

Clase 18 Introducción análisis multivariante

Diplomado en Análisis de datos con R para la Acuicultura

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

11 June 2022

PLAN DE LA CLASE

1.- Introducción

- ▶ ¿Qué son los análisis multivariantes?.
- ▶ Estudio de caso: Fauna béntica como indicador de calidad ambiental.
- ▶ Matrices de distancia.
- ▶ Análisis de cluster: jerárquico y no jerárquico.

2). Práctica con R y Rstudio cloud.

- ▶ Matriz de distancia: cálculo con R.
- ▶ Análisis de cluster.

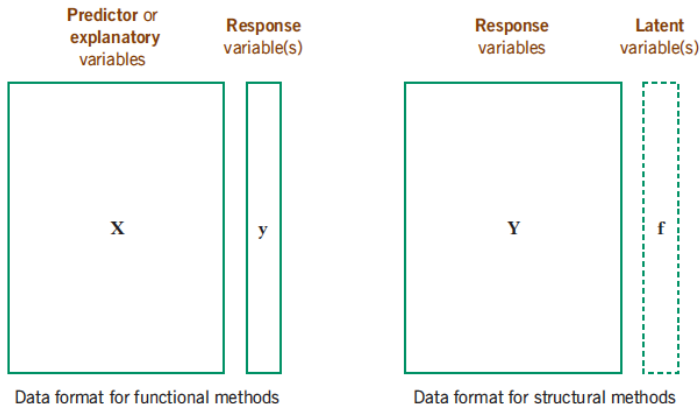
INTRODUCCIÓN ANÁLISIS MULTIVARIANTE

¿Qué son los análisis multivariantes?

Conjunto diverso de métodos estadísticos que estudian y examinan el efecto simultáneo de múltiples variables.

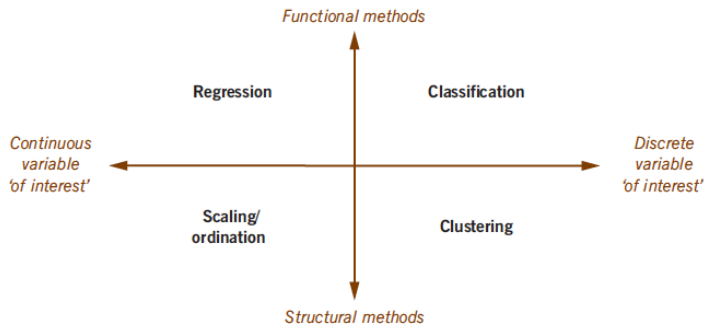
a	b	c	d	e	Depth	Pollution	Temperature
0	2	9	14	2	72	4.8	3.5
26	4	13	11	0	75	2.8	2.5
0	10	9	8	0	59	5.4	2.7
0	0	15	3	0	64	8.2	2.9
13	5	3	10	7	61	3.9	3.1
31	21	13	16	5	94	2.6	3.5

TIPOS DE MÉTODOS MULTIVARIANTES



Fuente: Multivariate Statistic, 2014

MÉTODOS MULTIVARIANTES SEGÚN TIPO DE VARIABLE



Fuente: Multivariate Statistic, 2014

ESTUDIO CASO: BENTOS Y AMBIENTE



Fuente FOTO: Mesa Merelo. 2014

ESTUDIO CASO: DATOS MULTIVARIADOS

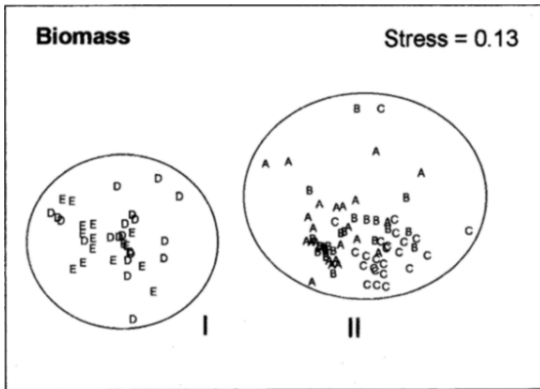
- ▶ Análisis del bentos.
- ▶ 21 variables respuesta (especies + biomasa) por tipo de cultivo y sitio.

Especies	POLICULTIVO			MONOCULTIVO	
	A	B	C	D	E
Nereis diversicolor (Po)	79	46	15	62	70
Syllides sp. (Po)	0	0	0	3	15
Polydora sp. (Po)	4	5	3	11	9
Streblospio srhubsolii (Po)	62	25	0	115	105
Alkmaria sp. (Po)	0	0	0	14	207
Capitella capitata (Po)	7	17	75	2	12

Fuente: Drake and Arias. 1997

ESTUDIO CASO: ANÁLISIS DE CLUSTER

- ▶ Análisis de cluster revela 2 grupos asociados al tipo de cultivo.



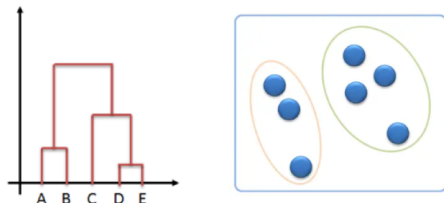
Fuente: De los Ríos-Escalante et al. 2013

ANÁLISIS DE CLUSTER

¿Qué son?: Herramientas de exploración que permiten agrupar y visualizar datos multivariados con base a su similitud (matriz de distancia).

Jerárquico: Los grupos se fusionan sucesivamente siguiendo una jerarquía de similitud (mayor a menor).

No jerárquico: Se forman grupos homogéneos sin establecer jerarquía entre ellos.



Fuente: Multivariate Statistic, 2014

VENTAJAS Y DESVENTAJAS

Tipo	Ventajas
Jerárquico	No requiere especificar grupos al inicio
No jerárquico	Útil cuando existen muchos elementos

Tipo	desventajas
Jerárquico	Difícil decidir que grupos son relevantes y cuales no.
Jerárquico	Difícil de interpretar cuando existen muchos elementos.
No jerárquico	El número de cluster que se define al inicio, podría no ser el adecuado.

ANÁLISIS JERÁRQUICO: MÉTODO

¿Qué hace el algoritmo estándar?

1. Agrupa dos elementos por su similitud (distancia).
2. Recalcula la matriz de distancia (muchas opciones).
3. Vuelve a punto 1.
4. Finaliza cuando todos los elementos han sido asignados a cluster.

¿Cómo recalculo la matriz?

1. Método de distancia máxima (vecino más lejano).
2. Método de distancia mínima (vecino más próximo).
3. Método UPGMA (unweighted Pair-group arithmetic averages).

MATRIZ DE DISTANCIA O SIMILARIDAD

¿Qué es y para que sirven?

- Las matrices de distancia o similaridad están en la base de todos los análisis multivariados de estructura.

Algunas consideraciones

- Las matrices de distancia se pueden elaborar tanto para variables cuantitativas continuas, como discretas.
 - ▶ Debido a que las variables pueden tener diferente escala o magnitud es necesario muchas veces transformar o estandarizar las variables antes de calcular las matrices de distancia.
 - ▶ Cuando una variable tiene muchos ceros también es conveniente transformarla.

TIPOS DE MATRICES DE DISTANCIA

- **Euclideana:** Para variables cuantitativas continuas.

Con base en el teorema de pitágoras

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

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

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

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

- **No euclideana:** Para variables cuantitativas discretas.

- a) Bray-Curtis (datos de conteo).
- b) Jacard (binarias).

EJEMPLO ESTUDIO DIVERSIDAD BENTOS

- ▶ ¿Cuán similares son las muestras entre si?
- ▶ ¿Qué muestras pertenecen a un mismo grupos (variable latente)?

SAMPLES	SPECIES									
	<i>sp1</i>	<i>sp2</i>	<i>sp3</i>	<i>sp4</i>	<i>sp5</i>	<i>sp6</i>	<i>sp7</i>	<i>sp8</i>	<i>sp9</i>	<i>sp10</i>
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

Fuente: Multivariate Statistic, 2014

INDICE DE JACARD

Índice de Similitud de Jaccard se usa para expresar el grado en que dos muestras son semejantes por las especies presentes en ellas.

- ▶ Co-presencias (a)
- ▶ Co-ausencias (d)
- ▶ No coincidentes ($b + c$)

		Sample 2		
		1	0	
Sample 1	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

CALCULE INDICE DE JACARD

Jaccard index dissimilarity:

$$\frac{b+c}{a+b+c} = 1 - \frac{a}{a+b+c}$$

Sitio	<i>sp1</i>	<i>sp2</i>	<i>sp3</i>	<i>sp4</i>	<i>sp5</i>	<i>sp6</i>	<i>sp7</i>	<i>sp8</i>	<i>sp9</i>	<i>sp10</i>
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1

Sitio	<i>sp1</i>	<i>sp2</i>	<i>sp3</i>	<i>sp4</i>	<i>sp5</i>	<i>sp6</i>	<i>sp7</i>	<i>sp8</i>	<i>sp9</i>	<i>sp10</i>
A	1	1	1	0	1	0	0	1	1	1
F	0	1	0	1	1	0	0	0	0	1

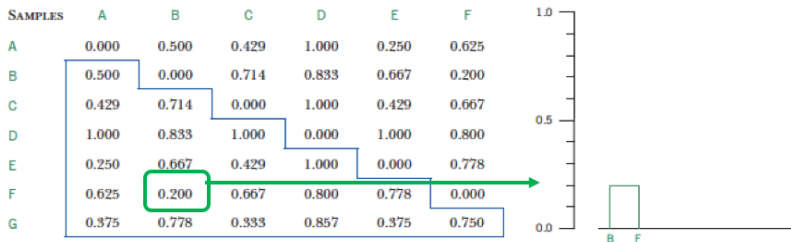
Sitio	<i>sp1</i>	<i>sp2</i>	<i>sp3</i>	<i>sp4</i>	<i>sp5</i>	<i>sp6</i>	<i>sp7</i>	<i>sp8</i>	<i>sp9</i>	<i>sp10</i>
B	1	1	0	1	1	0	0	0	0	1
F	0	1	0	1	1	0	0	0	0	1

MATRIZ DE SIMILARIDAD DE JACARD

SAMPLES	A	B	C	D	E	F	G
A	0.000	0.500	0.429	1.000	0.250	0.625	0.375
B	0.500	0.000	0.714	0.833	0.667	0.200	0.778
C	0.429	0.714	0.000	1.000	0.429	0.667	0.333
D	1.000	0.833	1.000	0.000	1.000	0.800	0.857
E	0.250	0.667	0.429	1.000	0.000	0.778	0.375
F	0.625	0.200	0.667	0.800	0.778	0.000	0.750
G	0.375	0.778	0.333	0.857	0.375	0.750	0.000

AGRUPAMIENTO JERARQUICO: PASO 1

Construcción del primer nodo: Mayor similitud entre B y F



AGRUPAMIENTO JERARQUICO: PASO 2

SAMPLES	A	B	C	D	E	F	G
A	0.000	0.500	0.429	1.000	0.250	0.625	0.375
B	0.500	0.000	0.714	0.833	0.667	0.200	0.778
C	0.429	0.714	0.000	1.000	0.429	0.667	0.333
D	1.000	0.833	1.000	0.000	1.000	0.800	0.857
E	0.250	0.667	0.429	1.000	0.000	0.778	0.375
F	0.625	0.200	0.667	0.800	0.778	0.000	0.750
G	0.375	0.778	0.333	0.857	0.375	0.750	0.000

**Construcción
nueva matriz
usando
método de
distancia
máxima.**

SAMPLES	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

**(B-F) -A
(B-F) -C
(B-F) -D
(B-F) -E
(B-F) -G**

AGRUPAMIENTO JERARQUICO: PASO 2.1

SAMPLES	A	B	C	D	E	F	G
A	0.000	0.500	0.429	1.000	0.250	0.625	0.375
B	0.500	0.000	0.714	0.833	0.667	0.200	0.778
C	0.429	0.714	0.000	1.000	0.429	0.667	0.333
D	1.000	0.833	1.000	0.000	1.000	0.800	0.857
E	0.250	0.667	0.429	1.000	0.000	0.778	0.375
F	0.625	0.200	0.667	0.800	0.778	0.000	0.750
G	0.375	0.778	0.333	0.857	0.375	0.750	0.000

**Construcción
nueva matriz
usando
método de
distancia
máxima.**

SAMPLES	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

**(B-F) -A
(B-F) -C
(B-F) -D
(B-F) -E
(B-F) -G**

AGRUPAMIENTO JERARQUICO: PASO 2.2

SAMPLES	A	B	C	D	E	F	G
A	0.000	0.500	0.429	1.000	0.250	0.625	0.375
B	0.500	0.000	0.714	0.833	0.667	0.200	0.778
C	0.429	0.714	0.000	1.000	0.429	0.667	0.333
D	1.000	0.833	1.000	0.000	1.000	0.800	0.857
E	0.250	0.667	0.429	1.000	0.000	0.778	0.375
F	0.625	0.200	0.667	0.800	0.778	0.000	0.750
G	0.375	0.778	0.333	0.857	0.375	0.750	0.000

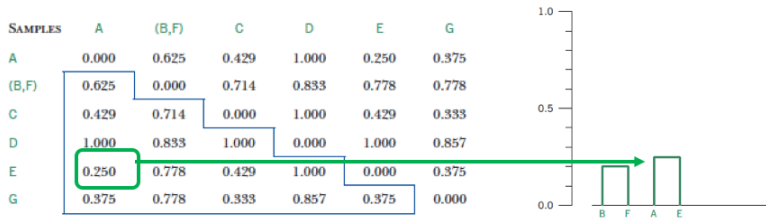
**Construcción
nueva matriz
usando
método de
distancia
máxima.**

SAMPLES	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

**(B-F) -A
(B-F) -C
(B-F) -D
(B-F) -E
(B-F) -G**

AGRUPAMIENTO JERARQUICO: PASO 3

Construcción del segundo nodo:
Mayor similitud entre A y E



AGRUPAMIENTO JERARQUICO: PASO 4

SAMPLES	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

SAMPLES	(A,E)	(B,F)	C	D	G
(A,E)	0.000	0.778	0.429	1.000	0.375
(B,F)	0.778	0.000	0.714	0.833	0.778
C	0.429	0.714	0.000	1.000	0.333
D	1.000	0.833	1.000	0.000	0.857
G	0.375	0.778	0.333	0.857	0.000

**Construcción
nueva matriz
usando
método de
distancia
máxima.**

(A-E) - (B-F)

(A-E) - C

(A-E) - D

(A-E) - G

AGRUPAMIENTO JERARQUICO: PASO 4.1

SAMPLES	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

SAMPLES	(A,E)	(B,F)	C	D	G
(A,E)	0.000	0.778	0.429	1.000	0.375
(B,F)	0.778	0.000	0.714	0.833	0.778
C	0.429	0.714	0.000	1.000	0.333
D	1.000	0.833	1.000	0.000	0.857
G	0.375	0.778	0.333	0.857	0.000

Construcción
nueva matriz
usando
método de
distancia
máxima.

(A-E) - (B-F)

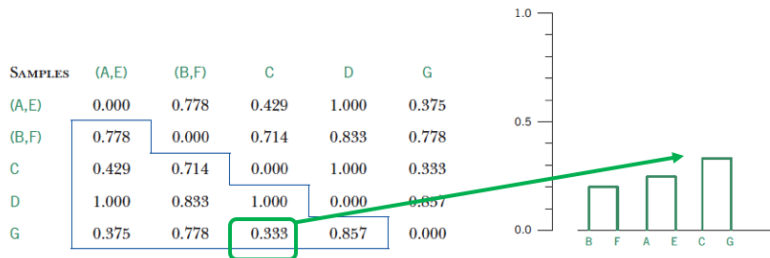
(A-E) - C

(A-E) - D

(A-E) - G

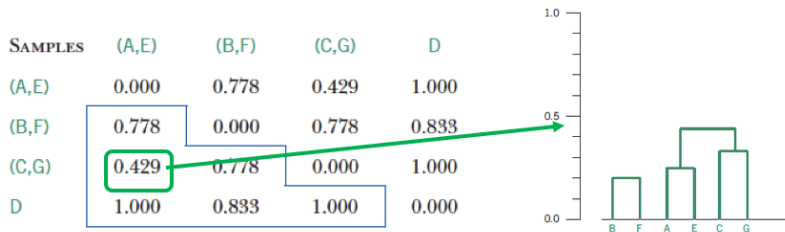
AGRUPAMIENTO JERARQUICO: PASO 5

Construcción del tercer nodo: Mayor similitud entre C y G



AGRUPAMIENTO JERARQUICO: PASO 6

Construcción del cuarto nodo: Mayor similitud entre A-E y C-G

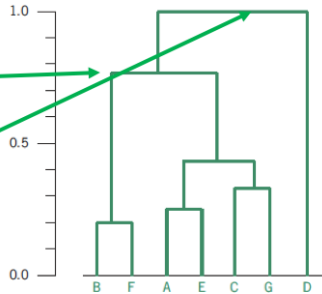


AGRUPAMIENTO JERARQUICO: PASO 7

Construcción del quinto y sexto nodo: Mayor similitud entre A-E-C-G con B-F y entre estos con D.

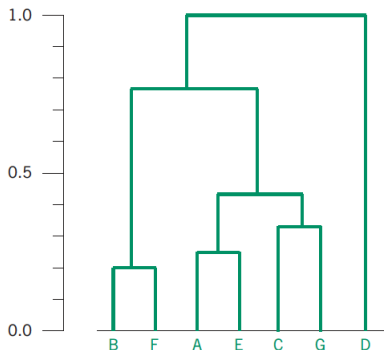
SAMPLES	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0.000	0.778	1.000
(B,F)	0.778	0.000	0.833
D	1.000	0.833	0.000

SAMPLES	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0.000	1.000
D	1.000	0.000



INTERPRETACIÓN CLUSTER JERÁRQUICO

- ▶ Establecemos nivel de agrupamiento = 0.5.
- ▶ Bajo 0.5 hay mas similaridad (Co-presencias).
- ▶ Se observan 3 grupos o cluster.



RESUMEN DE LA CLASE

- ▶ ¿Qué son los análisis multivariantes?.
- ▶ Estudio de caso 1: Fauna béntica como indicador de calidad ambiental.
- ▶ Matrices de distancia (Variables discretas): Jacard.
- ▶ Análisis de cluster: jerárquico y no jerárquico.