

Análisis espaciales y multivariantes en R aplicados a estudios de biodiversidad

Análisis multivariantes en R

Diego Nieto Lugilde. Profesor Titular de la Universidad de Córdoba
(España)



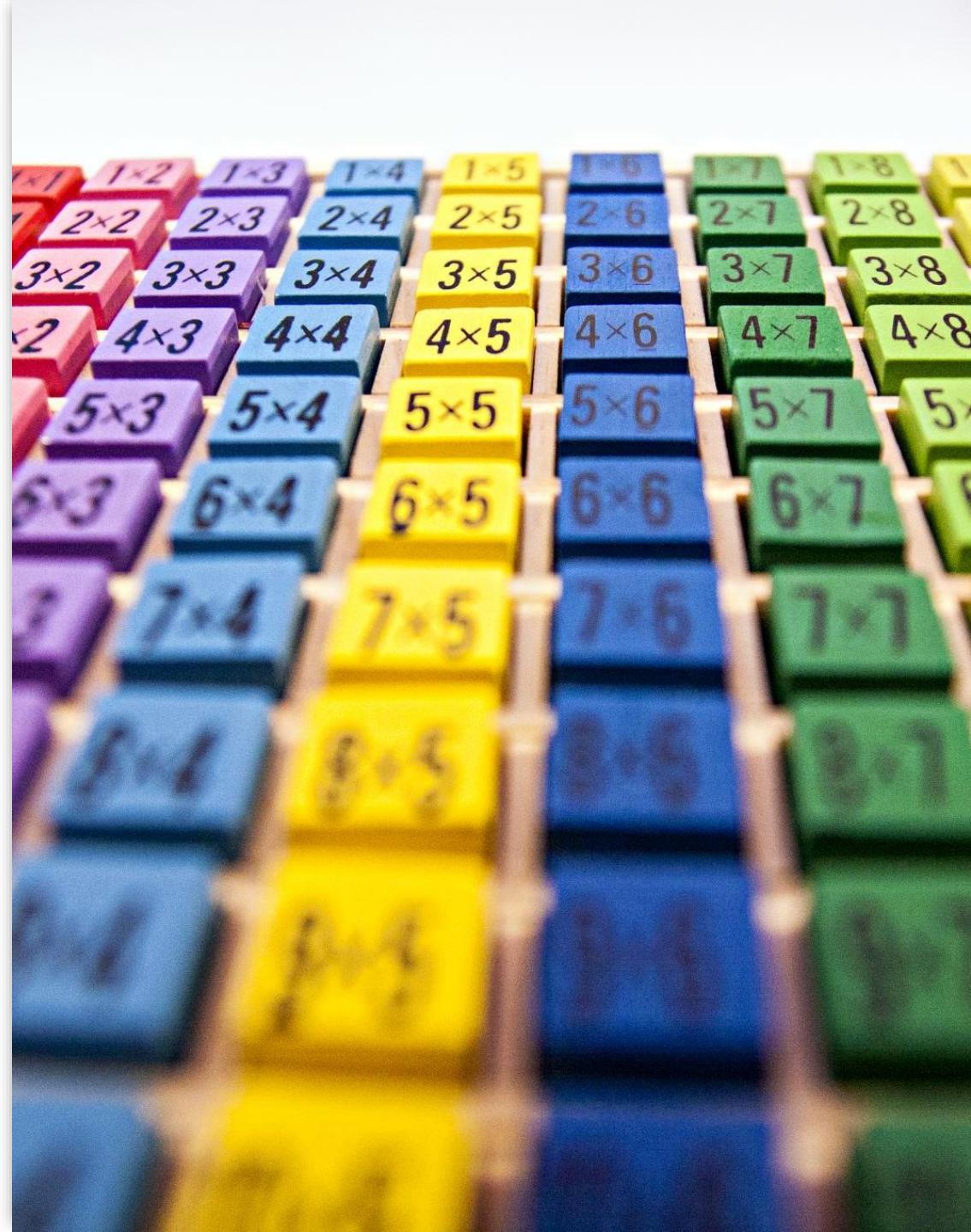
UNIVERSIDAD
DE
CÓRDOBA



Co-funded by
the European Union

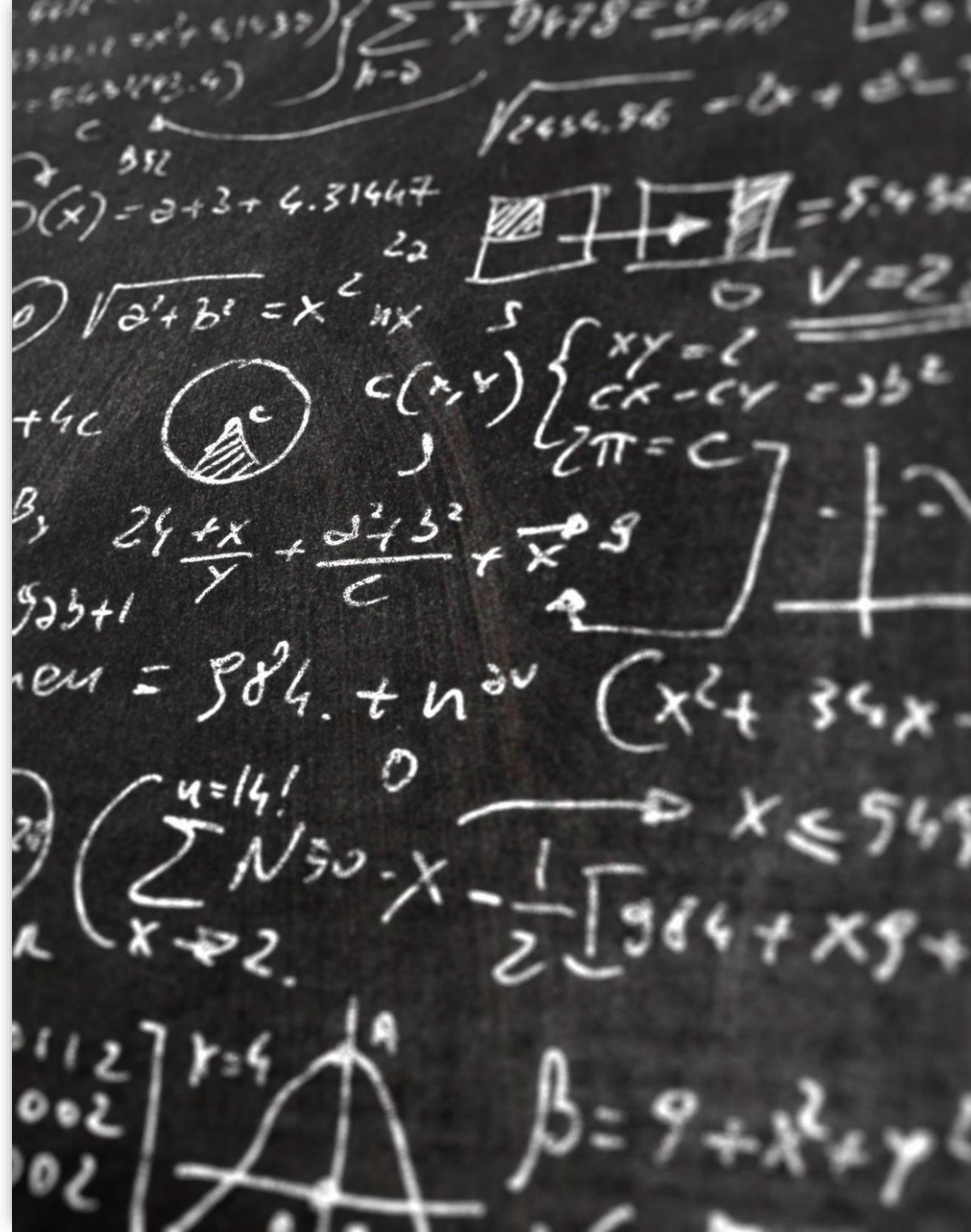
Contenidos

- ¿Qué son los análisis multivariantes?
 - Matrices de comunidades
- ¿Cuáles son los análisis multivariantes más frecuentes?
 - Ordenación
 - Agrupación
- ¿Cómo se realizan en R?
 - Paquete vegan



Análisis Multivariante

- Es un conjunto de métodos para describir e interpretar los datos que provienen de la observación de varias variables conjuntamente
- Estudiar, analizar, representar e interpretar los datos que resultan de observar más de una variable sobre una muestra de individuos
- Las variables deberían ser homogéneas y correlacionadas, sin que alguna predomine sobre las demás



Matrices multivariantes (variables ambientales)

Comunidad	T ^a annual	Prec. anual	C suelo	N suelo
1	10	100	5,5	25
2	8	100	7,5	15
...				
n	12	150	8,3	20

Matrices multivariantes (variables bióticas)

Comunidad	Sp 1	Sp 2	Sp 3	Sp 4
1	10	10	80	55
2	8	10	75	45
...				
n	12	15	83	70

Matrices multivariantes (variables bióticas)

Población	Genotipo 1	Genotipo 2	Genotipo 3	Genotipo 4
1	10	10	80	55
2	8	10	75	45
...				
n	12	15	83	70

Matriz de comunidades de las marismas saladas de Nueva Jersey con % cobertura

Species	Quadrats											
	1	2	3	4	5	6	7	8	9	10	11	12
1. <i>Atriplex patula</i> var <i>hastata</i>	1	10	2	1	1	2	5		1		5	2
2. <i>Distichlis spicata</i>		15	80	2	10	15	30	1	10	10	20	
3. <i>Iva frutescens</i>							5	1	2	1	20	10
4. <i>Juncus gerardii</i>			1			40	1					
5. <i>Phragmites communis</i>								1	10	20	5	30
6. <i>Salicornia europaea</i>	5	10	2	1	1		2			2		
7. <i>Salicornia virginica</i>				5	10							
8. <i>Scirpus olneyi</i>						5	20				1	
9. <i>Solidago sempervirens</i>									1	5	1	2
10. <i>Spartina alterniflora</i>	75	30	5	20	5	1		10	1	2		
11. <i>Spartina patens</i>							20	10	50		2	5
12. <i>Suaeda maritima</i>				20	10							

Matriz de comunidades de marismas saladas de Nueva Jersey con presencia/ausencia y transpuesta

Quadrats	Species											
	1	2	3	4	5	6	7	8	9	10	11	12
	<i>Atriplex patula</i> var. <i>hastata</i>	<i>Distichlis spicata</i>	<i>Iva frutescens</i>	<i>Juncus gerardii</i>	<i>Phragmites</i> <i>communis</i>	<i>Salicornia europaea</i>	<i>Salicornia</i> <i>virginica</i>	<i>Scirpus olneyi</i>	<i>Solidago</i> <i>sempervirens</i>	<i>Spartina alterniflora</i>	<i>Spartina patens</i>	<i>Suaeda maritima</i>
1	1	0	0	0	0	1	0	0	0	1	0	0
2	1	1	0	0	0	1	0	0	0	1	0	0
3	1	1	0	1	0	1	0	0	0	1	0	0
4	1	1	0	0	0	1	1	0	0	1	0	1
5	1	1	0	0	0	1	1	0	0	1	0	1
6	1	1	0	1	0	0	0	1	0	1	0	0
7	1	1	1	1	0	1	0	1	0	0	1	0
8	0	1	1	0	1	0	0	0	0	1	1	0
9	1	1	1	0	1	0	0	0	1	1	1	0
10	0	1	1	0	1	1	0	0	1	1	0	0
11	1	1	1	0	1	0	0	1	1	0	1	0
12	1	0	1	0	1	0	0	0	1	0	1	0

Tipos de análisis multivariantes

Análisis de regresión múltiple

Análisis de ordenación

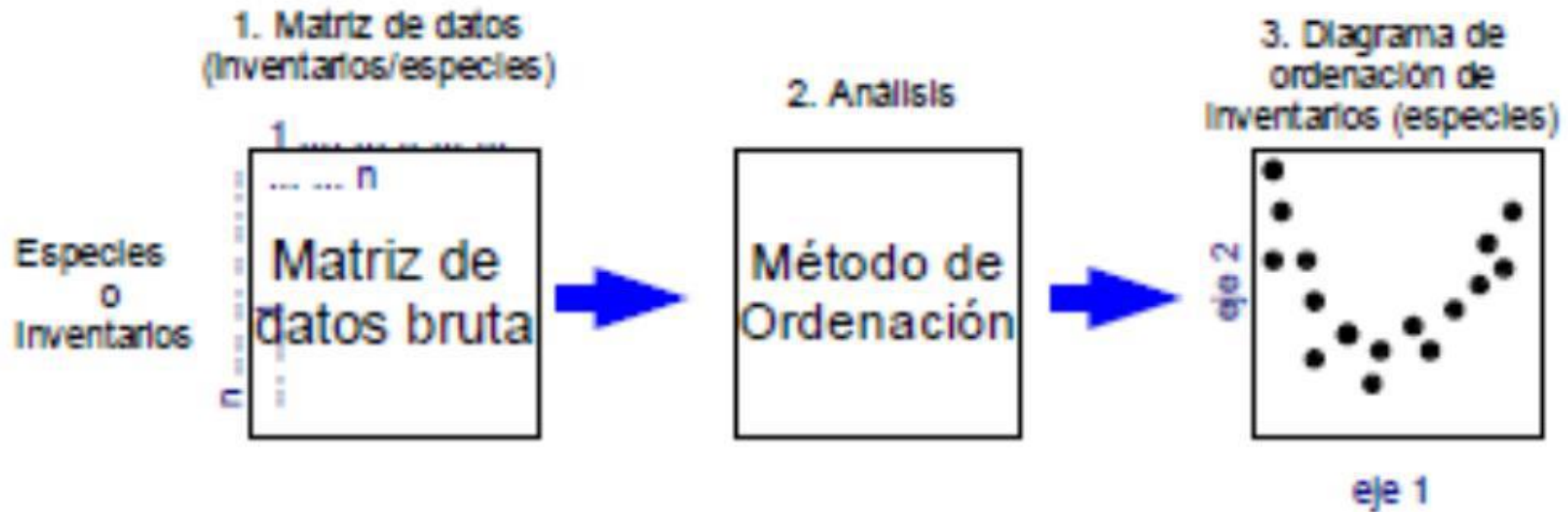
- Análisis de Componentes Principales (PCA) y sus derivadas (v.gr. RDA)
- Análisis de Correspondencias (AC) y sus derivadas (v.gr. DCA, CCA)
- Escalamiento Multidimensional No-Métrico

Análisis de agrupación

- Clasificación jerárquica
- K-means
- Twinspam

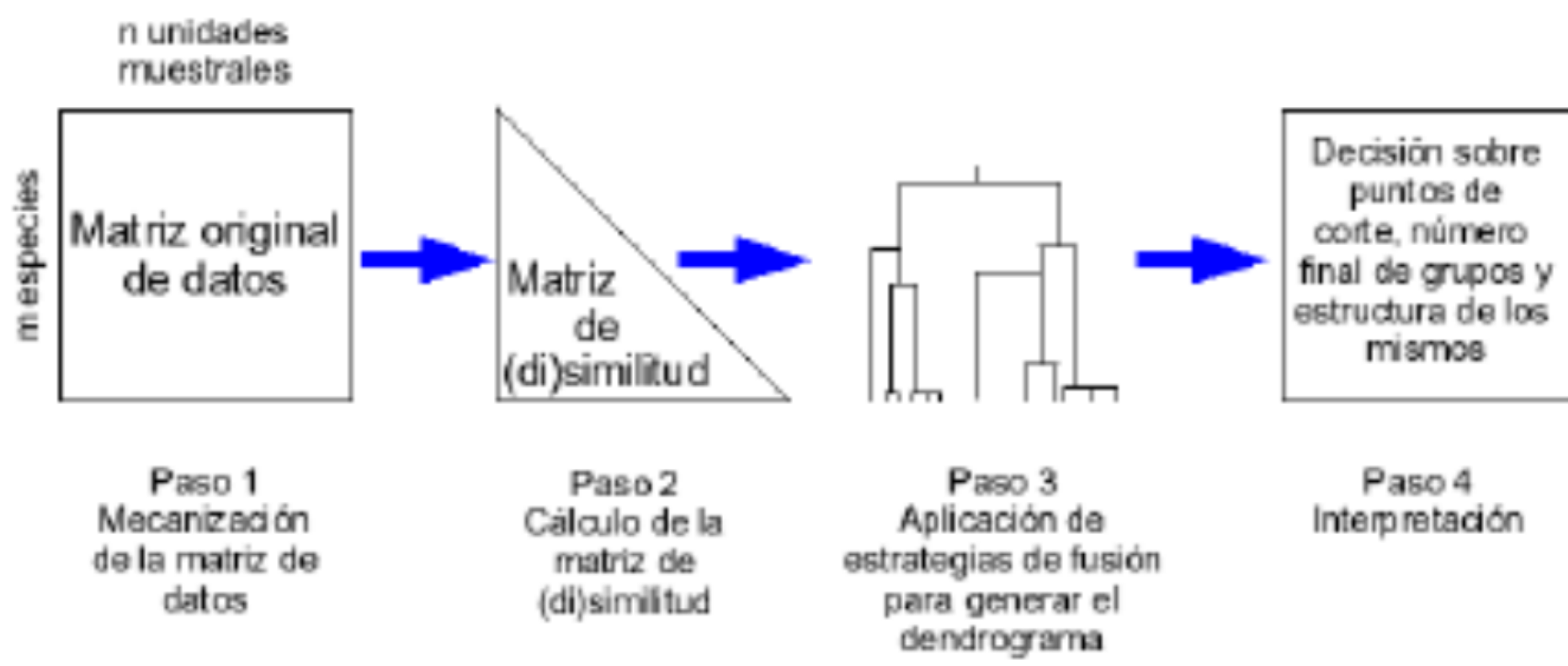
Análisis de ordenación

- Organizar los datos en el espacio ambiental
 - Permite encontrar tendencias en las variables medidas
 - ... y reducir la dimensionalidad de los datos al detectar tendencias compartidas

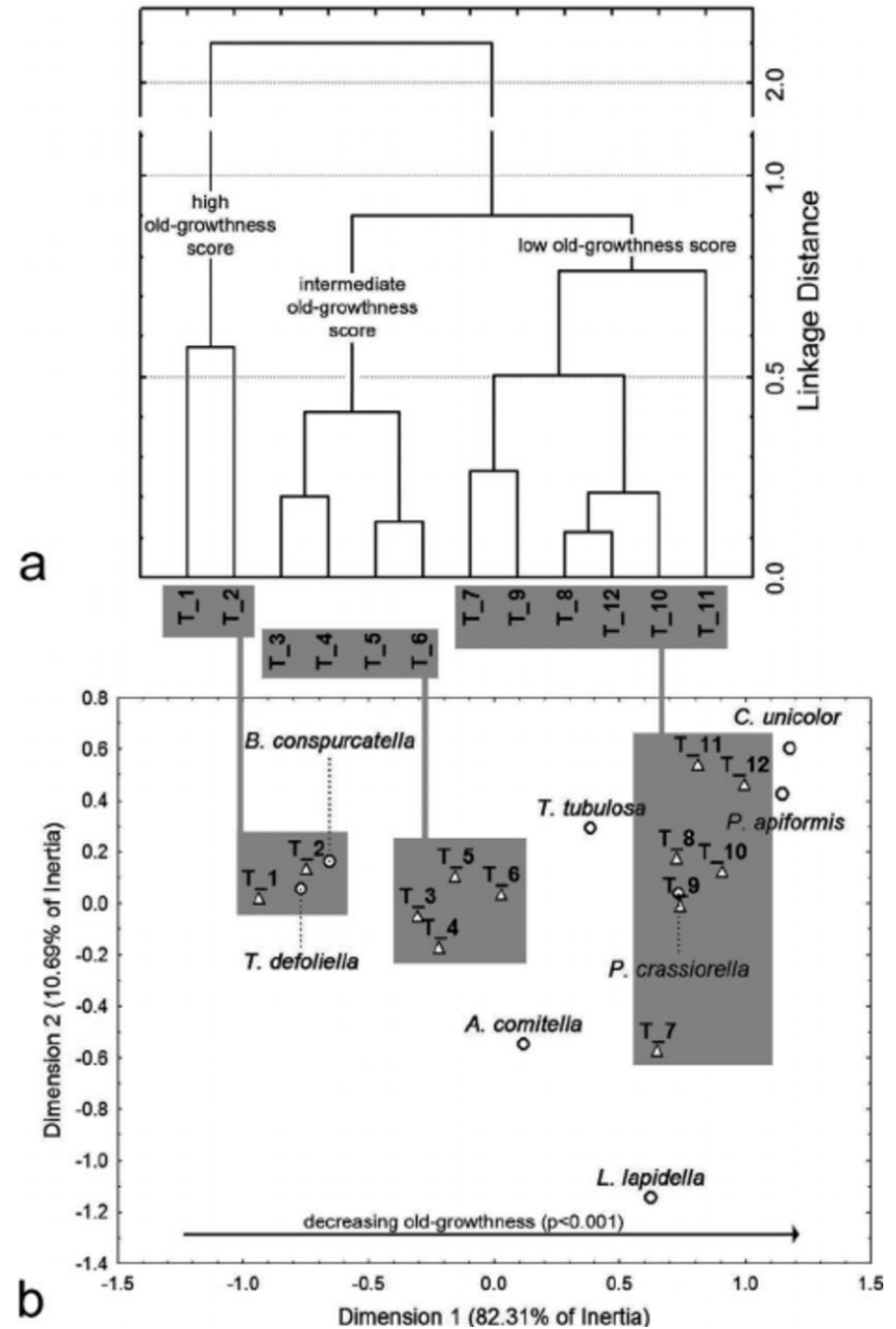
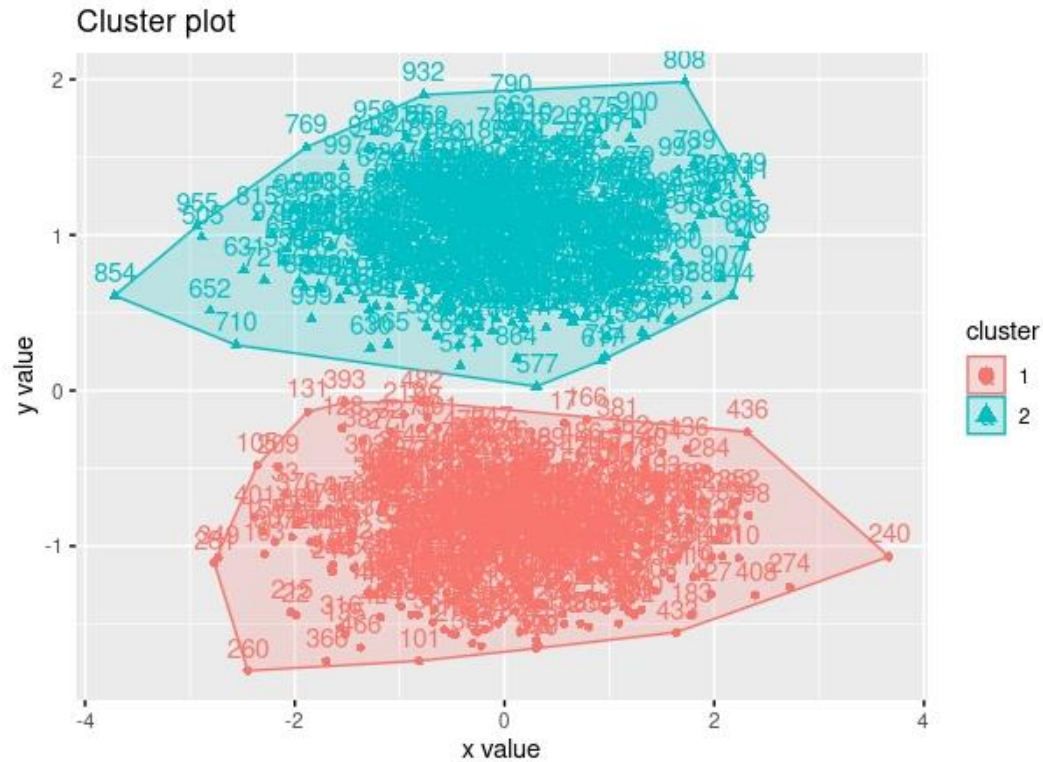



Análisis de clasificación

- Agrupa los datos para identificar conjuntos de datos muy similares entre si y, a su vez, lo suficientemente diferentes del resto de datos



Ordenación versus agrupamiento





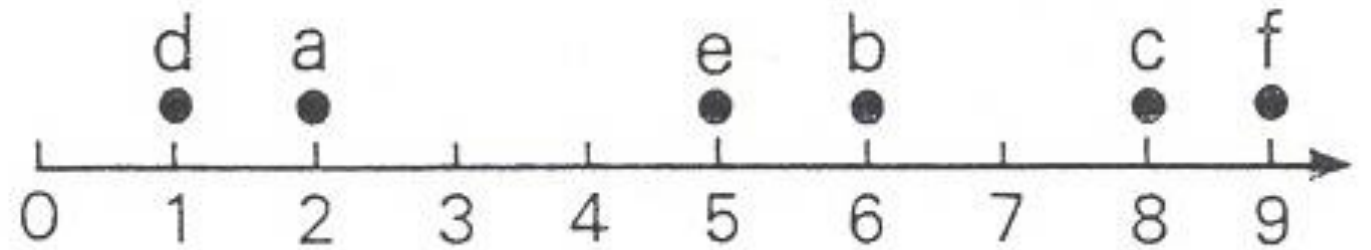
Análisis de ordenación

Jugamos a ordenar?

Localidad	Sp1
a	2
b	6
c	8
d	1
e	5
f	9

No es tan difícil ¿verdad?

Localidad	Sp1
a	2
b	6
c	8
d	1
e	5
f	9

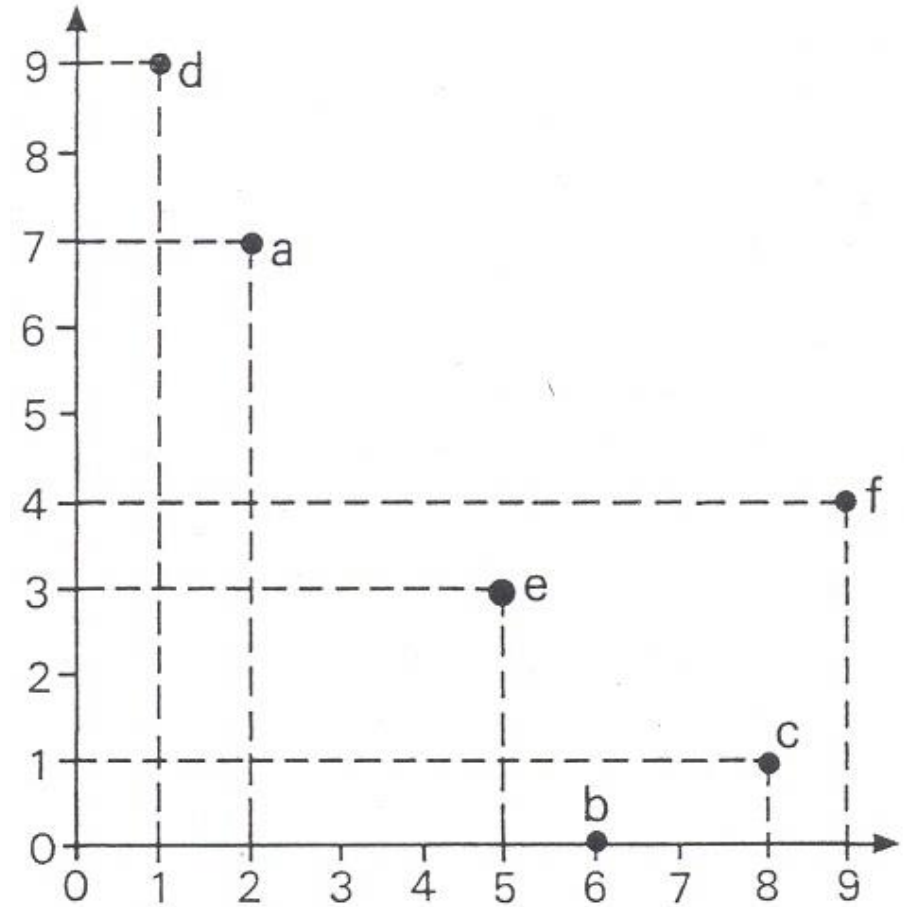


¿Os atrevéis con dos especies?

Localidad	Sp1	Sp2
a	2	7
b	6	0
c	8	1
d	1	9
e	5	3
f	9	4

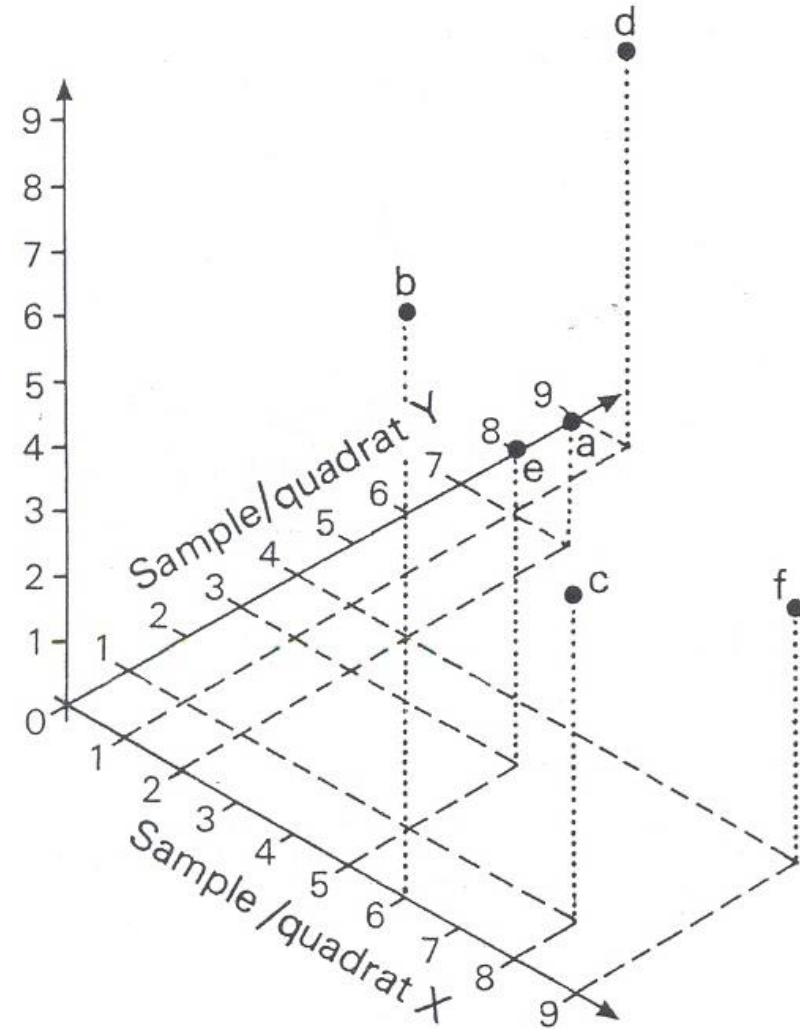
Tampoco fué tan difícil

Localidad	Sp1	Sp2
a	2	7
b	6	0
c	8	1
d	1	9
e	5	3
f	9	4



¿Y con una o dos especies más?

Localidad	Sp1	Sp2	Sp3
a	2	7	2
b	6	0	9
c	8	1	5
d	1	9	6
e	5	3	5
f	9	4	4



Redundancia



Normalmente los datos contienen cierta cantidad de redundancia: Algunas especies son similares en su respuesta a los gradientes ambientales, lo que duplica la información de la variación.



Por eso en los análisis de ordenación se realiza una reducción de información, encaminada a eliminar esta redundancia.



Ruido

- Dos muestras, incluso en la misma comunidad, es poco probable que sean idénticas:
 - diferencias micro espaciales en factores locales;
 - variaciones locales en presiones de origen biótico como incendios o pastoreo;
 - la distribución al azar de los individuos y otros procesos estocásticos (dispersión, etcétera);
 - errores en la toma de datos de la abundancia.
- Aportan variación, pero pequeña y de poco interés, especialmente en comparación con la variación entre muestras de distintas comunidades.

3 tipos de variación



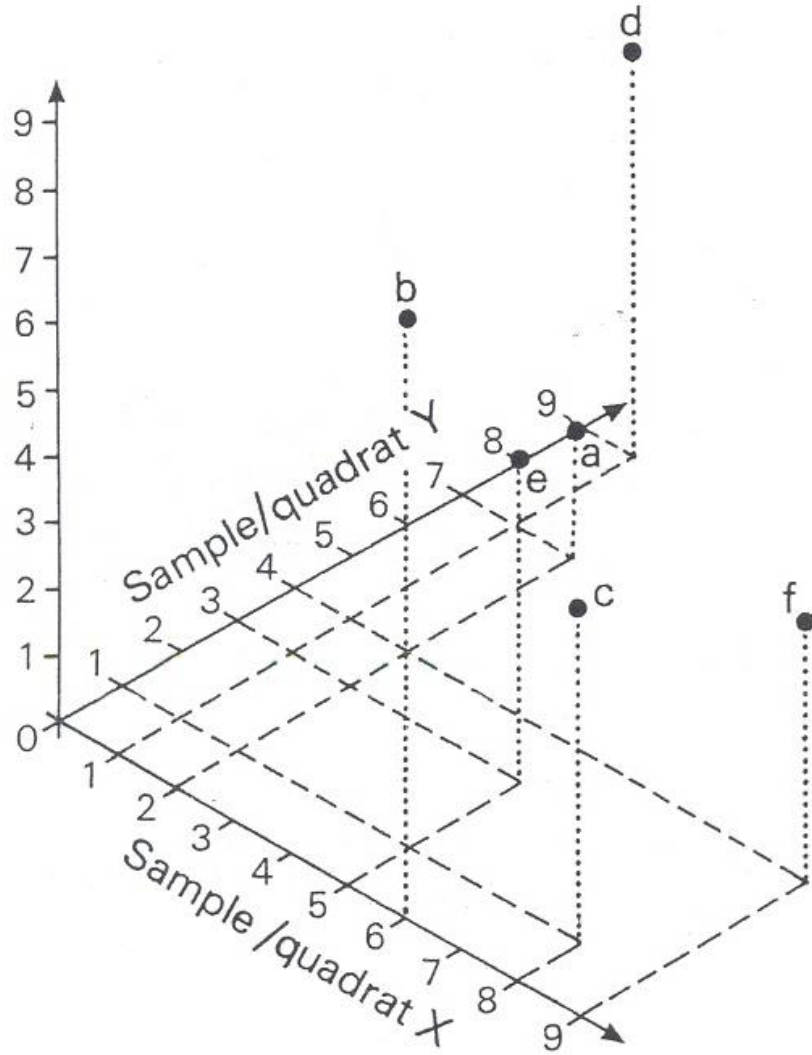
Variabilidad importante, permite explicar y comprender la relación ambiente/comunidad



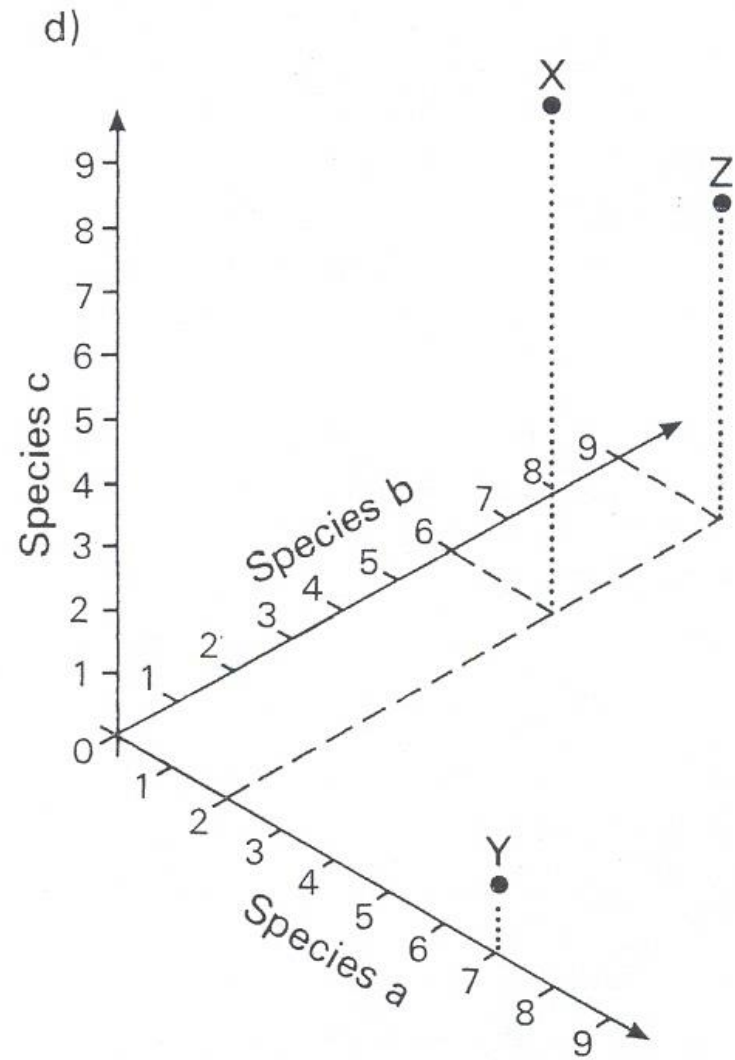
Variabilidad redundante, aparece por acumulación de especies con respuestas parecidas a las condiciones ambientales



Variabilidad de menor detalle, y que puede interferir con la visión general



-
- Estudio de las relaciones que existen entre los objetos (las muestras)

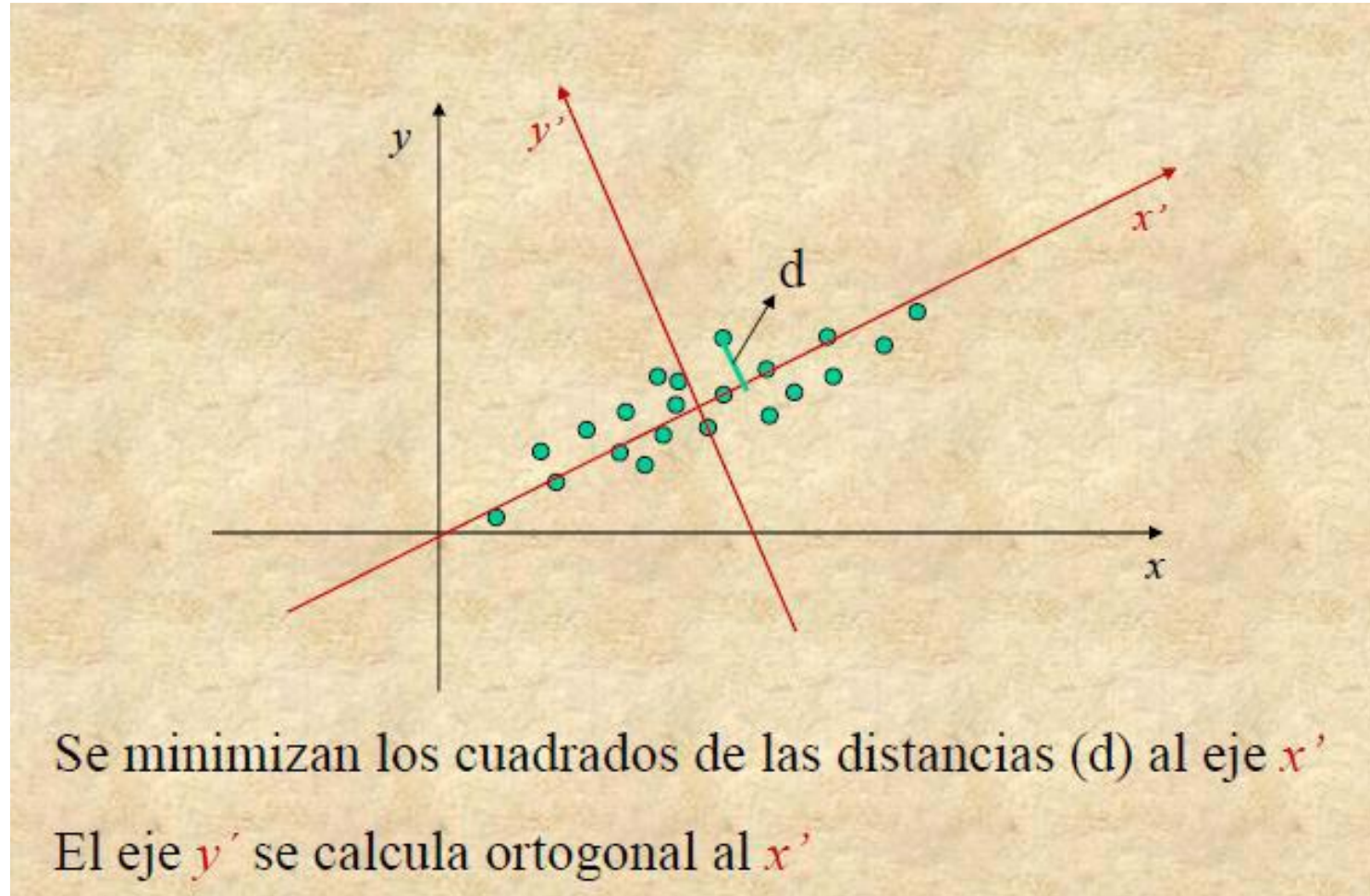


-
- Estudio de las relaciones que existen entre las variables o descriptores (las especies)

	(a) Basado en datos brutos (aproximación clásica)		(b) Basado en transformaciones	(c) Basado en distancias
	Lineal	Unimodal		
(1) Unconstrained (indirecto)	Análisis de Componentes Principales (PCA)	Análisis de Correspondencia (CA) Análisis de Correspondencia Sin Tendencia (DCA)	Análisis de Componentes Principales con transformación (tb-PCA)	Non-metric Multidimensional Scaling (NMDS)
(2) Constrained (directo, canónico)	Análisis de Redundancia (RDA)	Análisis de Correspondencia Canónico (CCA)	Análisis de Redundancia con transformación (tb-RDA)	Análisis de Redundancia basado en distancias (db-RDA)

tb-: Los datos de composición se modifican usando la transformación de Hellinger, resultando en distancias de Hellinger que son más adecuadas para datos ecológicos por que son asimétricas (no les afectan los dobles ceros)

PCA



PCA

- Se aplica sólo con datos cuantitativos
- No es necesario establecer jerarquías ni comprobar la normalidad
- Las variables originales deben estar correlacionadas
 - Si no, el análisis no tiene sentido
- Se usa la varianza para medir la cantidad de información incorporada en cada componente
- Los ejes se ordenan de mayor a menor varianza

	x1	x2
1	72	70
2	67	69
3	67	70
4	75	69
5	70	70
6	74	71
7	71	70
8	60	70
9	69	70
10	75	69

media	70.00	69.80
varianza	21.11	0.40

Genera un componente por cada variable (especies o muestras) en el conjunto de datos

- Hasta que toda la variabilidad original queda recogida

Sólo unas pocas componentes recojan la mayor parte de la información (variabilidad) de los datos

- Si las variables no están correlacionadas, esto no ocurre y por eso no tiene sentido el análisis

Hay diferentes criterios para decidir el número de componentes a retener:

- Broken stick

PCA - Debilidades

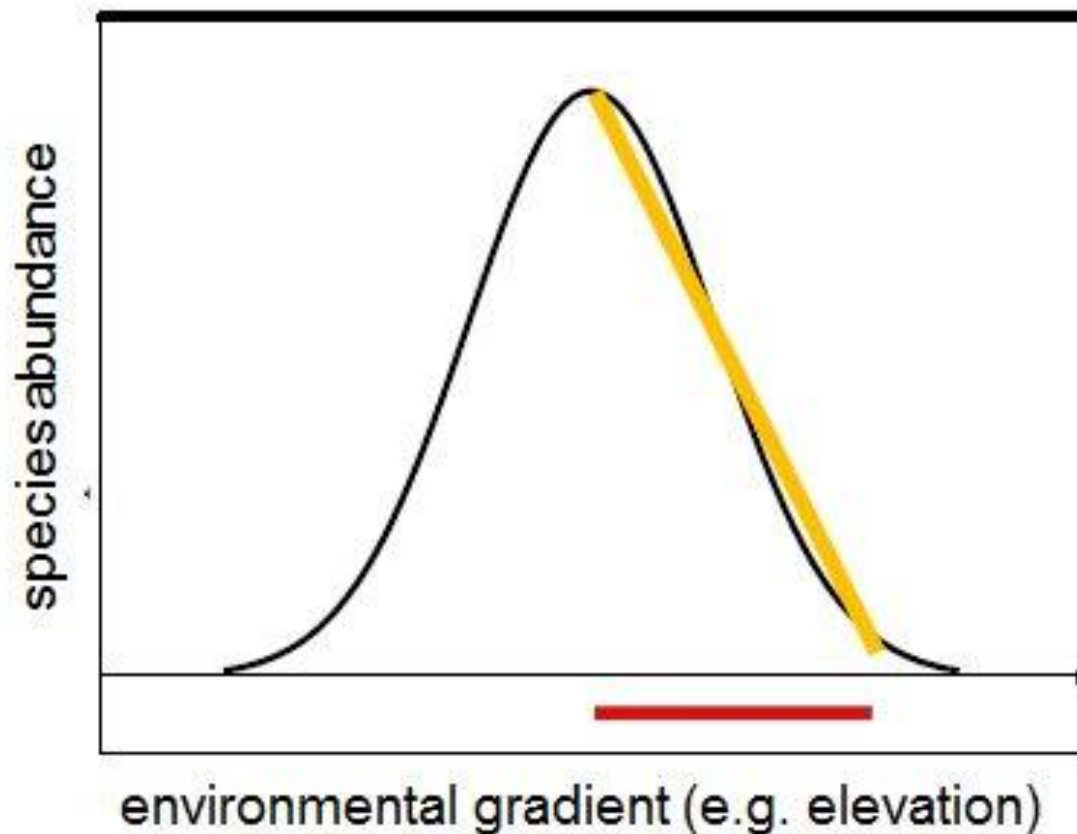
Define la disimilaridad sólo con la distancia Euclídea

- La distancia euclídea tiene problemas cuando hay especies ausentes en dos comunidades (dobles ceros)

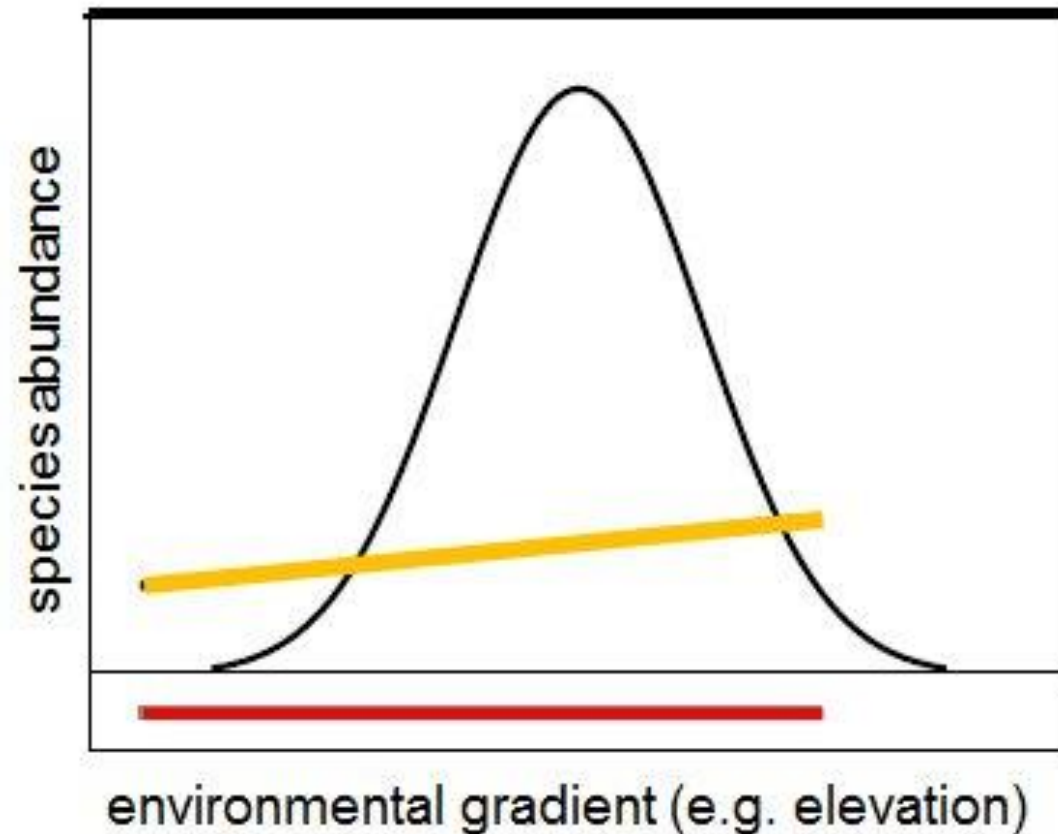
No explica bien las relaciones no lineales (curvilíneas, etc.) entre variables.

- Por tanto, sólo sirve para estudiar gradientes ambientales relativamente homogéneos, sin grandes cambios en la composición.

short ecological gradient

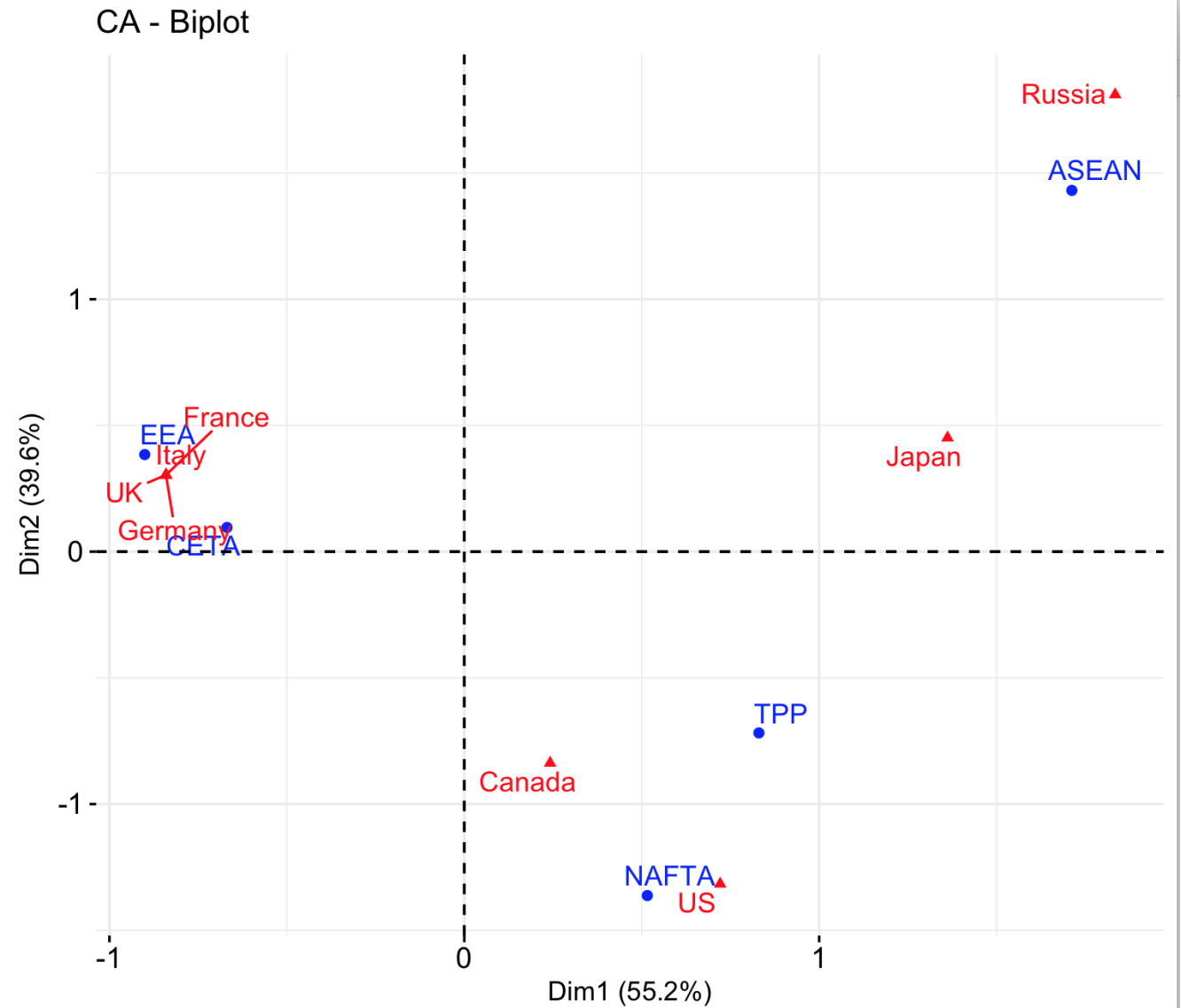


long ecological gradient



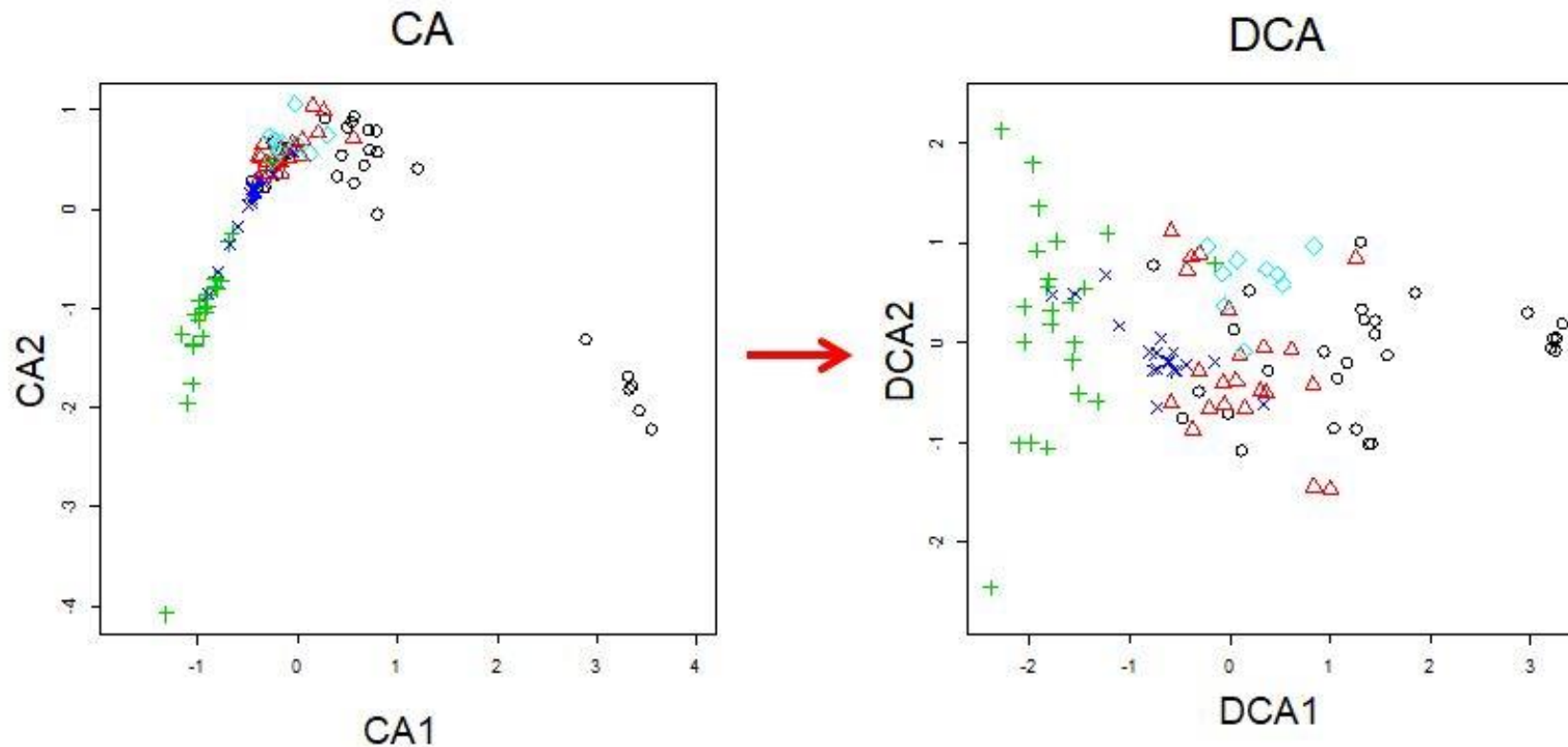
Análisis de Correspondencias (CA)

- Se basa en Chi cuadrado
- No sufre de simetría (doble cero)
- Sufre de arqueado de los gráficos, por falta de linealidad entre el primer componente y el resto

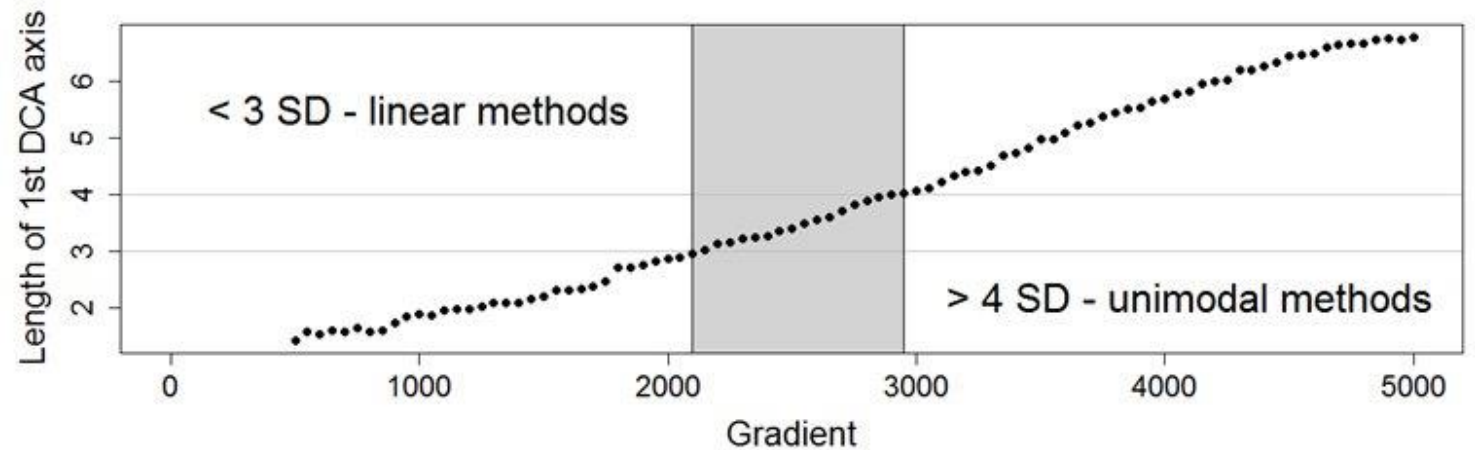
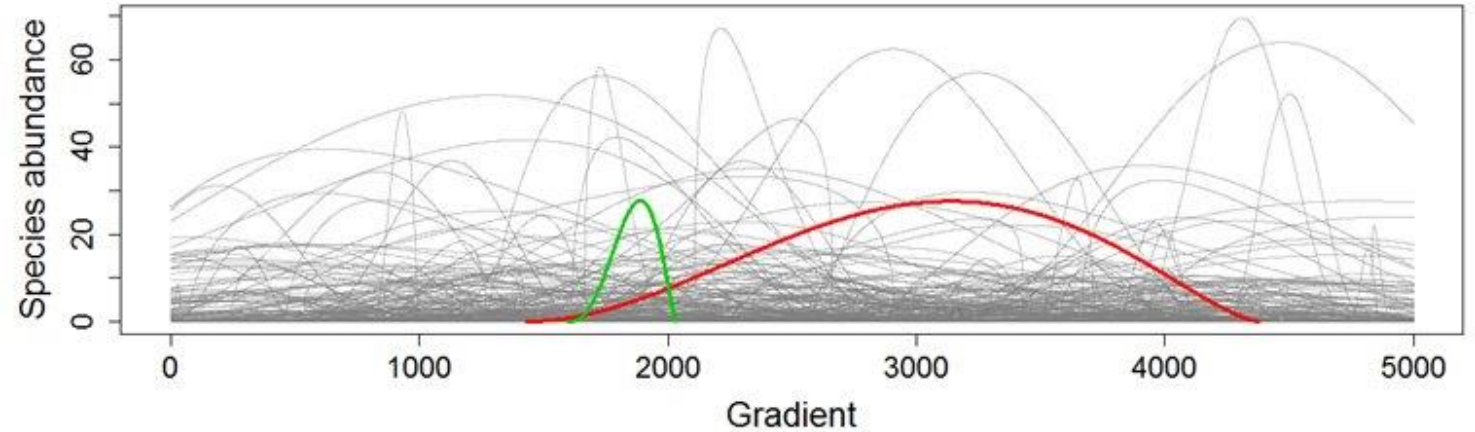


Análisis de Correspondencia Sin Tendencia (DCA)

- Surge para corregir el arqueado de los gráficos de CA.
- Lo hace segmentando el primer eje y “linealizando” a la fuerza. Demasiado tosco y mucha gente no lo recomienda.
- No obstante, es útil para decidir si es necesario una aproximación lineal o una unimodal (PCA o CA).
- Una alternativa es usar las aproximaciones basadas en transformaciones (dt-PCA).



DCA y DS del primer componente



Métodos de ordenación directos

Incorporan la información ambiental para “condicionar” la ordenación en base a dichos gradientes

RDA

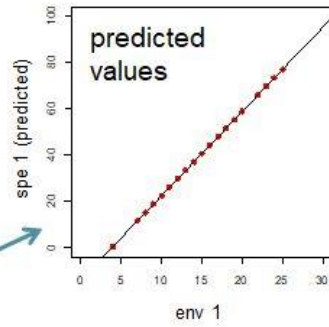
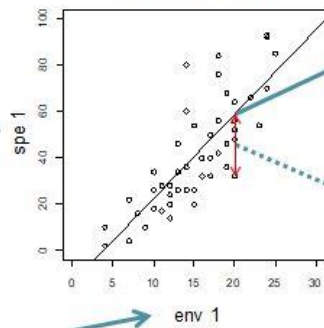
sample × species matrix

	spe1	spe2	spe3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

	env 1
sam 1	
sam 2	
sam 3	
sam 4	
sam 5	
sam 6	
sam 7	

matrix of environmental variables
(single variable in this case)

regression of species abundances on env. variable

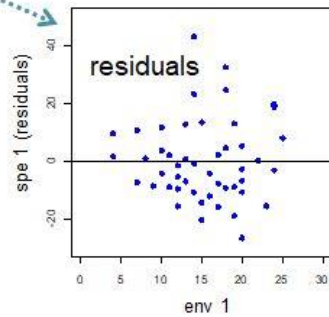


matrix of predicted values

	spe1	spe2	spe3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

	spe1	spe2	spe3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

matrix of residuals



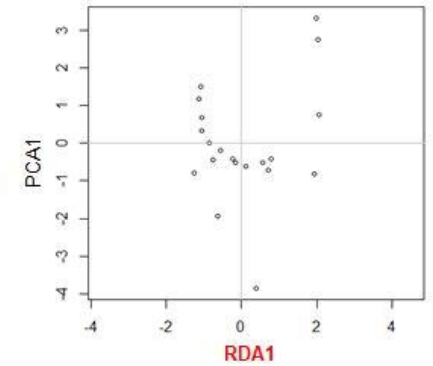
matrix of predicted values

	spe1	spe2	spe3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

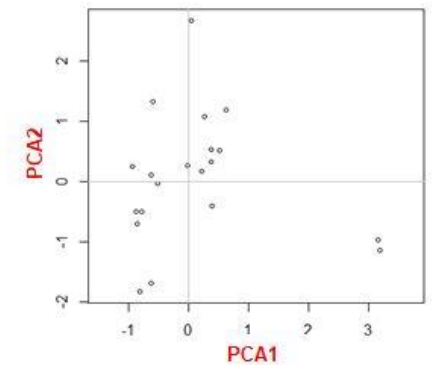
	spe1	spe2	spe3
sam 1			
sam 2			
sam 3			
sam 4			
sam 5			
sam 6			
sam 7			

matrix of residuals

PCA on predicted values



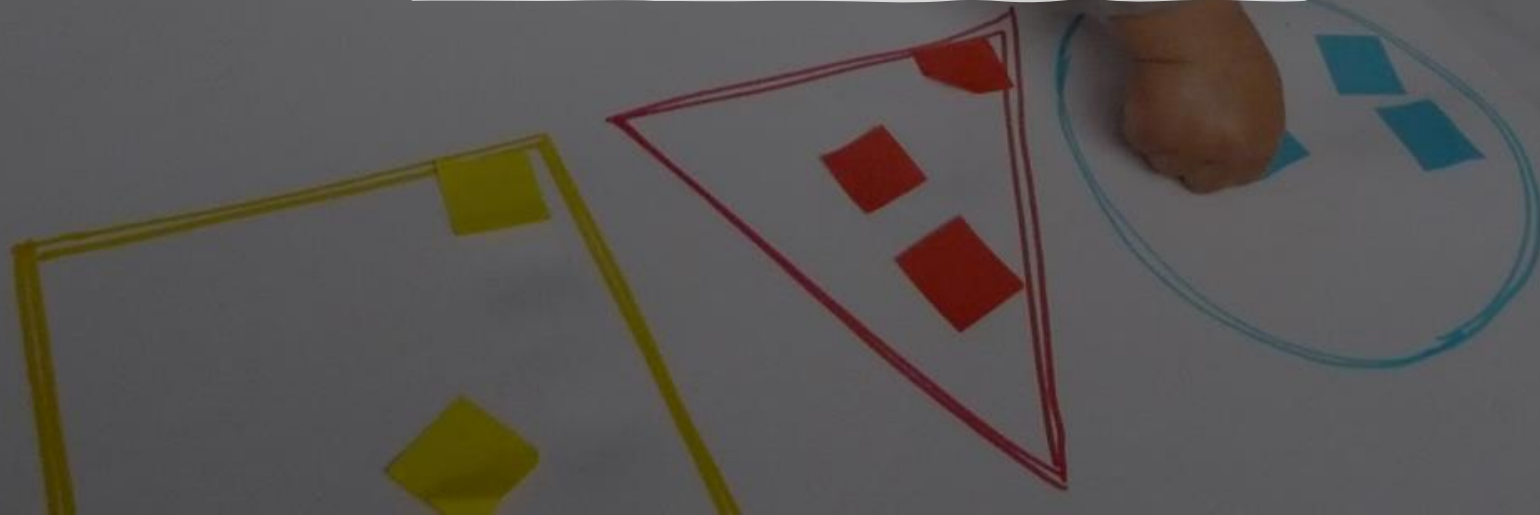
PCA on residuals



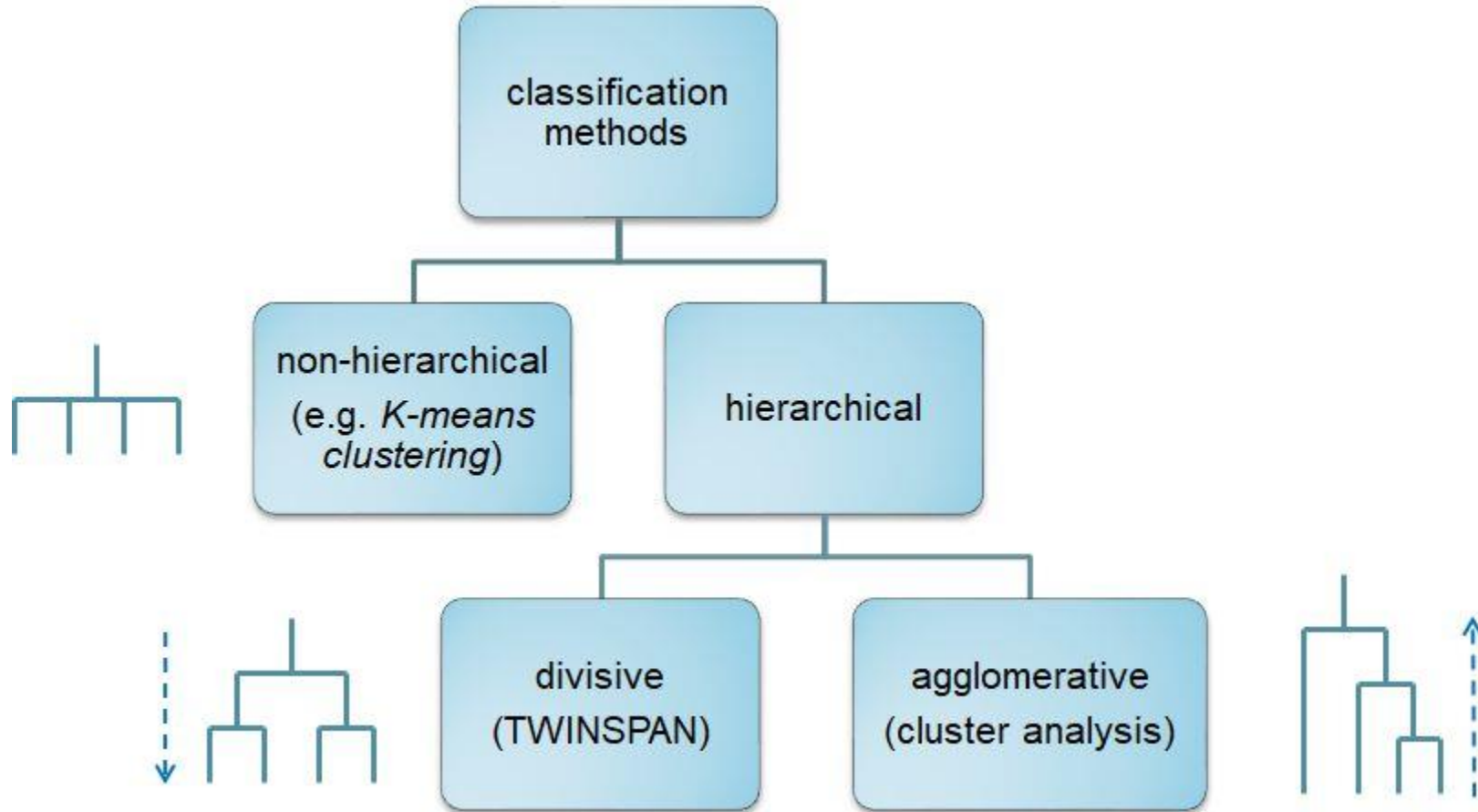
constrained ordination axes

unconstrained ordination axes

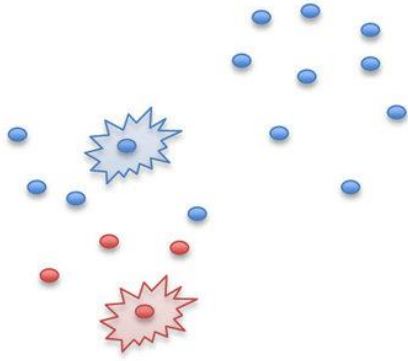
Análisis de agrupación



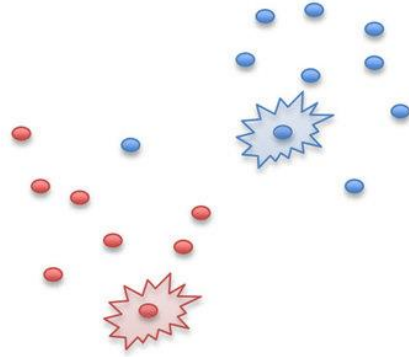
Tipos de análisis de clasificación



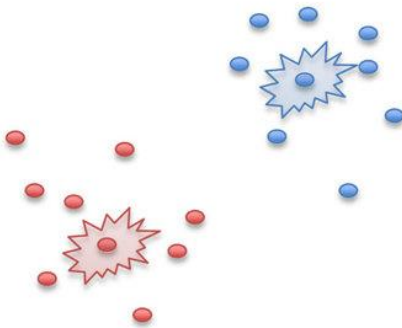
Initial Seeding



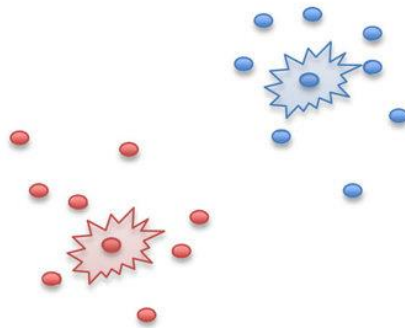
After Round 1



After Round 2



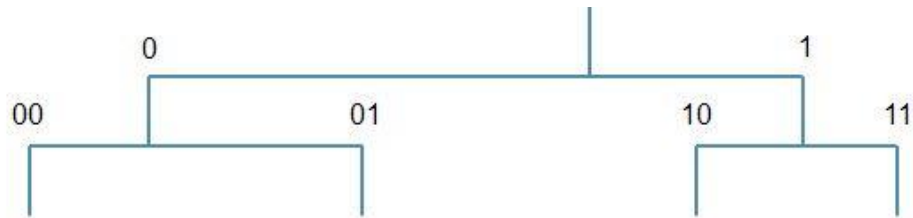
Final



K-means

- El número de grupos se establecen a priori por el investigador
- Proceso iterativo
 - Sensible a soluciones locales
 - Mejor repetir varias veces

TWO-WAY ORDERED TABLE



	112333344567788999	1122222233555555556666666677777788888999	111222334444778899	113344448
44 Galespe6	-22-22--112--2---	1-2-22-222-2222---	2--1---221-2-1-2--2-2-1---	22-----12-----2-1-
48 Impagla6	5222242--22522222221-	1-12-2-1-----1-----	2-----2-----1-1-----	1-1-----1-----
61 Seneova6	-1---2---21--22--2---	1-----222-22222222221-	222222-2-----1-----	2-----2-----
86 Urtidio6	2222222222522-2-2221-	222--222--222-1222---	2-2--22-2-2-212	22-----2-----2-----
94 Galemon6	-22--422-2--2-2---	22--52-22-52---2225242-	22-2-2---25-----	2-----2-----
113 Sambnig4	-22-2222-2-22-----	2-2-2522-2-2--2--255---	25554-22222--2-22	22--2-2-1-----2-2-----
121 Asareur6	--2222-2-22-2222-22-2-	2222222222-22--2-2-2-	2-----2-----	2-1-----2-1-----
124 Dryofil6	--2--2-222--222222242-	22222222224222222222-	22222242325433232-	2-1-----2-2-22-----
149 Dryodil6	--21-----12222--2-22222-	-----2-22-22222-22--22-	22221--211-----1-----	1-----1-----
215 Oxalace6	-----22-2-22222212-----	2-----2-22222222222222-	222-22-2212-----	2-----2-----
227 Abiealb1	-----2-----22-----	22244255222255222--	5533455-3-----	2-2-----2-2-----
8 Coryave4	-225544-55--445455-	22-5-22222-2222222--	2552--52224-24343353422	222--2-2222--22-2-2252422-
16 Galebif6	2--2-2-1-----22--22-1-1221-	122--12-1-1--2221--	221221-----22222--	2-2-2-12-1-----
55 Gerarob6	2-2--222-2--22-22122-	22222-2-2--22222-22-2--	1--2222-1--42-----	2-----2-----2-1-22--22
56 Impapar6	-2222222--22222-3222222442222-	1-22--2222--2222-222-22-2-222	122222--2-----22--2-2-22222--	01
65 Tilicor1	-545--4--22-45-54-	255-4-522-525522-----	222--24242--5-4-453-5-	2-----3-----2-2-25445252
89 Acerpse7	-2-222-----21-2122222-	22--2-2--2-----22-2-1--	222--22-22--1-----122--	22--2--2--1
151 Stelhol6	-222222-2--222-22-----	22222-----2--2--2-----	2-----2-----	2222--221
26 Poa nem6	2-22-222--22-2-2-22--2-	22-22222-----22222-2222-	2-2-22--22-----2242222224422222	10
33 Solivir6	1-1-----2--11222-	2221-1-----112-----	12-2--21-1-----2-211-----	222-21-1-----1-221
34 Tilicor4	-----222-2-2-----	2222222-25-----22-----	22-----2222--22-2-2--	25222222-
63 Polyvul6	-----22-2-22--22-----	1-1-----221-1--122-----	24222--2-2-----1-222222--	1-----111222
126 Sorbauc7	-----2-----2--2-1-----	11-2-1222-222-1--2-21222-2-	21-2--122--2222--	21--2
128 Fagusyl1	-----2-----2-----	5-22-----5-----52-----	55-5--5-----422-5453--	2-----2
195 Calaarue6	-----2-----2-22-1-22--	222-1-2-----2222--	2222-233-25-22--2-2--	224-122225-----2-12
21 Luzuluz6	-2-----2-----222-2-2--	2222-----1-2-222-222-2-2-23-	222-222222-2--222-2-2-22--	22223
24 Pinusyl1	-----2-----2-21-----	2-----2-----2-----	22-122224224332-2-222222--	2-2
41 Avenfle6	--2-1--21-1-212--2-22--	222-----222-----	1-1-22-2-2-2-3--222222422222323322-	2-22-----22-3
50 Querpet1	--22-2-45-2-2-----	22--2-2422-2-----	4554-----2445-54545552-----	52-222522525--
162 Convmaej6	-----21-----2--221-2-----	2-----22222-22-----	2-52-2212-22-----22-	2-2-22--2

classification of species

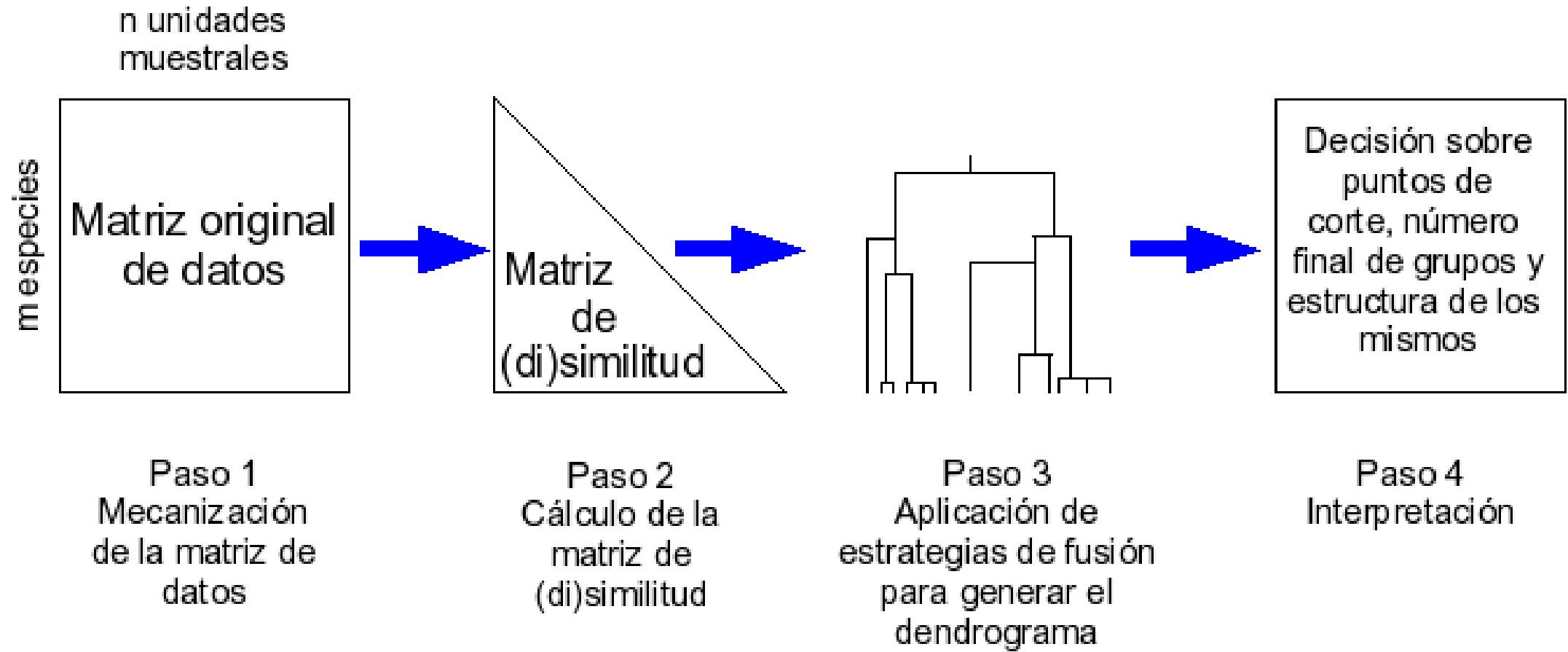
00000000000000000000	00000000000000000000	11111111111111111111	11111111111111111111
00000000000000000000	11111111111111111111	00000000000000000000	11111111111111111111

classification of samples

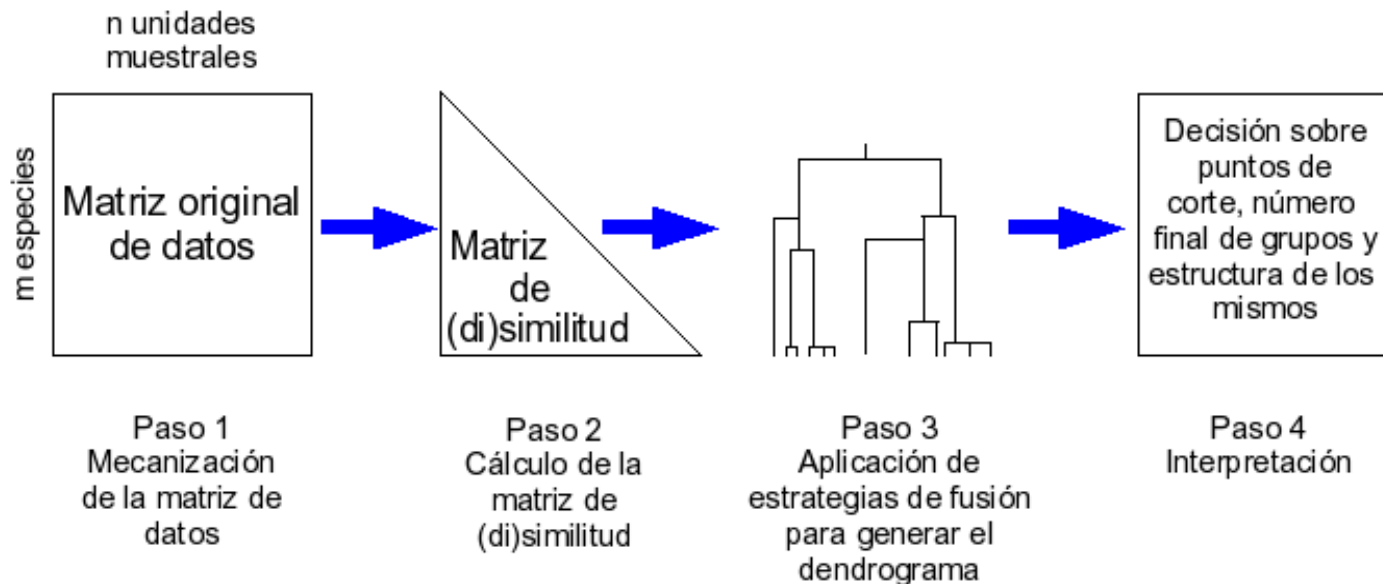
Twinspan

Se basa en la distribución de las especies a lo largo del primer eje de un análisis de ordenación.

Análisis de clasificación jerárquica



Matriz de (di)similitud: índices de betadiversidad



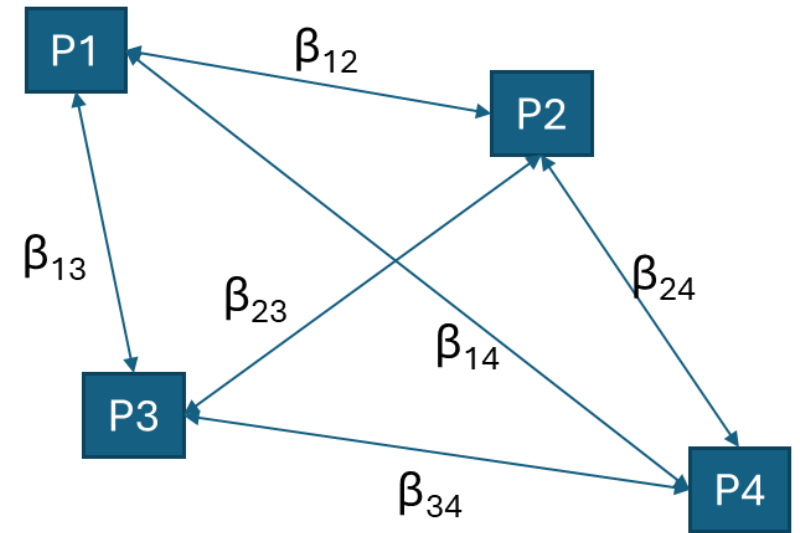
- Jaccard: $S_J = \frac{c}{a+b+c}$
- Sorensen: $S_S = \frac{2c}{a+b}$
- ...
 - a: N^o especies exclusivas del punto 1
 - b: N^o especies exclusivas del punto 2
 - c: N^o especies comunes en ambos puntos

Matriz de similitud del coeficiente de Czekanowski

	1	2	3	4	5	6	7	8	9	10	11	12
1	0.00											
2	0.51	0.00										
3	0.91	0.69	0.00									
4	0.66	0.58	0.87	0.00								
5	0.88	0.67	0.73	0.44	0.00							
6	0.97	0.72	0.75	0.93	0.76	0.00						
7	0.96	0.70	0.60	0.94	0.80	0.69	0.00					
8	0.81	0.75	0.89	0.69	0.80	0.95	0.77	0.00				
9	0.97	0.83	0.86	0.94	0.79	0.83	0.58	0.71	0.00			
10	0.93	0.73	0.79	0.89	0.66	0.79	0.79	0.84	0.60	0.00		
11	0.99	0.66	0.69	0.94	0.76	0.69	0.52	0.87	0.67	0.64	0.00	
12	0.99	0.97	0.97	0.98	0.98	0.96	0.82	0.81	0.69	0.48	0.61	0.00

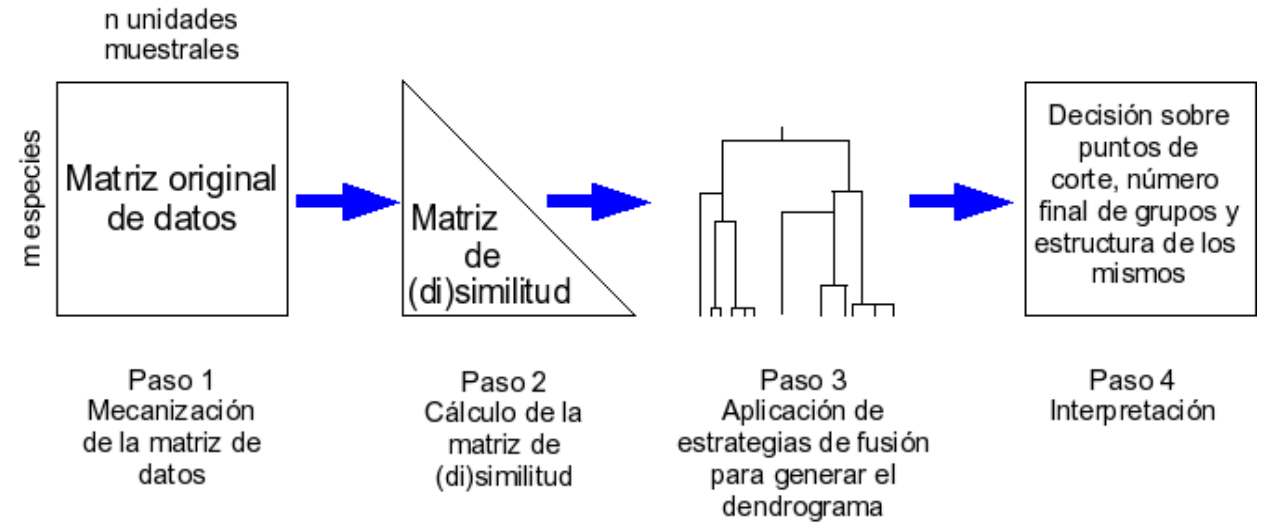
Similitud = 1 - Disimilitud

- Betadiversidad (β)
 - Reemplazo (turnover) de especies entre dos sitios
 - Término ecológico
 - Característico para parejas de puntos
- Interpretación
 - $\beta = 1$: Sitios completamente diferentes
 - $\beta = 0$: Sitios completamente similares
- Similitud = 1 - Disimilitud



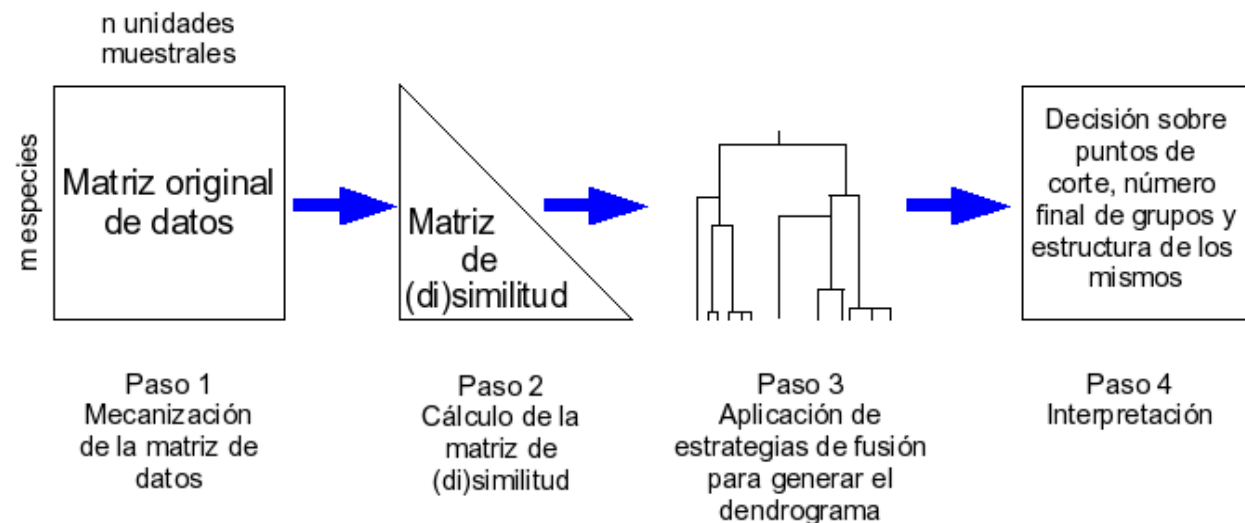
	P1	P2	P3	P4
P1	0	β_{12}	β_{13}	β_{14}
P2	β_{12}	0	β_{23}	β_{24}
P3	β_{13}	β_{23}	0	β_{34}
P4	β_{14}	β_{24}	β_{34}	0

Estrategias para agrupar observaciones



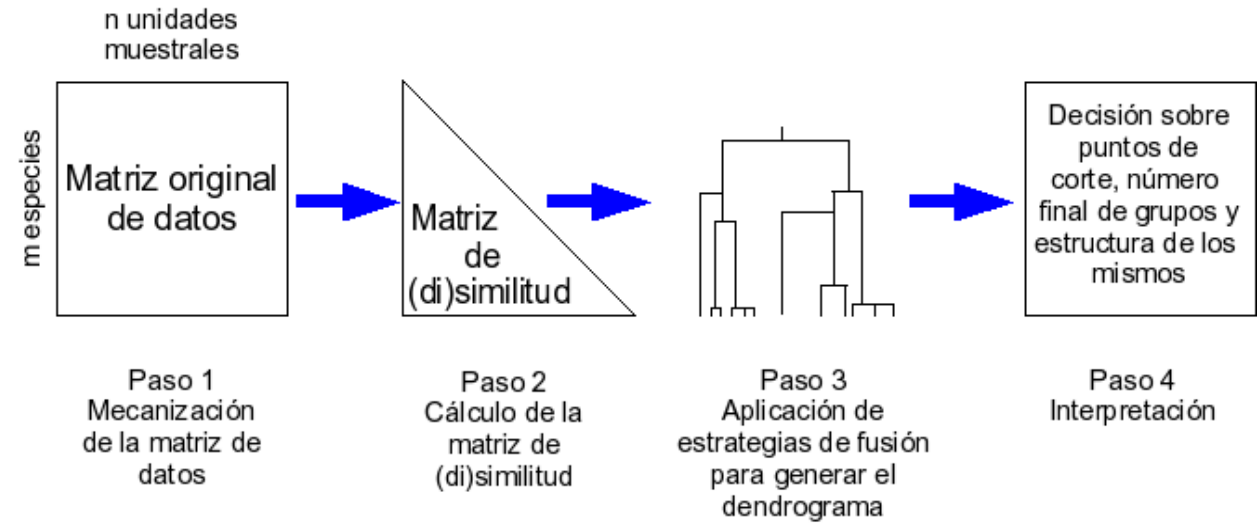
	P1	P2	P3
P1			
P2	0.1		
P3	0.5	0.6	

Estrategias para agrupar observaciones



	P1	P2	P3	P4
P1				
P2	0.1			
P3	0.5	0.6		
P4	0.7	0.8	0.3	

Estrategias para agrupar observaciones



	P1	P2	P3	P4	P5
P1					
P2	0.1				
P3	0.5	0.6			
P4	0.7	0.8	0.3		
P5	0.4	0.6	0.5	0.7	

Estrategias para agrupar observaciones

- Single linkage
- Complete linkage
- Average linkage (UPGMA)
- Centroid linkage

