Clase 21 - Componentes principales y PERMANOVA

Diplomado en Análisis de Datos con R e Investigación reproducible para Biociencias.

Dr. José Gallardo Matus | https://genomics.pucv.cl/

Pontificia Universidad Católica de Valparaíso

19 November 2022

PLAN DE LA CLASE

1.- Introducción

- ¿Qué son los análisis de componentes principales?
- Estudio de caso 1: Detectar fraude alimentario en salmón.
- Estudio de caso 2: Cacho de cabra v/s ají chileno negro.
- Estudio de caso 3: Contaminación ambiental Humedal Yali.
- Etapas para realizar un ACP.
- Varianza explicada.
- Graficas biplot.
- PERMANOVA
- Estudio de caso 4: Tortugas y contaminación por plástico.

2). Práctica con R y Rstudio cloud.

- Elaborar análisis de componentes principales con R.

ANÁLISIS DE COMPONENTES PRINCIPALES

- **¿Qué son los análisis de componentes principales?**

a) Son una herramienta estadística multivariada que se utiliza para realizar análisis exploratorio de datos y para construir modelos predictivos.

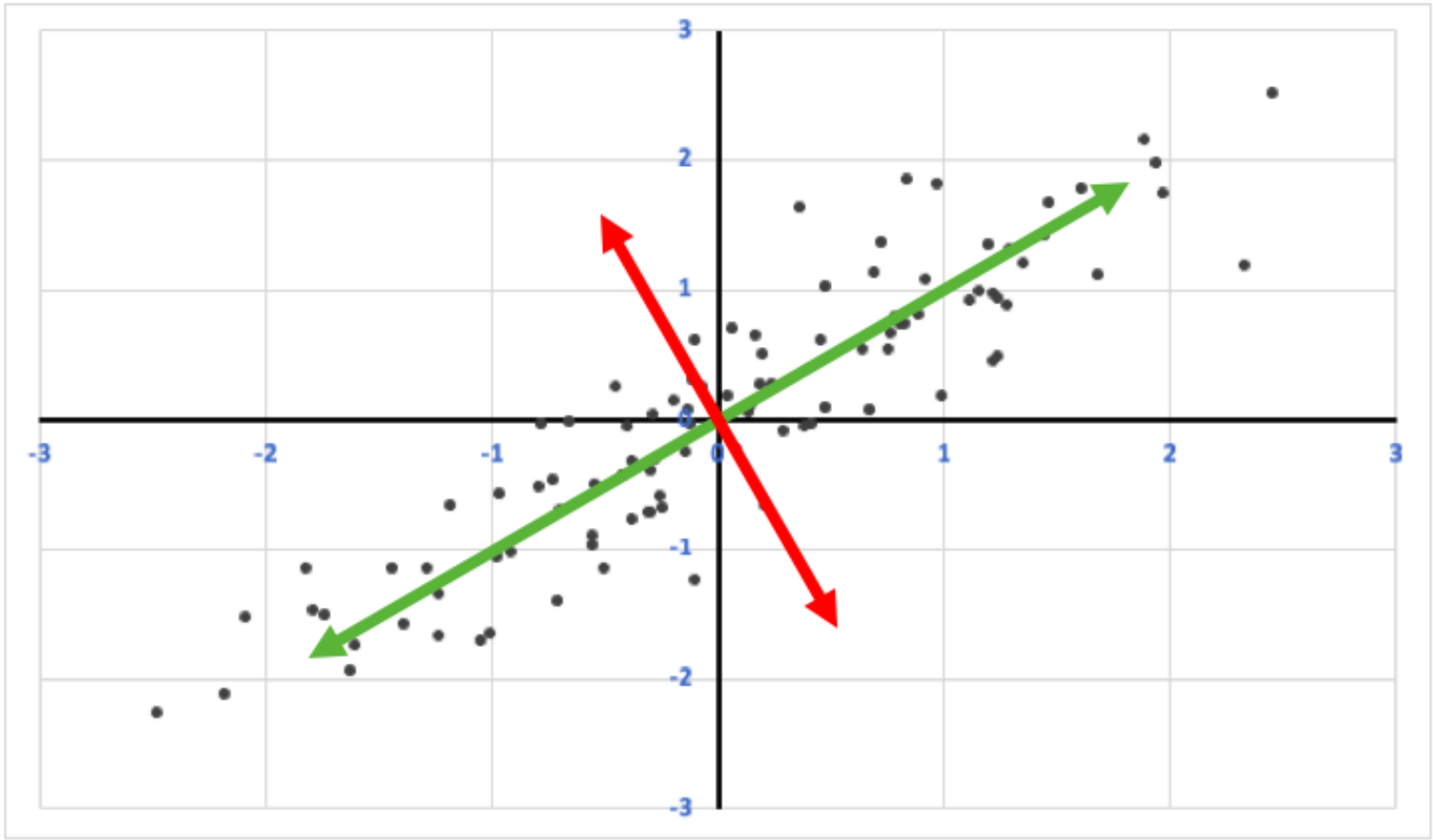
b) Permite reducir la dimensionalidad de un set de datos con muchas variables respuesta, sin perder mucha información.

c) Permite encontrar patrones en un set de datos mediante el calculo de los "componentes principales".

COMPONENTES PRINCIPALES

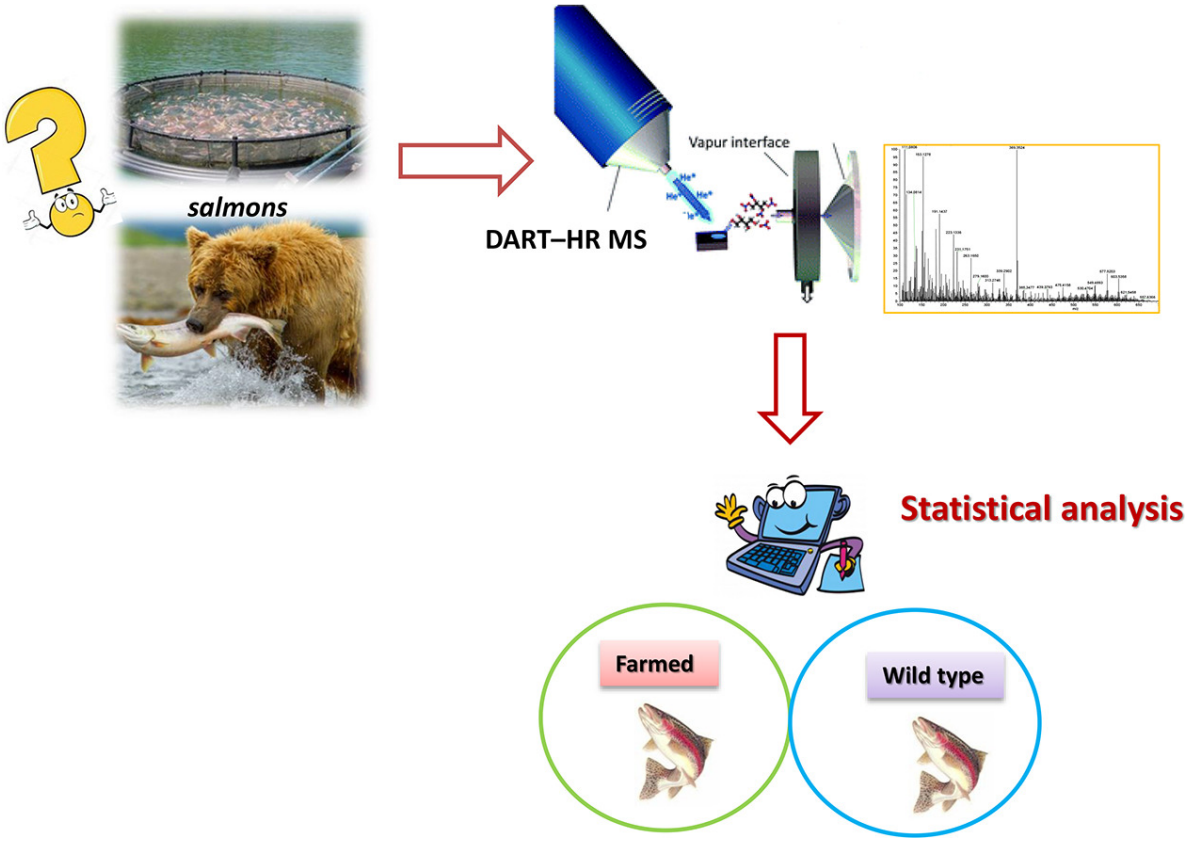
- **¿Qué son los componentes principales?**

a) Combinación lineal de las variables originales no correlacionadas entre si (perpendiculares / ortogonales).



CASO 1: FRAUDE ALIMENTARIO SALMÓN.

- ¿Cómo distinguir filetes de salmón silvestre y de cultivo?



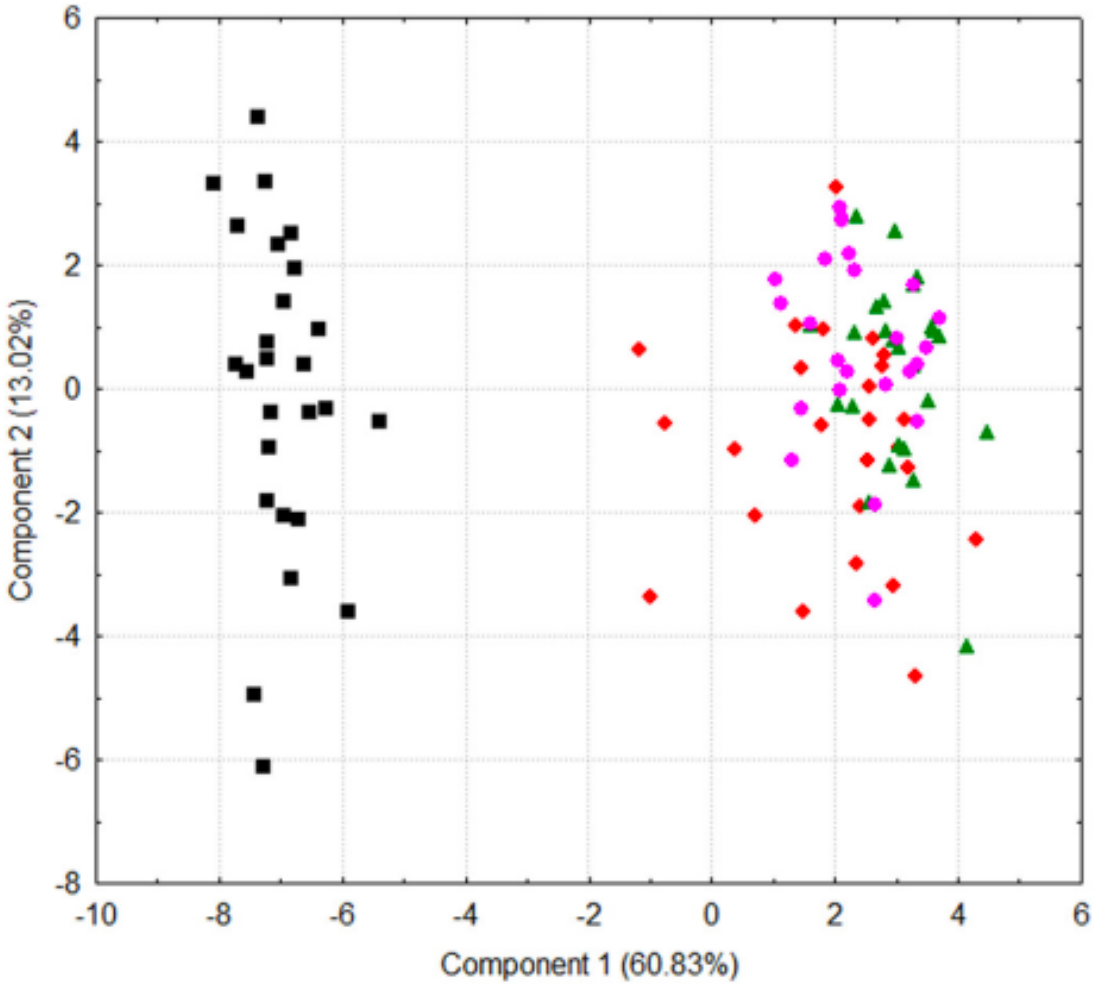
Fiorino et al. 2019

ANÁLISIS UNIVARIADO ACIDOS GRASOS

Table 1
Summary of MS-related data, chemical formulas and possible chain compositions inferred for the 30 fatty acids that were considered for the discrimination between wild-type and farmed salmons during the present study. In the last two columns average values and standard deviations referred to normalized abundances observed for each fatty acid in the 26 wild-type and the 74 farmed salmons are reported.

#	Average Experimental m/z	Theoretical m/z	Chemical formula	Mass accuracy (ppm)	Chain composition(s)	Wild type	Farmed
1	157.0862	157.0870	$C_8H_{13}O_3$	-5.09	Hydroxy-8:1 Oxo-8:0	0.45 ± 0.16	0.35 ± 0.11
2	171.1018	171.1027	$C_9H_{15}O_3$	-5.26	Hydroxy-9:1 Oxo-9:0	1.10 ± 0.30	1.67 ± 0.37
3	181.0860	181.0870	$C_{10}H_{13}O_3$	-5.52	Oxo-10:2	0.50 ± 0.09	0.21 ± 0.07
4	211.1329	211.1340	$C_{12}H_{19}O_3$	-5.21	Hydroxy-12:2 Oxo-12:1	0.17 ± 0.03	0.44 ± 0.07
5	227.2008	227.2017	$C_{14}H_{27}O_2$	-3.96	14:0	4.09 ± 1.61	1.80 ± 0.91
6	253.2167	253.2173	$C_{16}H_{29}O_2$	-2.36	16:1	5.40 ± 0.82	3.97 ± 1.09
7	255.2323	255.2330	$C_{16}H_{31}O_2$	-2.74	16:0	25.37 ± 3.13	15.04 ± 2.67
8	269.2116	269.2122	$C_{16}H_{29}O_3$	-2.23	Hydroxy-16:1 Oxo-16:0	5.22 ± 0.88	3.04 ± 0.77
9	271.2271	271.2279	$C_{16}H_{31}O_3$	-2.95	Hydroxy-16:0	1.59 ± 0.32	0.75 ± 0.15
10	275.2006	275.2017	$C_{18}H_{27}O_2$	-3.99	18:4	1.86 ± 0.57	0.57 ± 0.13
11	277.2167	277.2173	$C_{18}H_{29}O_2$	-2.16	18:3	1.13 ± 0.17	3.76 ± 0.68
12	279.2322	279.2330	$C_{18}H_{31}O_2$	-2.86	18:2	2.08 ± 0.22	11.82 ± 1.39
13	281.2479	281.2486	$C_{18}H_{33}O_2$	-2.49	18:1	15.64 ± 2.10	27.85 ± 2.52
14	283.2633	283.2643	$C_{18}H_{35}O_2$	-3.18	18:0	3.86 ± 0.36	2.63 ± 0.39
15	285.2066	285.2071	$C_{16}H_{29}O_4$	-1.75	Hydroxy, oxo -16:0	1.17 ± 0.23	0.81 ± 0.30
16	287.2222	287.2228	$C_{16}H_{31}O_4$	-2.09	Dihydroxy-16:0	1.27 ± 0.15	1.87 ± 5.18
17	293.2114	293.2122	$C_{18}H_{29}O_3$	-2.73	Hydroxy-18:3 Epoxy-18:2 Oxo-18:2	0.46 ± 0.04	1.85 ± 0.30
18	295.2270	295.2279	$C_{18}H_{31}O_3$	-3.04	Hydroxy-18:2 Epoxy-18:1 Oxo-18:1	2.64 ± 0.34	4.82 ± 0.70

PCA SALMON

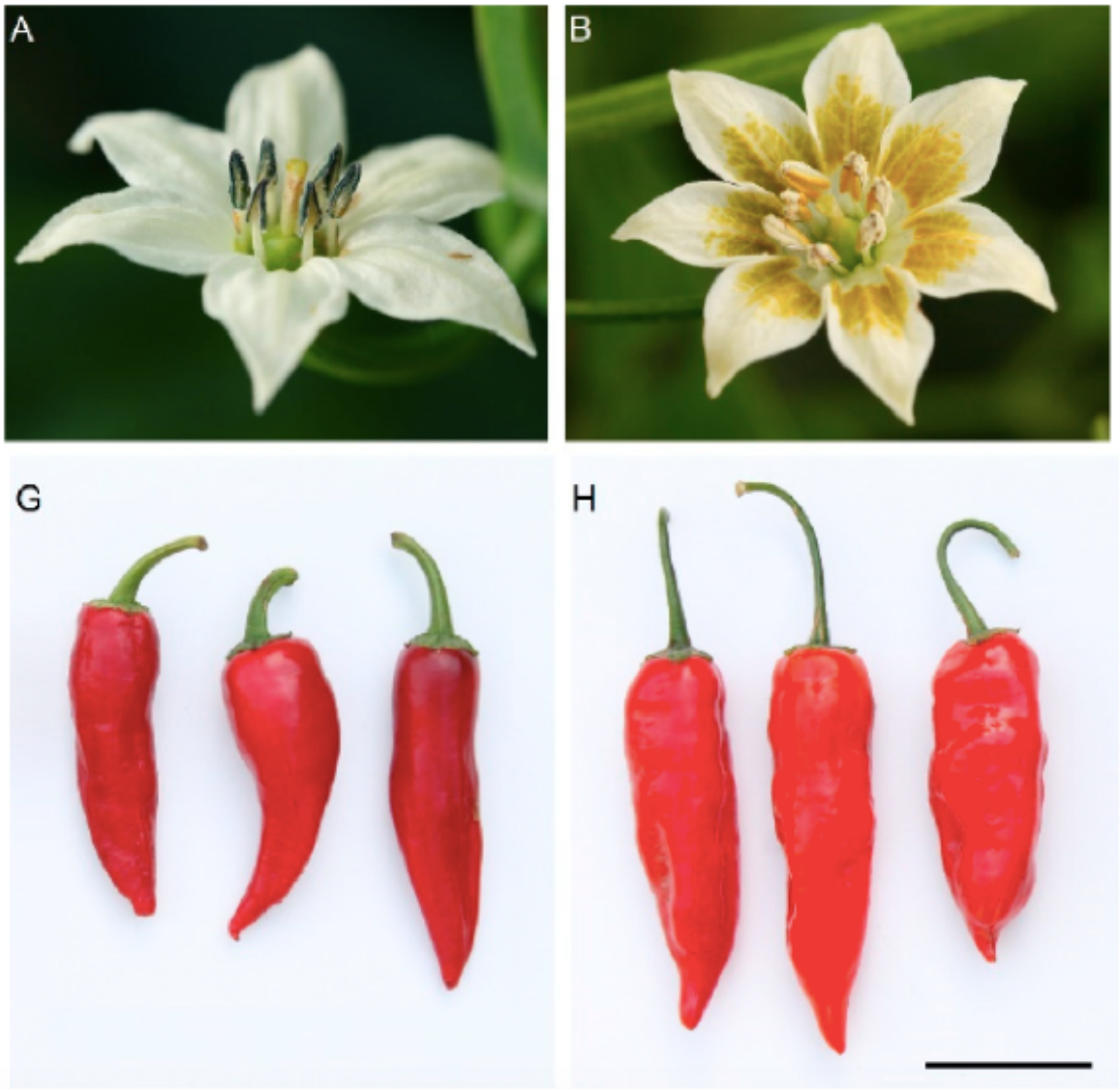


CP1 + CP2 explican 74 % de la varianza en ácidos grasos analizadas.

CP1 separa grupos silvestre y de cultivo, pero CP2 no.

El modelo puede ser usado para predecir fraude con base a los ácidos grasos.

CASO 2: AJI CACHO DE CABRA.



Comparación morfológica entre Cacho de cabra (izq.) v/s ají chileno negro (der.).

Muñoz-Concha et al. 2019

ANÁLISIS FUNCIONAL UNIVARIADO DE MORFOMETRÍA

Table 4. Average values for each measured variable in chili pepper fruits of different fields and landraces.

Variable	Landrace						Statistical Comparisons			
	cacho de cabra			*chileno negro*			between Landraces		among Fields	
	Average	SE	*n*	Average	SE	*n*	Test	*p* Value	Test	*p* Value
Fruit weight (g)	25.2	1.57	4	30.8	0.71	3	T	0.057	ANOVA	<0.001
Fruit length (cm)	11.5	0.80	5	11.8	0.31	3	T	0.841	ANOVA	<0.001
Fruit diameter (mm)	28.6	0.87	5	26.8	1.12	3	T	0.313	ANOVA	<0.001
Fruit volume (mL)	41.4	4.21	5	49.8	0.71	3	T	0.228	Kruskal–Wallis	<0.001
Pericarp thickness (mm)	2.2	0.07	4	2.3	0.06	3	T	0.584	Kruskal–Wallis	0.001
Petiole length (mm)	35.7	2.67	5	71.4	4.57	3	T	0.001	Kruskal–Wallis	<0.001
Petiole diameter (mm)	4.7	0.32	5	2.9	0.03	3	Mann–Whitney	0.025	Kruskal–Wallis	<0.001
Number of seeds per fruit	272.4	26.60	5	187.7	9.49	3	T	0.082	ANOVA	<0.001
Seed weight (g)	3.8	0.35	4	3.7	0.15	3	T	0.815	Kruskal–Wallis	<0.001
Fruit weight per seed (mg)	113.1	18.83	4	178.2	8.73	3	T	0.063	Kruskal–Wallis	<0.001
Fruit-to-petiole length ratio	3.4	0.11	5	1.7	0.16	3	T	<0.001	Kruskal–Wallis	<0.001
Color parameter L*	33.2	0.50	4	34.6	0.35	3	T	0.117	ANOVA	<0.001
Color parameter a*	32.1	0.64	4	30.5	0.18	3	T	0.134	ANOVA	<0.001
Color parameter b*	14.2	0.62	4	17.7	0.25	3	T	0.012	ANOVA	<0.001
Color parameter chroma	35.1	0.83	4	35.3	0.09	3	T	0.868	ANOVA	0.003
Color parameter hue	23.7	0.51	4	30.0	0.43	3	T	0.001	ANOVA	<0.001

ANÁLISIS ESTRUCTURAL MULTIVARIADO: CLUSTER Y PCA

- Note que los primeros 2 componentes principales explican un 69 % de la varianza explicada por las variables morfométricas analizadas.

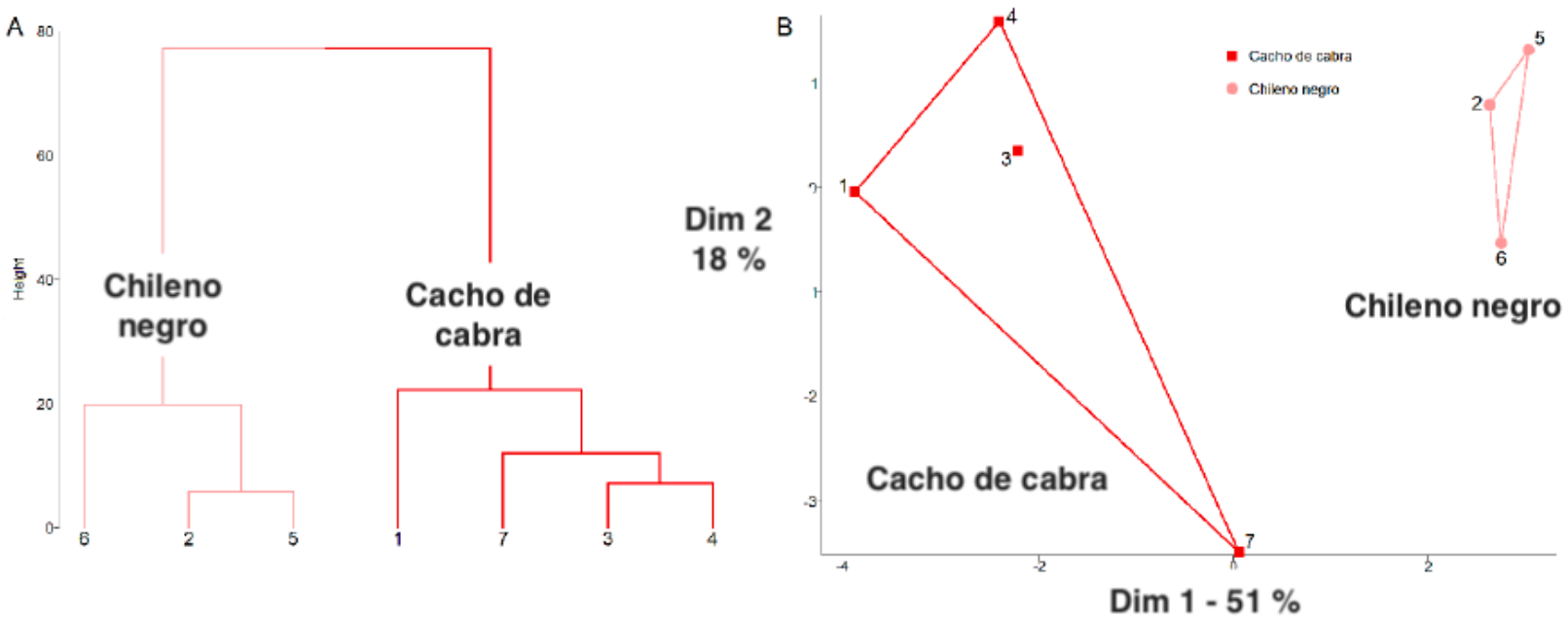
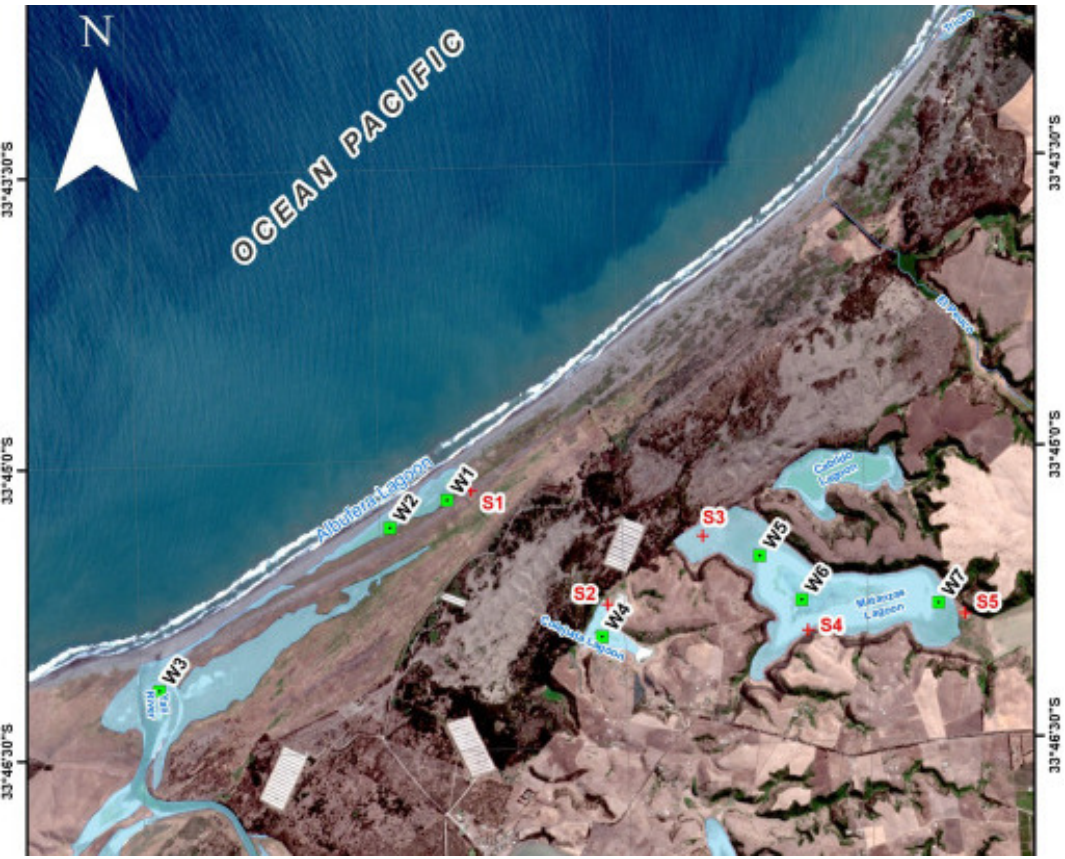


Figure 8. Cluster analysis for field data using morphological and color parameters. (**A**) Dendrogram and (**B**) partition plot generated by principal component analysis. Fields with the same landrace are grouped distinctly.

CASO 3: CONTAMINACIÓN HUMEDAL YALI.



Comparación calidad de agua lagunas Reserva el Yali.

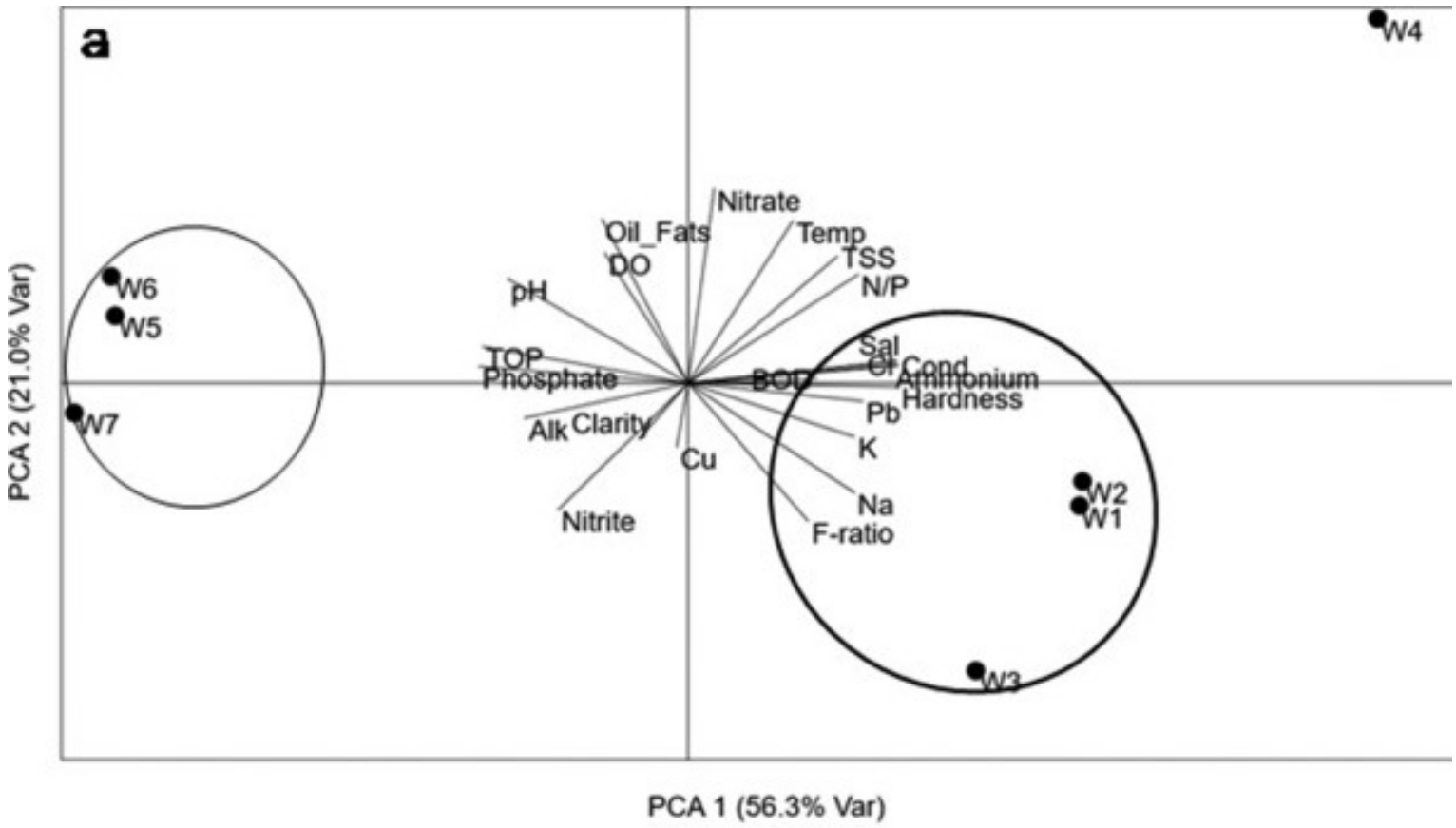
Rivera et al. 2019

Table 1: Organic and inorganic parameters measured in the water.

Station	Nitrate	Nitrite	Ammonium	Phosphate	N/P	TOP
W1	12.8	0.05	25.4	0.67	0.05	0.04
W2	13.1	0.05	19.5	4.37	0.33	0.05
W3	3.8	0.08	20.2	4.34	1.14	0.06
W4	50.0	0.03	75.5	0.18	0.01	0.04
W5	14.4	0.06	1.8	756	52.50	0.24
W6	15.8	0.06	2.4	724	45.82	0.31

PCA YALI: GRAFICA BIPLOT

- Las gráficas biplot permiten observar simultáneamente la clasificación de muestras en grupos, la varianza explicada, la correlación entre variables y la contribución que ellas hacen a cada componente principal.



ETAPAS PARA REALIZAR UN ACP

1) Estandarizar datos: Media 0 y varianza 1. .

2) Calcular matriz de distancia (euclideana) de valores estandarizados.

3) Calcular valores y vectores propios (Eigenvalue y Eigenvector) de la matriz estandarizada.

4) Interpretación varianza explicada y gráficas biplot.

5) Opcional: Prueba de hipótesis multivariada.

MATRIZ DE DISTANCIA EUCLIDEANA

- Usar con variables cuantitativas continuas.
- Comparar efecto escala de las variables.

Sitio	Depth	Pollution	Temperature
s29	51	6.0	3.0
s30	99	1.9	2.9

$$s29 - s30 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}.$$

$$s29 - s30 = \sqrt{(51 - 99)^2 + (6.0 - 1.9)^2 + (3.0 - 2.9)^2}.$$

$$s29 - s30 = \sqrt{(2304) + (18.81) + (0.01)} = 48.17$$

ESTANDARIZACIÓN

	Depth	Pollution	Temperature
Mean	74,43	4,52	3,06
sd	15,61	2,14	0,28

Valor estandarizado : (valor original – mean) / sd
Valor estandarizado s29 : (51 – 74,43) / 16,61 = -1,501

Sitio	Depth	Pollution	Temperature
s29	-1,501	0,693	-0,201
s30	1,573	-1,222	-0,557

DISTANCIA EUCLIDEANA ESTANDARIZADA

Sitio	Depth	Pollution	Temperature
s29	-1,501	0,693	-0,201
s30	1,573	-1,222	-0,557

$$s29 - s30 = \sqrt{(-1,50 - 1,57)^2 + (0,69 - 1,22)^2 + (0,20 - 0,55)^2}.$$

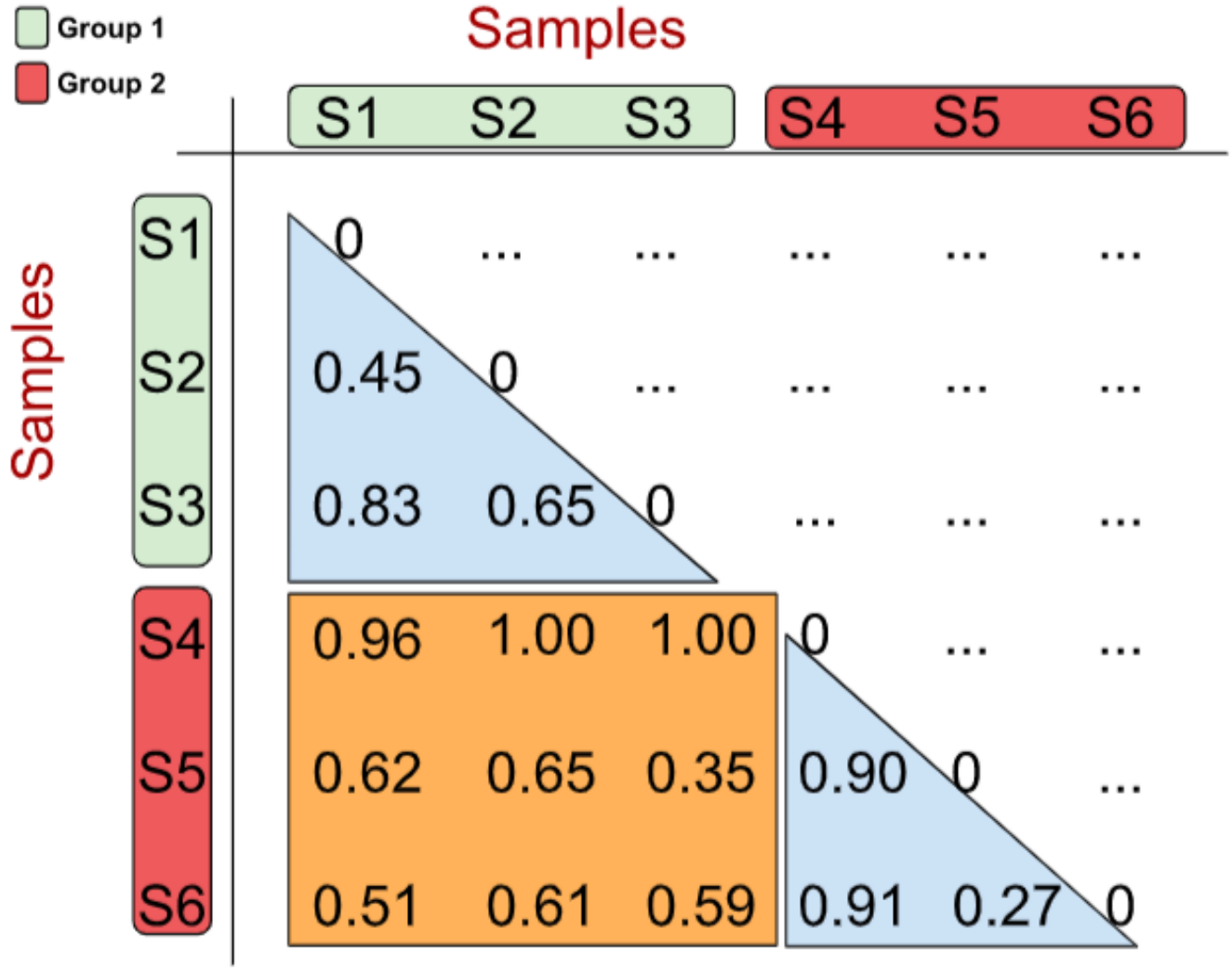
Distancia estandarizada.

$$s29 - s30 = \sqrt{(9,499) + (3,667) + (0,127)} = 3,639.$$

Distancia no estandarizada.

$$s29 - s30 = \sqrt{(2304) + (18.81) + (0.01)} = 48.17$$

MATRIZ DE DISTANCIA ENTRE GRUPOS



ANÁLISIS DE VARIANZA MULTIVARIANTE PERMUTACIONAL

- **¿Qué es un PERMANOVA?**

a) Es una prueba estadística multivariante No paramétrica.

b) Determina, en términos simples, si existen o no diferencias entre grupos.

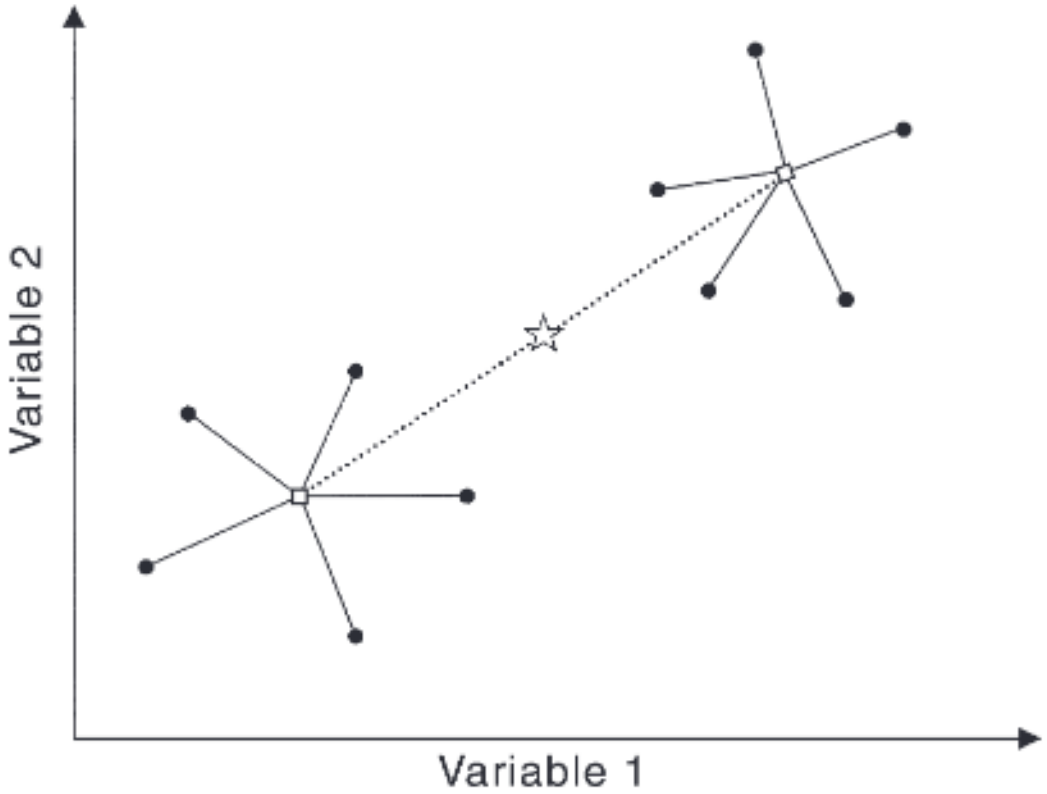
c) Usa la matriz de distancia y no los datos originales de las variables analizadas.

Fuente: Anderson, 2001

HIPÓTESIS PERMANOVA

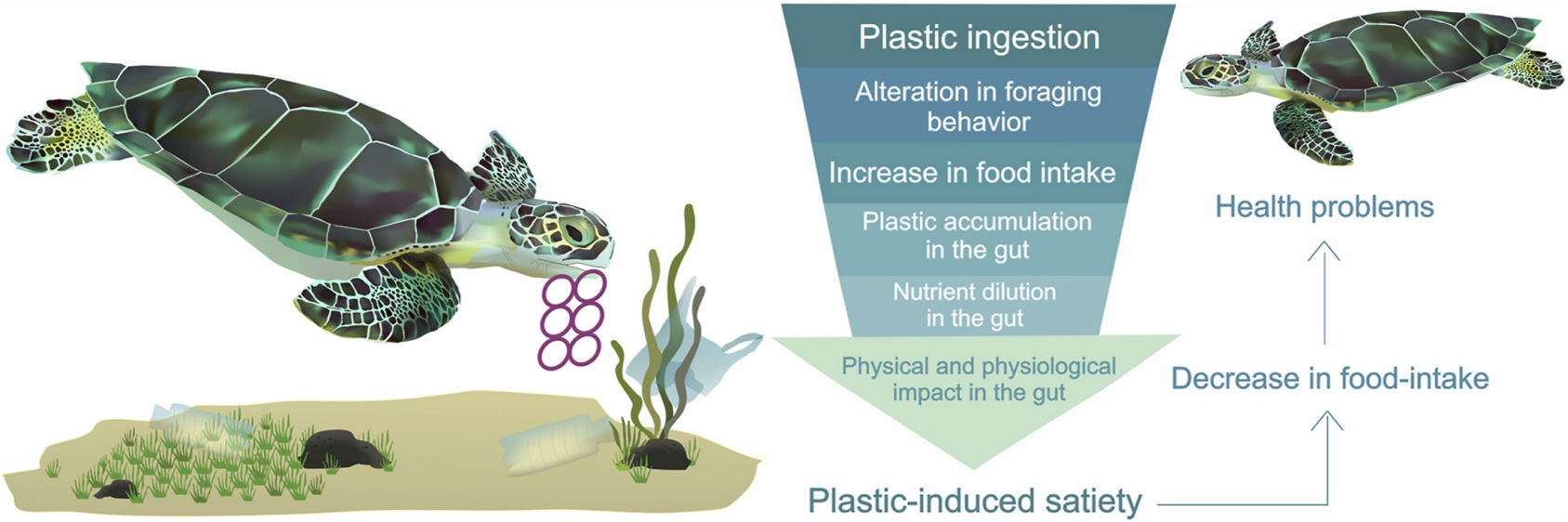
- Hipótesis.

a) H_0= No existe diferencia entre los centroides de los grupos.
b) H_1= Al menos dos centroides son diferentes.



ESTUDIO DE CASO 4: TORTUGAS Y PLÁSTICO

- Comparación acumulación de plástico intestinal en tortugas marinas brasil.



Robson et al. 2020

DATOS CONTAMINACIÓN

Table 6: Ingesta de plástico y dieta.

Area	Plastic	Chlorophyta	Rhdophyta	Phaeophyceae	Land	Animal
Fundao	No	99.32	0	0.6	0	0
Alagoas	No	0	0	12	0	0
Alagoas	No	88	0	12	0	0
Alagoas	No	70	0	20	0	0
Fundao	Yes	99.83	0.13	0	0	0
Fundao	No	97.67	0.4	1.93	0	0

PERMANOVA DATOS CONTAMINACIÓN

- Sobre matriz de distancia estandarizada.

Table 7: Permutation test for adonis under reduced model

	Df	SumOfSqs	R2	F	Pr(>F)
Area	2	54.37	0.1812	6.319	0.001
Plastic	1	1.962	0.006541	0.4561	0.819
Area:Plastic	2	7.042	0.02347	0.8184	0.567
Residual	55	236.6	0.7887	NA	NA
Total	60	300	1	NA	NA

RESUMEN DE LA CLASE

- ¿Qué es un análisis de componentes principales?.
- Etapas para realizar un ACP.
- Varianza explicada.
- Graficas biplot.
- PERMANOVA.
- 4 estudios de caso: Fraude alimentario en salmón; Ají chileno; Contaminación Humedal; Tortugas y contaminación por plástico.