OCE 313
TÉCNICAS DE ANÁLISIS NO PARAMÉTRICO

CLASE 11 – Introducción a los métodos multivariantes: Matriz de distancia

Dr. José Gallardo

Mayo 2021

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Contenidos de la clase

- Datos multivariantes.
- Estudios de caso: Análisis de cluster, análisis de componentes principales.
- Matriz de distancia (similaridad): cálculo manual
- Matriz de distancia con R

Datos multivariantes.

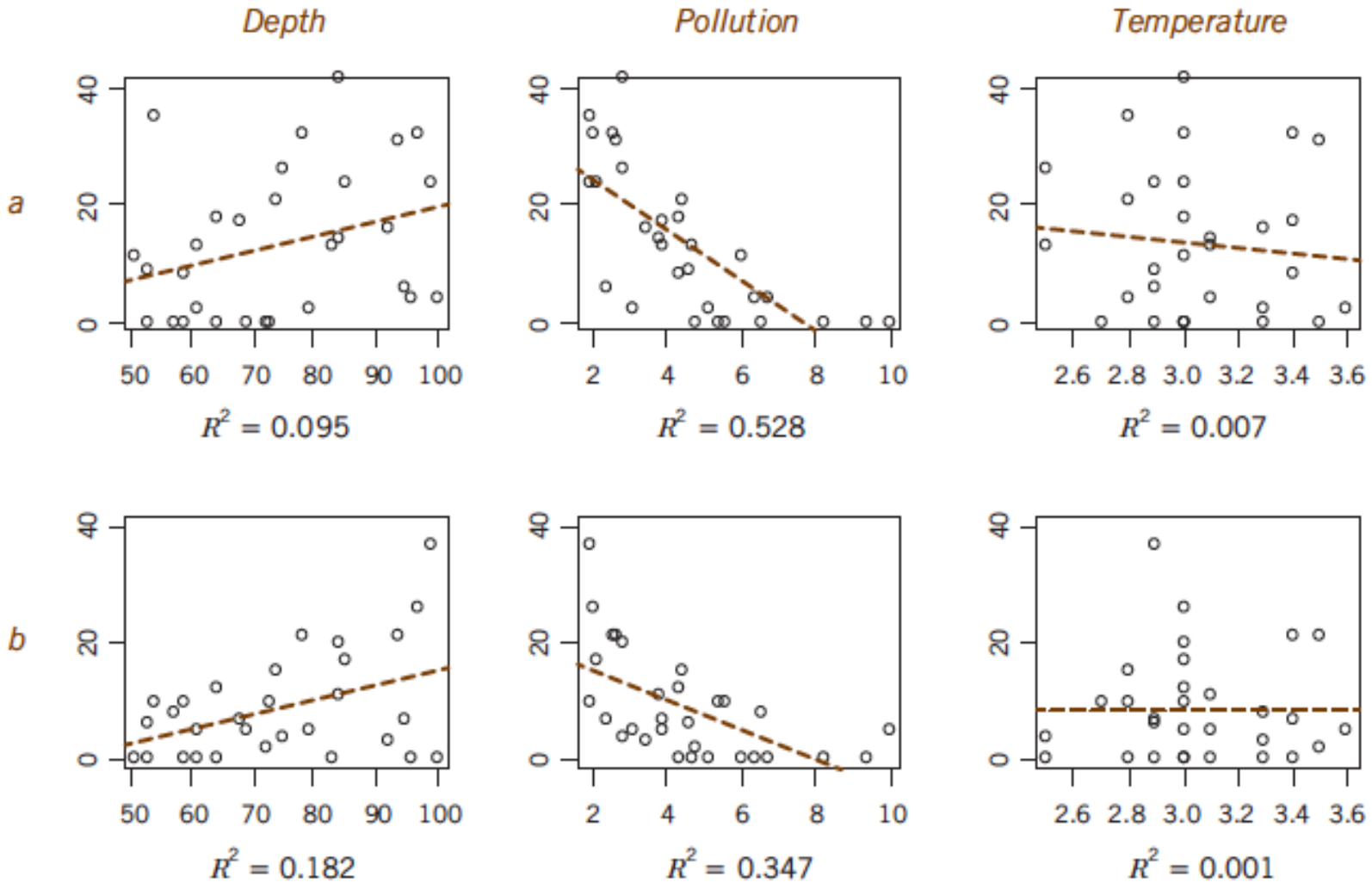
Datos multivariantes (toy dataset)

Site No.	Species counts					Environmental variables			
	a	*b*	*c*	*d*	*e*	*Depth (x)*	*Pollution (y)*	*Temperature (z)*	*Sediment (s)*
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

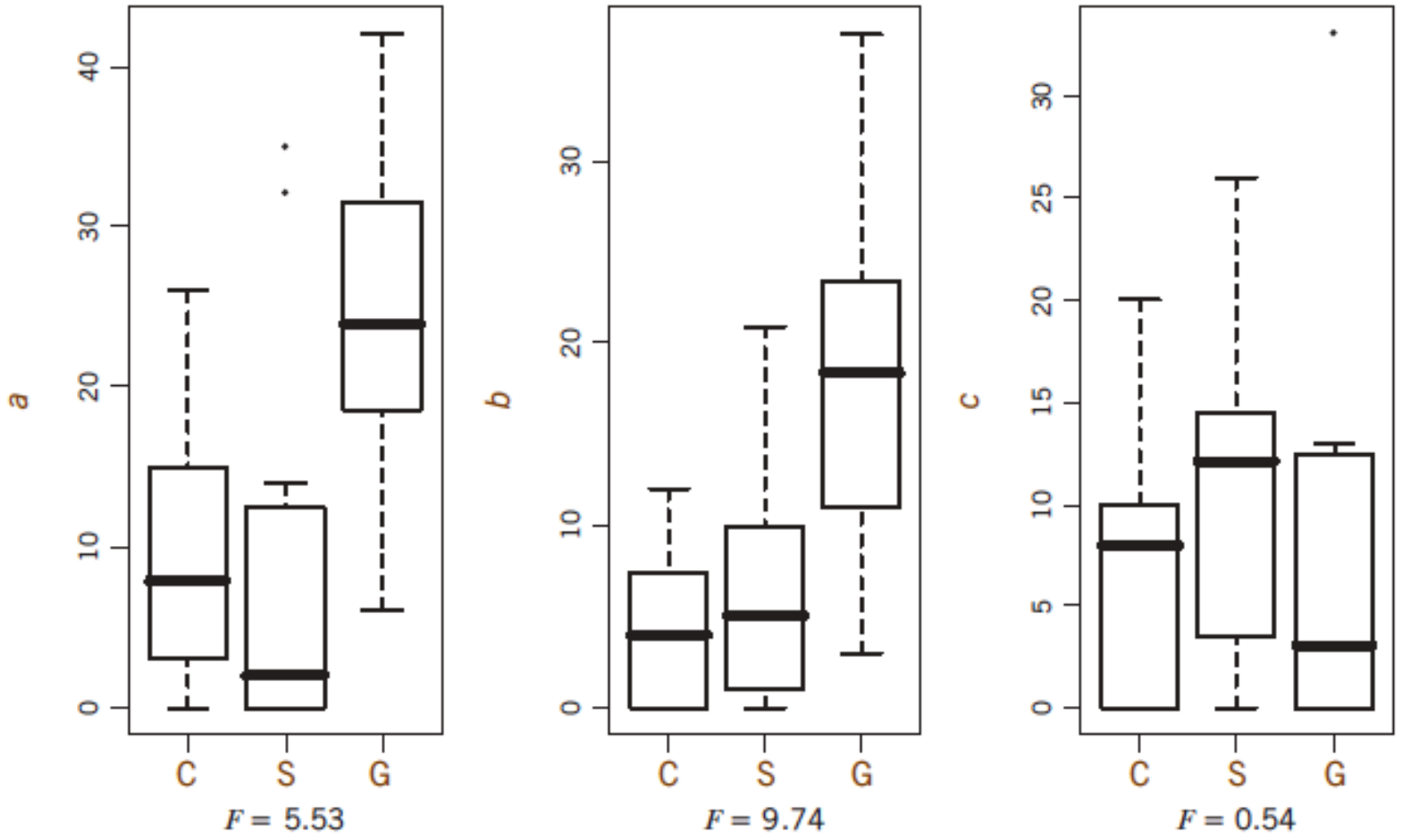
C: Arcilla
S: Arena
G: Gravilla

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Hipótesis básicas: Regresión y-x.

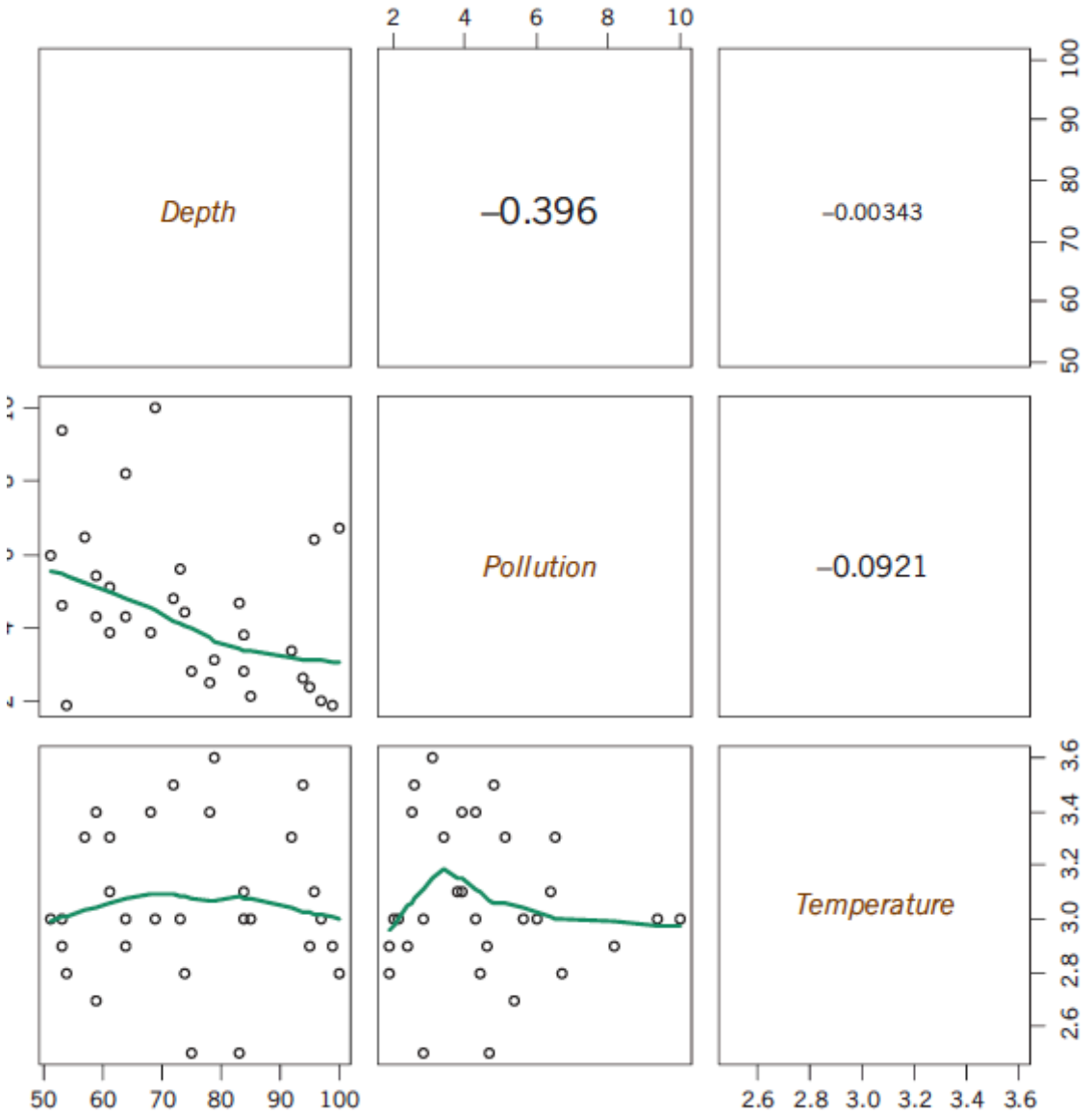


Hipótesis básicas: Comparación de medias (y-x).



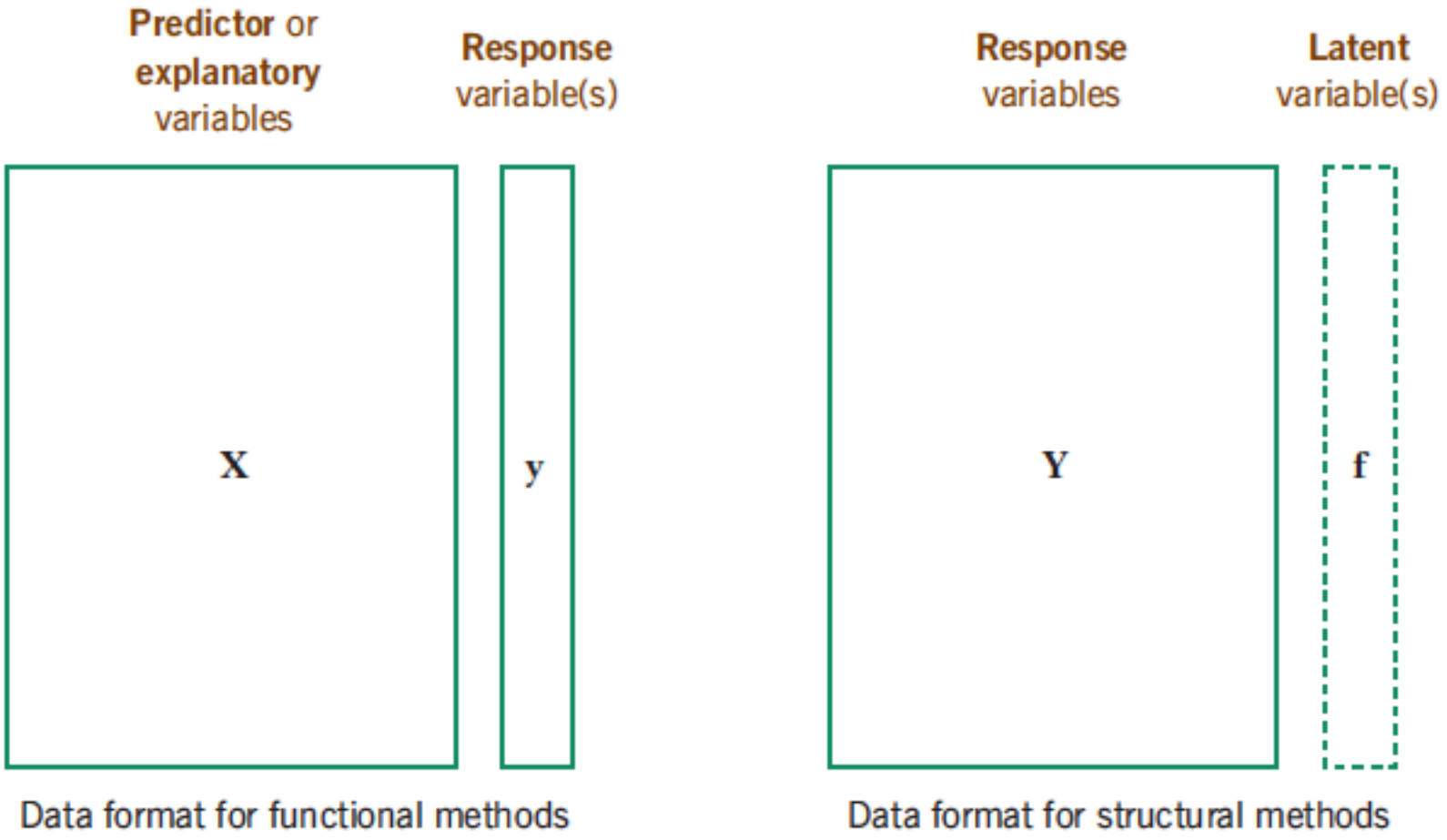
PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Hipótesis básicas: Correlación(y-x)

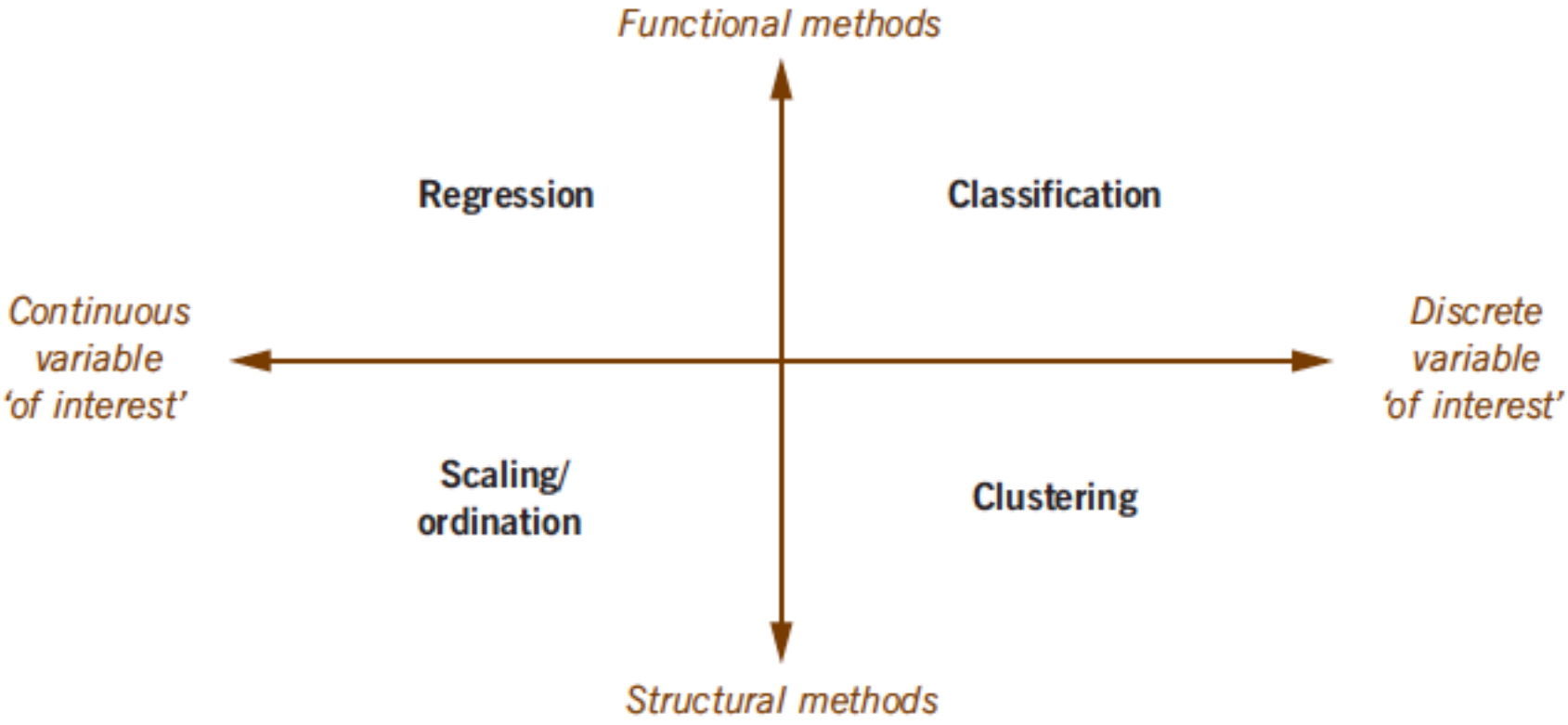


PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Análisis multivariado: Función v/s estructura



Las cuatro esquinas del análisis multivariado.



Estudios de caso

Análisis de agrupamiento (clustering)

Descubriendo una variable latente categórica

Problema: Almejas de distintas especies ($V_{latente}$)

¿Será posible agruparlas por especie?

¿Qué variables respuestas (V_r) medir?

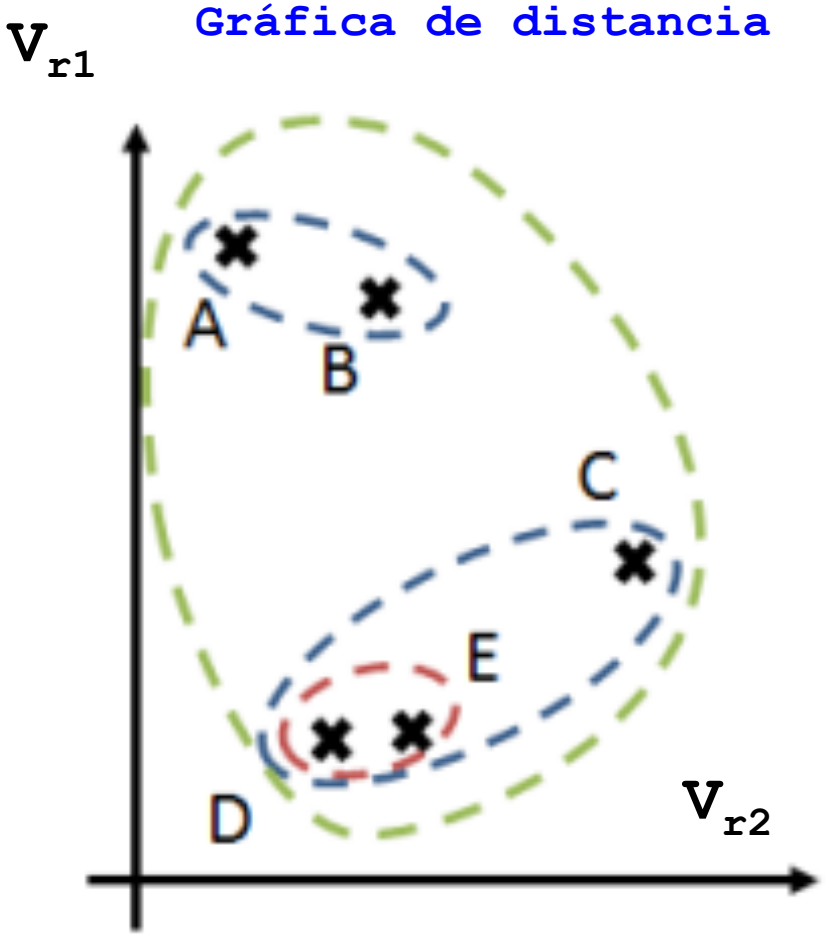
¿Cuántas V_r o varias?

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

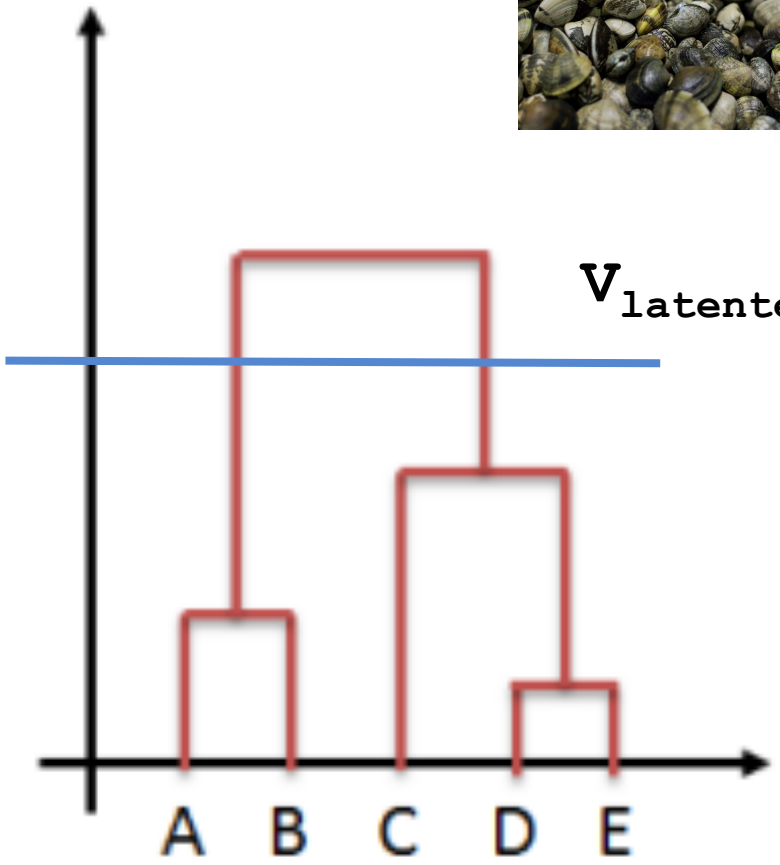
Análisis de agrupamiento 2

Descubriendo una variable latente categórica

Gráfica de distancia



Agrupamiento jerarquico



Análisis de componentes principales (ACP)

Descubriendo variables latentes continuas

Problema cambio climático: demasiadas variables respuesta, muchas de ellas correlacionadas

¿Puedo reducir las variables?

¿Existe algún patrón?

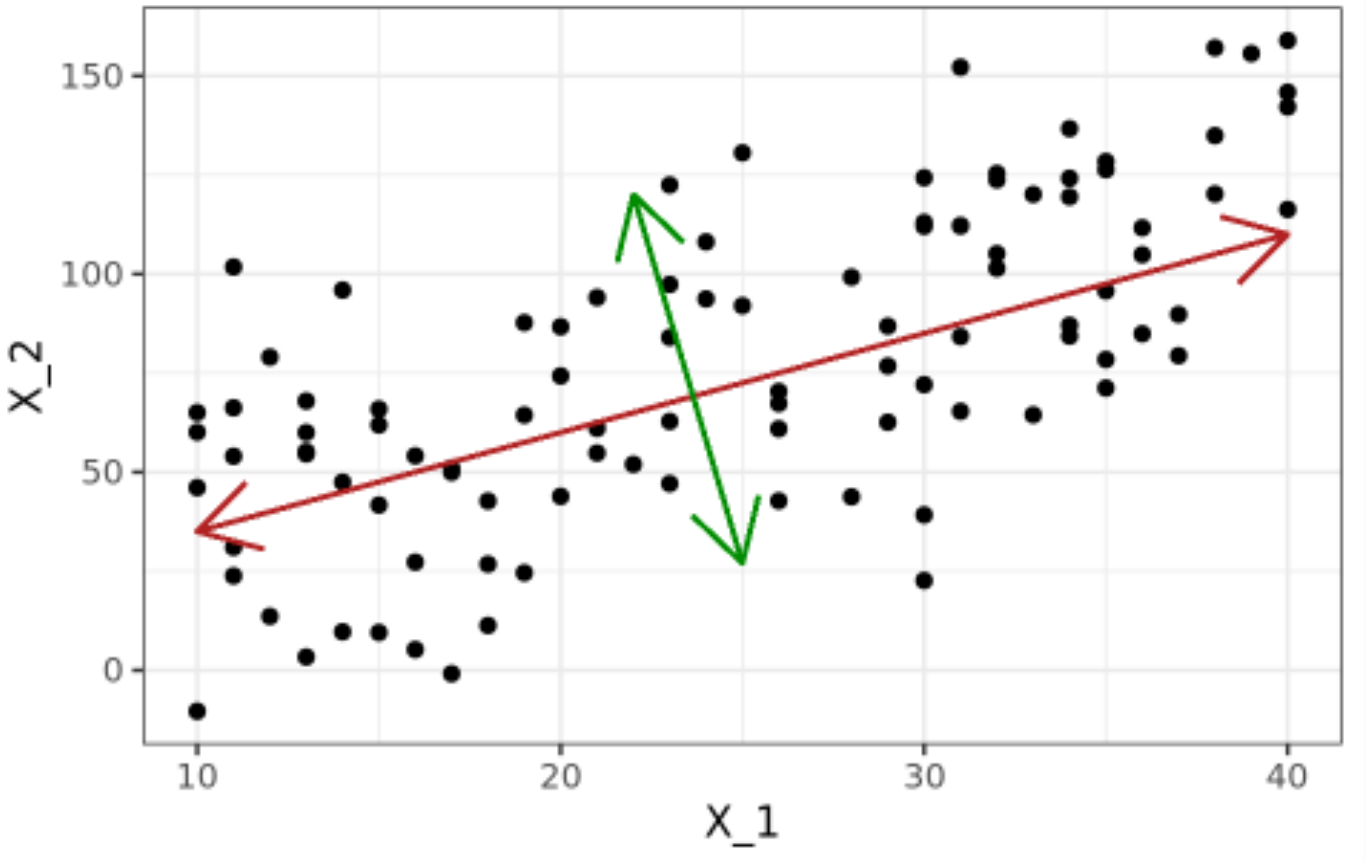
Year	*AO*	*AO_winter*	*AO_summer*	*NPI*	*NPI_spring*	*NPI_winter*	*Temp*	⋯	*IceCover*	*IceFreeDays*
1981	–0.4346	–0.1683	–0.2410	–2.09	–0.15	–4.46	–3.9	⋯	–0.64	140
1982	0.2977	–0.3750	0.3083	0.75	0.13	1.70	–4.7	⋯	–1.65	144
1983	0.0319	0.1733	0.4653	–2.54	0.30	–5.44	–4.4	⋯	–0.34	116
1984	–0.1917	0.2627	0.0240	–1.20	–0.23	–2.62	–7.0	⋯	0.15	134
1985	–0.5192	–1.2667	0.2678	0.52	–0.43	1.11	–5.9	⋯	–0.21	120
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2002	0.0717	0.4543	0.0187	0.13	–0.18	0.30	–3.3	⋯	0.78	203
2003	0.1521	–0.6453	0.0399	–1.67	–0.40	–3.84	–3.8	⋯	–1.60	179
mean	0.0466	0.0587	0.1652	–0.440	0.023	–0.950	–5.15		–0.317	151.8
variance	0.1699	1.1687	1.0505	1.166	0.491	5.603	1.08		0.888	398.5

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

ACP 2: Definición e interpretación geométrica.

Descubriendo variables latentes continuas

CP: Combinación lineal de las variables originales no corr. entre si (perpendiculares / ortogonales).

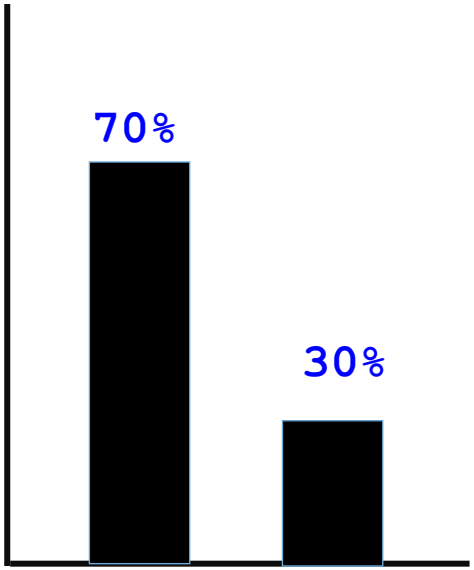


Ejemplo
2 var.cor.
2 CP

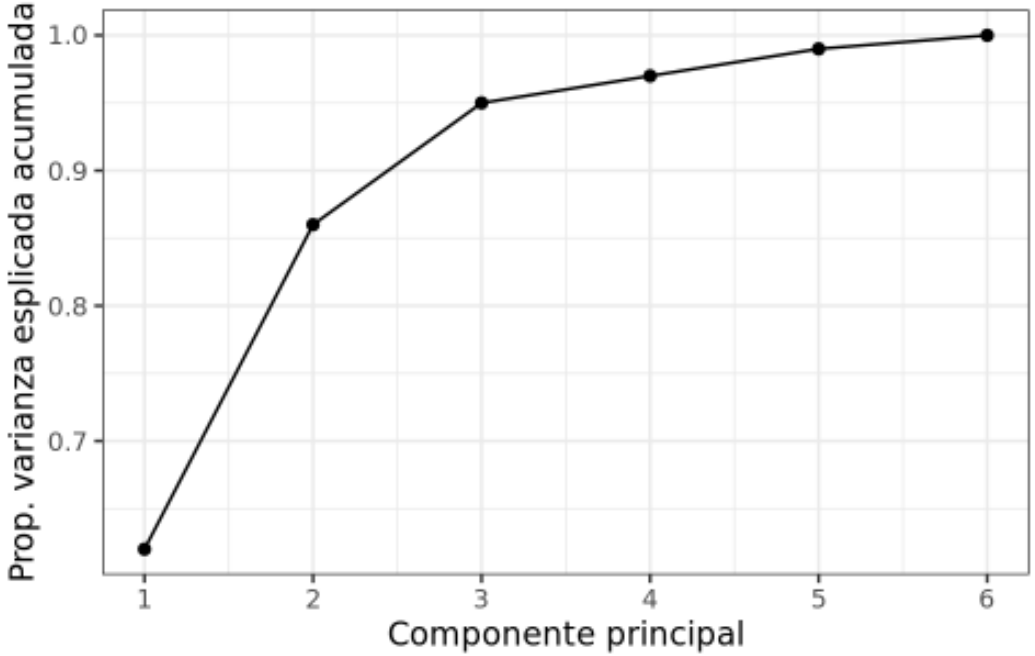
ACP 2: Varianza explicada

Descubriendo variables latentes continuas

Varianza explicada 2 var cor.



Varianza explicada por 6 variables



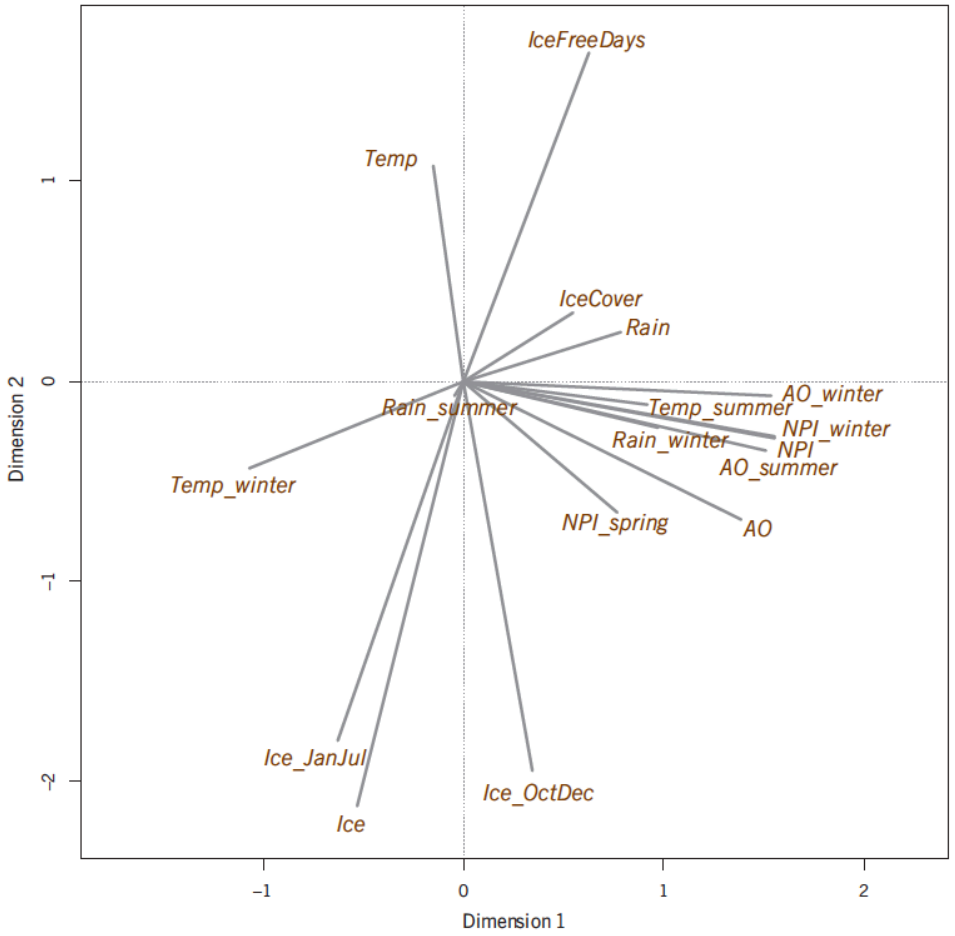
ACP 3: Gráficas biplot

Descubriendo variables latentes continuas

Ventajas

- Reducción de dimensionalidad.
- Mayor varianza explicada por los primeros 2 CP.

Gráficas biplot
(Muchas variables en 2 CP)

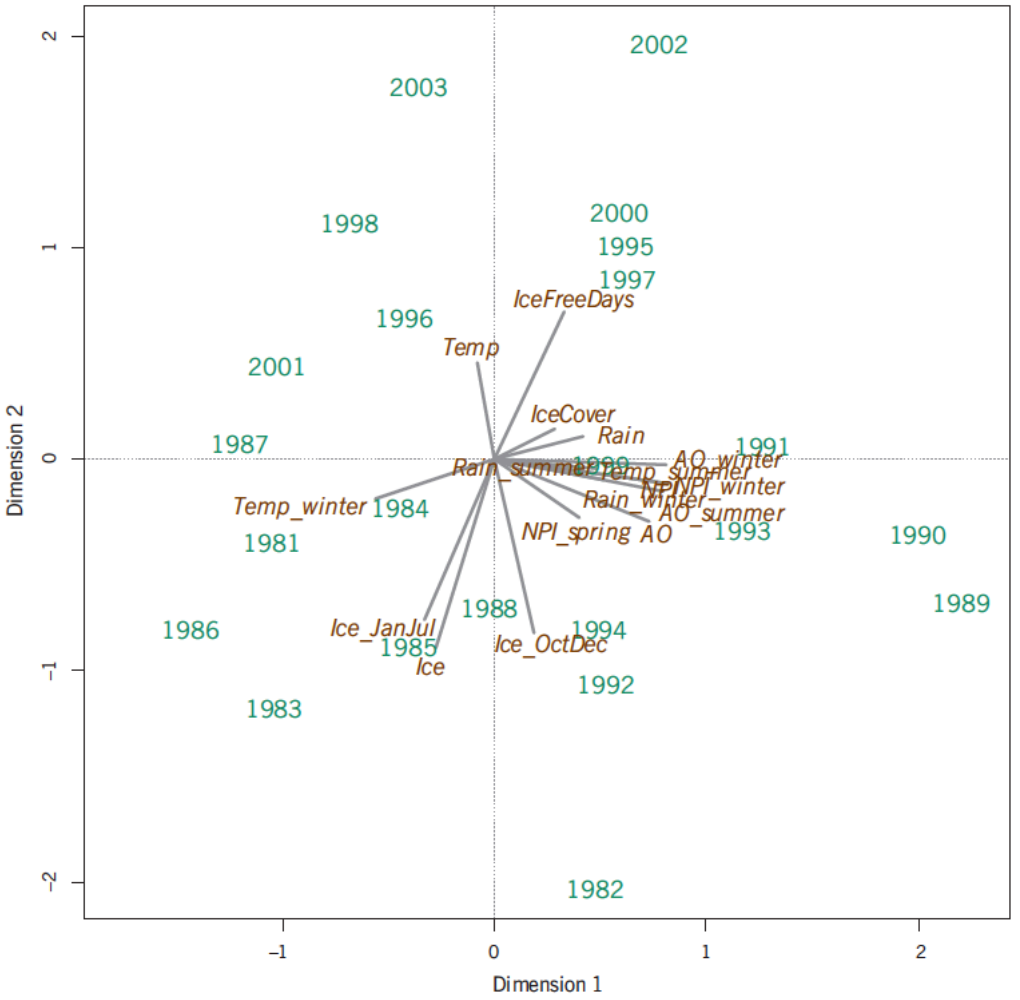


ACP 4: Descubriendo patrones

Descubriendo varias variables latentes continuas

Existe un patrón de cambio climático: Si.

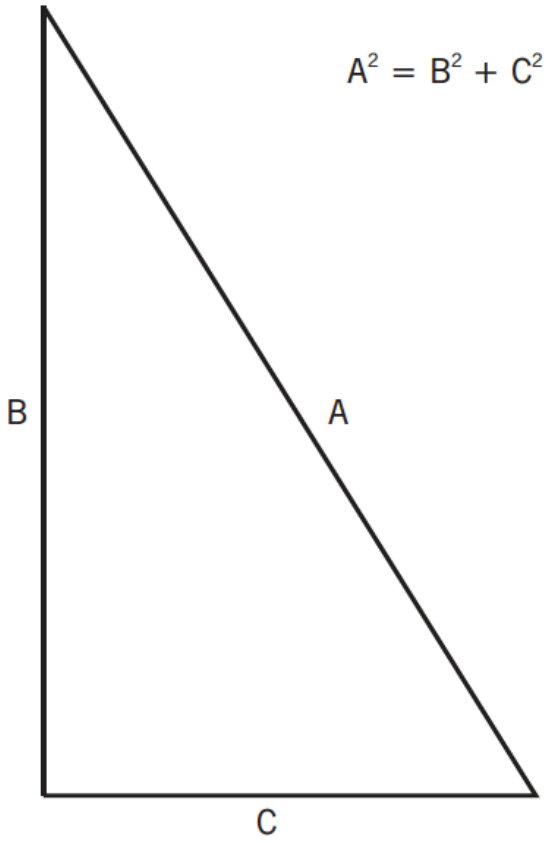
Qué variables influyen: cada cuadrante se puede asociar a una combinación de variables correlacionadas



Matriz de distancia (similaridad)

Variables continuas

Teorema de pitágoras



TEOREMA DE PITÁGORAS

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

De donde se extrae que

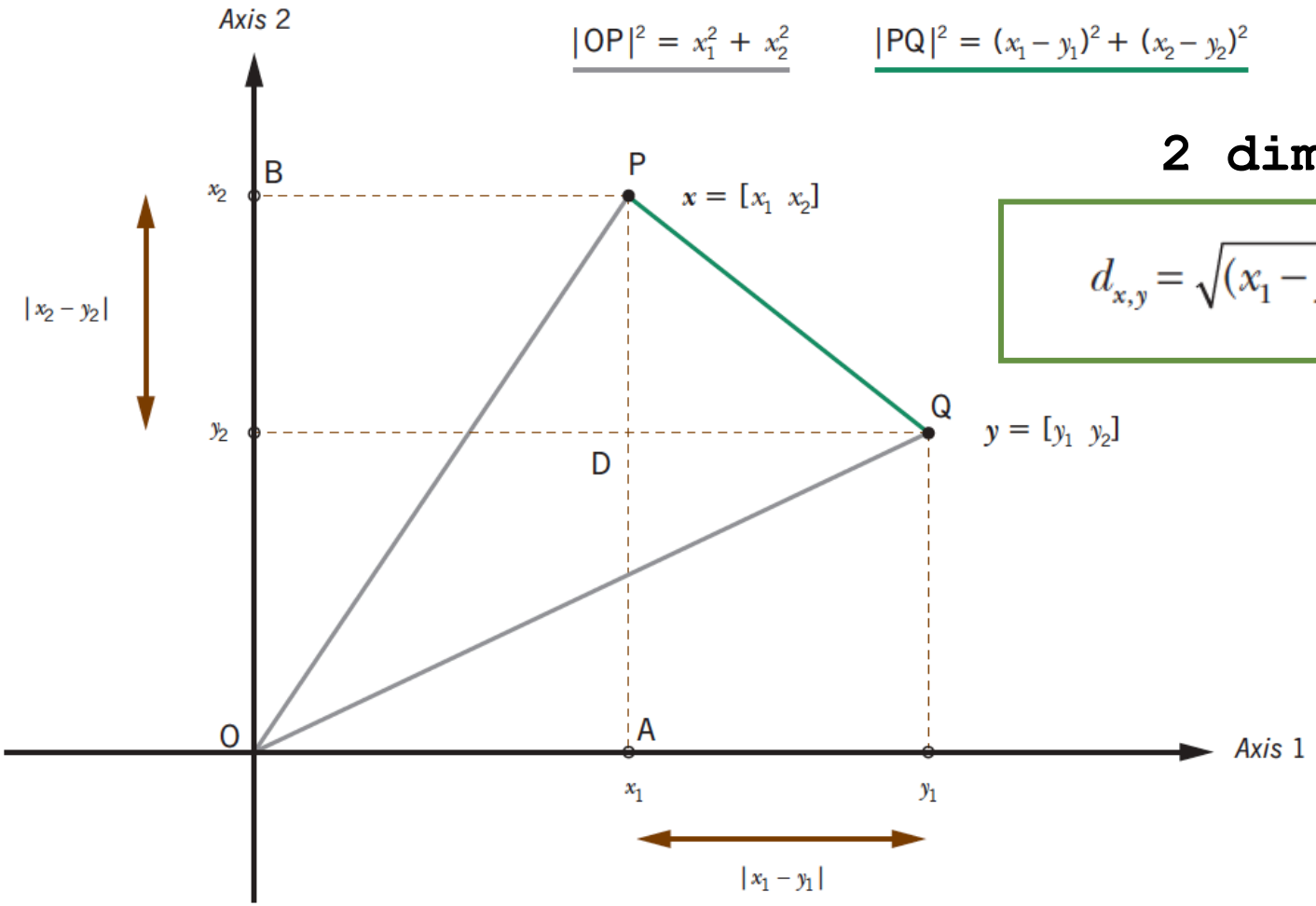
$$c = \sqrt{a^2 + b^2}$$

$$b = \sqrt{c^2 - a^2}$$

$$a = \sqrt{c^2 - b^2}$$

https://www.educapeques.com/recursos-para-el-aula/teorema-de-pitagoras.html

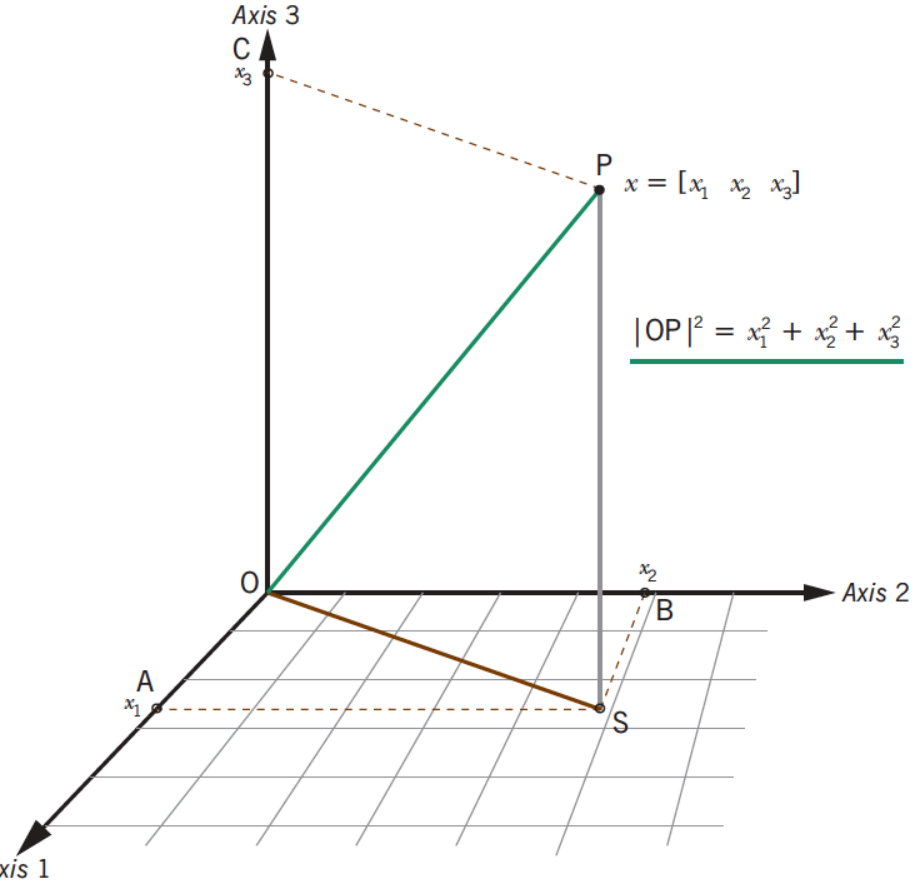
Distancia euclideana: 2 dim.



2 dimensiones

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Distancia euclideana: 3 dim.



3 dimensiones

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

> 3 dimensiones

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

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Calculo distancia euclideana

para variables continuas

Sitio	*Depth*	*Pollution*	*Temperature*
s29	51	6,0	3,0
s30	99	1,9	2,9

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

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

$$= \sqrt{2304 + 16.81 + 0.01} = \sqrt{2320.82} = 48.17$$

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Estandarice antes de calcular distancia euclideana

Sitio	*Depth*	*Pollution*	*Temperatura*
media	74,433	4,517	3,057
ds	15,615	2,141	0,281

Valor estandarizado : (valor original – media) / ds

Sitio	*Depth*	*Pollution*	*Temperatura*
s29	51	6,0	3,0
s30	99	1,9	2,9
Sitio	Estandarizado		
S29			
s30			

PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO

Calcule distancia estandarizada

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

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2}$$

$$d_{s29,s30} = \sqrt{[-1.501 - 1.573]^2 + [0.693 - (-1.222)]^2 + [-0.201 - (-.557)]^2}$$

$$= \sqrt{9.449 + 3.667 + 0.127} = \sqrt{13.243} = 3.639$$

Matriz de distancia (similaridad)

No euclideana

Disimilaridad de Bray-Curtis

para variables discretas

Sitio	*a*	*b*	*c*	*d*	*e*	*Suma*
s29	11	0	7	8	0	26
s30	24	37	5	18	1	85

$$b_{ii'} = \frac{\sum_{j=1}^{J} \left| n_{ij} - n_{i'j} \right|}{n_{i+} + n_{i'+}}$$

$$b_{s29,s30} = \frac{|11-24| + |0-37| + |7-5| + |8-18| + |0-1|}{26+85} = \frac{63}{111} = 0.568$$

Resumen de la clase

- Revisión e importancia de datos multivariantes.

- Introducción al Análisis de cluster y al análisis de componentes principales.

- Calculo de matriz de distancia (similaridad)