Clustering

Análisis de segmentación

Prof. Dr. Jose Jacobo Zubcoff
Departamento de Ciencias del Mar y
Biología Aplicada

Análisis Multivariante

Uso de medidas de similaridad/disimilaridad

- En general los conjuntos de datos contienen:
  - Variables morfológicas (e.g. tamaño)
  - Fisiológicas
  - Medidas ambientales fisico-químicas del entorno (e.g. temperatura)
  - Medidas de riqueza de especies, biomasa/abundancia……
- Así como especies (en filas o en columnas)

Análisis multivariante

Samples	Species				
	A	B	C	D	E
X1	5	0	0	7	25
X2	4	0	0	8	25
X3	7	0	0	6	33
Y1	7	6	8	4	3
Y2	5	9	7	3	4
Y3	8	9	9	4	3
Z1	7	1	0	7	4
Z2	6	0	0	8	5
Z3	7	1	0	6	5

Análisis multivariante

Samples	Environmental Variables				
	A	B	C	D	E
X1	5	0	0	7	25
X2	4	0	0	8	25
X3	7	0	0	6	33
Y1	7	6	8	4	3
Y2	5	9	7	3	4
Y3	8	9	9	4	3
Z1	7	1	0	7	4
Z2	6	0	0	8	5
Z3	7	1	0	6	5

Análisis multivariante

Samples	Species			Environmental Variables	
	A	B	C	O	P
X1	5	0	0	7	25
X2	4	0	0	8	25
X3	7	0	0	6	33
Y1	7	6	8	4	3
Y2	5	9	7	3	4
Y3	8	9	9	4	3
Z1	7	1	0	7	4
Z2	6	0	0	8	5
Z3	7	1	0	6	5

Análisis Multivariante

Medidas de similaridad

- Concepto de similaridad entre pares de datos
  - Por ej. en términos de comunidades biológicas que contiene cada muestra (fila o columna)

Hay muchas formas de definir la similaridad

- En fc. del peso que se otorgue a algún aspecto de la comunidad
- Valores absolutos: (biomasa/cobertura/…)
- Valores relativos: (biomasa/cobertura/…) estandarizado $n_i/total$ (cuando no son directamente comparables)
- Presencia/Ausencia

Matriz de similaridad

- Depende de la medida de similaridad usada (o disimilaridad en su caso)
- Por ejemplo: el índice de Bray-Curtis (1957)

$$S_{jk} = 100 \left\{ 1 - \frac{\sum_{i=1}^{n} | y_{ij} - y_{ik} |}{\sum_{i=1}^{n} | y_{ij} + y_{ik} |} \right\}$$

Donde *y* representa Abundancia/Cobertura/…
de la fila *i-ésima*

- Si se acerca al 100% son similares completamente

Medidas de similaridad

Indice de Bray-Curtis (1957)

$$S_{jk} = 100 \left\{ 1 - \frac{\sum_{i=1}^{n} | y_{ij} - y_{ik} |}{\sum_{i=1}^{n} | y_{ij} + y_{ik} |} \right\}$$

Samples	Species				
	A	B	C	D	E
X2	4	0	0	8	25
X3	7	0	0	6	33
Y1	7	6	8	4	3

$$S_{(X2,X3)} = 100 \left\{ 1 \frac{3+0+0+2+8}{11+0+0+14+58} \right\} = 84$$

Medidas de similaridad

Indice de Bray-Curtis (1957)

$$S_{jk} = 100 \left\{ 1 - \frac{\sum_{i=1}^{n} | y_{ij} - y_{ik} |}{\sum_{i=1}^{n} | y_{ij} + y_{ik} |} \right\}$$

Samples	Species				
	A	B	C	D	E
X2	4	0	0	8	25
X3	7	0	0	6	33
Y1	7	6	8	4	3

$$S_{(X3,Y1)} = 100 \left\{ 1 - \frac{0+6+8+2+30}{14+6+8+10+36} \right\} = 38$$

Medidas de similaridad

Indice de Bray-Curtis (1957)

$$S_{jk} = 100 \left\{ 1 - \frac{\sum_{i=1}^{n} | y_{ij} - y_{ik} |}{\sum_{i=1}^{n} | y_{ij} + y_{ik} |} \right\}$$

Samples	Species				
	A	B	C	D	E
X2	4	0	0	8	25
X3	7	0	0	6	33
Y1	7	6	8	4	3

$$S_{(X2,X3)} = 100 \left\{ 1 - \frac{3+0+0+2+8}{11+0+0+14+58} \right\} = 84$$

$$S_{(X3,Y1)} = 100 \left\{ 1 - \frac{0+6+8+2+30}{14+6+8+10+36} \right\} = 38$$

Medidas de similaridad

- Matriz de Similaridad: Bray-Curtis (especies)

	X1	X2	X3	Y1	Y2	Y3	Z1	Z2
X1								
X2	97							
X3	87	84						
Y1	37	34	38					
Y2	37	34	32	86				
Y3	34	31	35	92	89			
Z1	57	54	52	64	55	58		
Z2	61	61	52	55	51	50	89	
Z3	57	54	55	64	55	58	95	89

Medidas de similaridad

El porqué del “éxito” del índice de Bray-Curtis

- Cuando dos muestras son iguales toma el valor 100 (*)
- Si no tienen especies en común vale 0 (**)
- No le afectan los cambios en la unidad de medida (*)
- No se ve afectado si se incluyen/excluyen especies ausentes (**)
- No se ve afectado por la inclusión de mas muestras(**)
- Detecta diferencias en Abundancia Total cuando las Abundancias relativas are idénticas (**)

(*) esto es así para la mayoría de coeficientes

(**) esto lo fallan la mayoría de coeficientes

Medidas de similaridad

- Transformaciones de datos

- Cuando similaridades “parecen” muy bajas
- Cuando las especies mas frecuentes tienen demasiado “peso” en el conjunto de datos

Estrategia:
- Transformar ANTES de aplicar Bray-Curtis
- Eliminar especies “raras” ANTES de Bray-Curtis

Por ejemplo: una transf. cuadrática (raiz cuadrada) “suaviza” el peso demasiado elevado de especie muy frecuentes.

- Se puede aplicar doble transformación en los datos (doble cuadrática), etc.

Clustering

- Se pretende encontrar grupos de comportamiento homogéneo
- Además, representar de una manera visual la similaridad (o disimilaridad)
  - Puede ser entre diferentes sitios
  - Entre diferentes tiempos de un mismo sitio
- Cuando encuentra diferencias entre sitios/tiempos genera un nuevo clúster

Clustering

Tipos de clustering

- Jerárquicos (iterativo)
- Ténicas de optimización (grupos mutuamente excluyentes k=pre-establecido)
- Modo-búsqueda (basado en la densidad de la muestra)
- …

Clustering jerárquico

Representación gráfica: Dendogramas

Ilustra cómo se van haciendo las subdivisiones o los agrupamientos, etapa a etapa, o por nivel

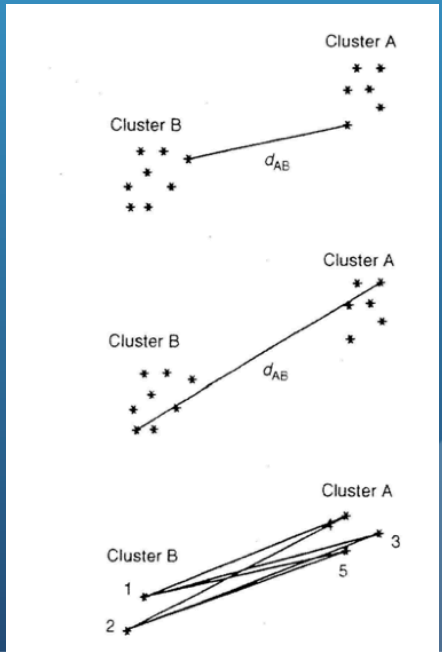
Utiliza distintos métodos aglomerativos con diferentes métodos de unión (linkage methods).

Los más importantes son:

(i) Mínima distancia o vecino más próximo (nearest neightbor).

(i) Máxima distancia o vecino más lejano (farthest neightbor).

(i) Distancia media (average distance).



Clustering

Dendogramas

Suponemos unos datos originales, y calculamos la matriz de similitud

	Year	10	11	12	13
	Sample	1	2	3	4
	Sp1.	1.7	0	0	0
	Sp2.	2.1	0	0	1.3
Species	Sp3.	1.7	2.5	0	1.8
	Sp4.	0	1.9	3.5	1.7
	Sp5.	0	3.4	4.3	1.2
	Sp6.	0	0	0	0

Sample	1	2	3	4
1	-			
2	25.6	-		
3	0.0	67.9	-	
4	52.2	**68.1**	42.0	-

Máximo

Group average link	Sample	1	2&4	3
	1	-		
	2&4	**38.9**	-	
	3	0.0	**55.0**	-

S(2&4,1)=(25.6+52.2)/2 = 38.9

S(2&3&4,1)=(38.9*2+0*1)/3 = 25.9

Sample	1	2&3&4
1	-	
2&3&4	**25.9**	-

Ejemplo de cálculo de matriz de:

(i) Máxima distancia (farthest neightbor)(single linkage)

(ii) Mínima distancia o vecino más próximo (nearest neightbor) (Complete linkage)

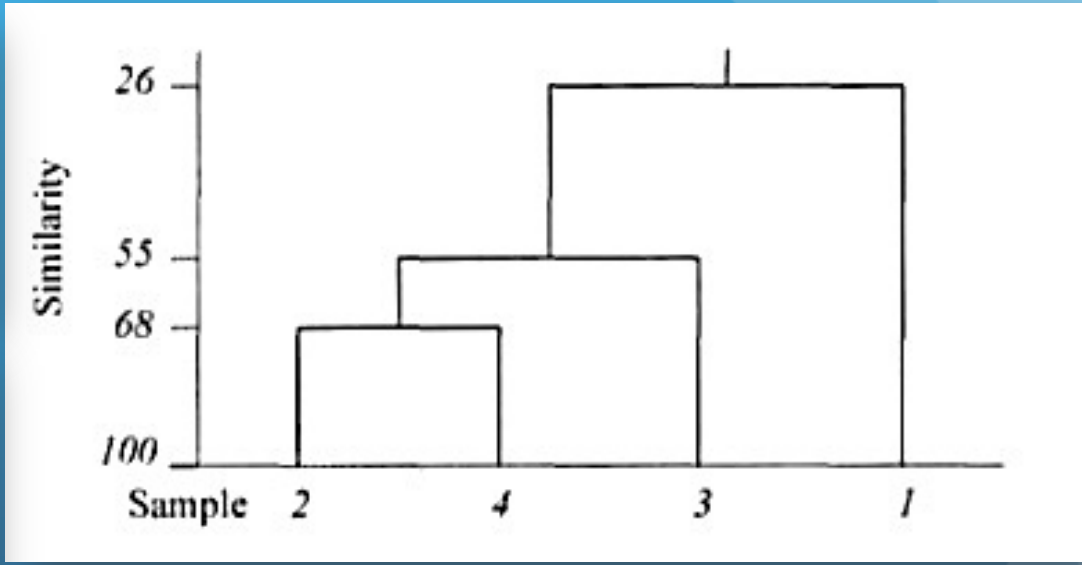
(iii) Distancia promedio (**group average link**)

Clustering: dendograma

Sample	1	2	3	4
1	-			
2	25.6	-		
3	0.0	67.9	-	
4	52.2	**68.1**	42.0	-

Sample	1	2&4	3
1	-		
2&4	**38.9**	-	
3	0.0	**55.0**	-

Sample	1	2&3&4
1	-	
2&3&4	**25.9**	-



Clustering

- Notas finales

- No funciona bien cuando los objetos están próximos
- Se obtienen dendogramas similares si se utiliza la distancia máxima, o la distancia media
- Las fuentes de error y variación no afectan a los métodos jerárquicos
- Gran sensibilidad a observaciones anómalas o outliers
- Si un objeto se ha colocado erróneamente en un grupo al principio del proceso, ya no se puede arreglar en una etapa posterior
- El orden de representación no es único

Consejo de utilización

- Usar varias distancias o similitudes con los mismos objetos y observar si se mantienen los mismos clusters o grupos
- Eliminar especies raras u outliers
- Reordenar eje x *a posteriori* para mejorar visualización