.70789603
```



Análisis de Componentes Principales (PCA)

El PCA es un método de reducción de dimensionalidad que se utiliza a menudo para reducir la dimensionalidad de grandes conjuntos de datos, transformando un gran conjunto de variables en uno pequeño que aún contiene la mayor parte de la información del conjunto grande.

1. Estandarizar la base de datos ($\mu = 0, \sigma^2 = 1$).
2. Cálculo de las matrices de covarianzas.
3. Cálculo de los eigenvectores y eigenvalores.
4. **Reordenamiento de los eigenvalores y eigenvectores**
5. Reproyectar los datos en los ejes principales.

pca

```
## Standard deviations (1, ..., p=3):  
## [1] 1.1932076 0.9958414 0.7645623  
##  
## Rotation (n x k) = (3 x 3):  
##          PC1        PC2        PC3  
## Aspect  0.1396642  0.99004058  0.01770777  
## Slope   0.6993188 -0.11128145  0.70609465  
## Snow    -0.7010329  0.08623276  0.70789603
```



Desarrollemos la sección:

5) Análisis de Componentes Principales (PCA)

60 : 00



Análisis de Coordenadas Principales (MDS)

- Conocida como Escalamiento Multidimensional Métrico.
- PCA conserva distancias euclidianas entre las muestras y CA las distancias chi-cuadrado, PCoA proporciona una representación euclidiana de un conjunto de objetos cuya relación se mide mediante cualquier método de distancias.
- Tanto PCA y CA y PCoA devuelven un conjunto de ejes ortogonales.

Eigenvalores negativos

En caso de utilizar un índice de distancia que no sea métrico, el PCoA puede producir ejes con valores propios negativos que no se pueden trazar. Soluciones:

1. Convertir la distancia no métrica a métrica mediante transformación (**BC + sqrt = métrica**).
2. Utilizar correcciones para evitar eigenvectores no negativos (**e.g. lingoes**).
3. PCoA no proyecta las columnas de la tabla (especies), solo observaciones, pero podemos aplicar código para extraer esa información en R.



Desarrollemos la sección:

6) Análisis de Coordenadas Principales (PCoA, MDS)

60 : 00



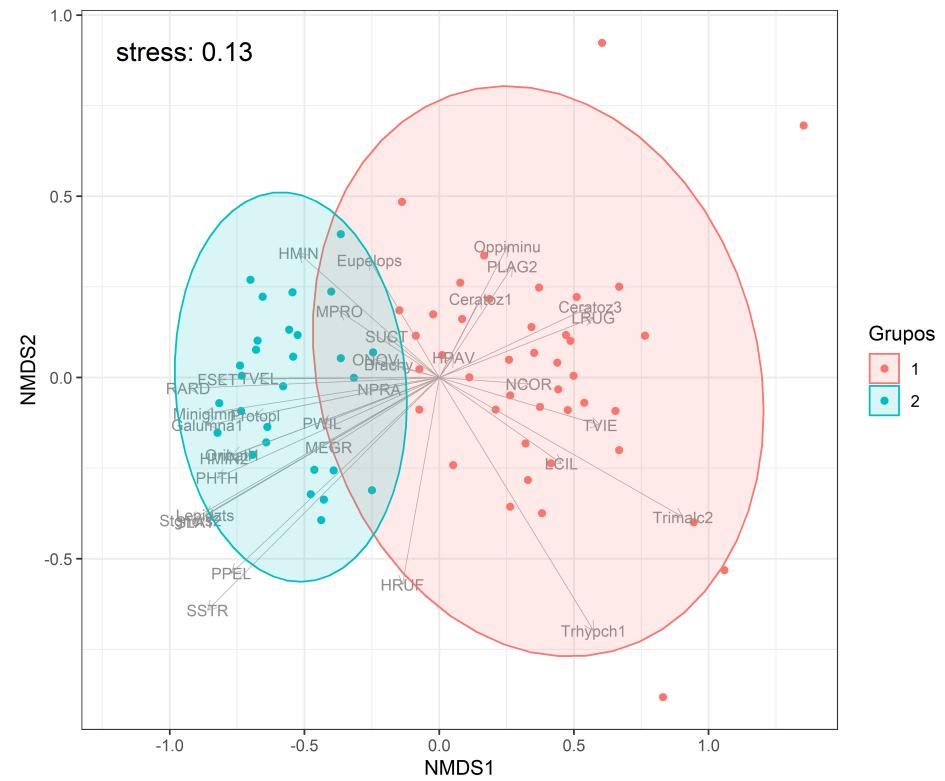
Escalamiento Multidimensional No Métrico (NMDS)

- Es muy útil cuando queremos **representar en 2D las relaciones entre observaciones**, siendo que obtuvimos un PCA o PCoA que nos indicaban que 3 a más dimensiones eran necesarias para capturar un gran porcentaje de la variabilidad (>50%).
- Es **más robusto que el PCoA** cuando se tiene que representar datos que no sean adecuados para un PCA (principalmente datos de conteos)
- Utiliza un algoritmo iterativo aleatorio, no paramétrico, para hallar con “prueba y error” la mejor posición para los puntos (la solución final). **No se basa en descomposición de los eigenvalores como el PCA.**
- No está restringido a un tipo de matriz de distancia, **acepta cualquier matriz de distancia.**



Estrés del NMDS

- Como regla general, el valor del estrés del NMDS deberá ser interpretado como
 - ≤ 0.05 = ideal
 - <0.1 y >0.05 = bueno
 - >0.1 y <0.2 = muy justo
 - >0.2 y <0.3 = sospechoso
 - >0.3 = ordenamiento arbitrario.



Desarrollemos la sección:

7) Escalamiento Multidimensional No Métrico (NMDS)

60 : 00



Gracias por tu atención

