# CLASE 09 - ANÁLISIS DE CLUSTER

## OCE 313 - Técnicas de análisis no paramétricos.

Dr. José Gallardo Matus

Pontificia Universidad Católica de Valparaíso

22 May 2022

# PLAN DE LA CLASE

## 1.- Introducción

- ¿Qué son los análisis de cluster?
- Clasificación: jerárquico v/s no jerárquico
- Elaborar análisis de cluster jerárquico manual

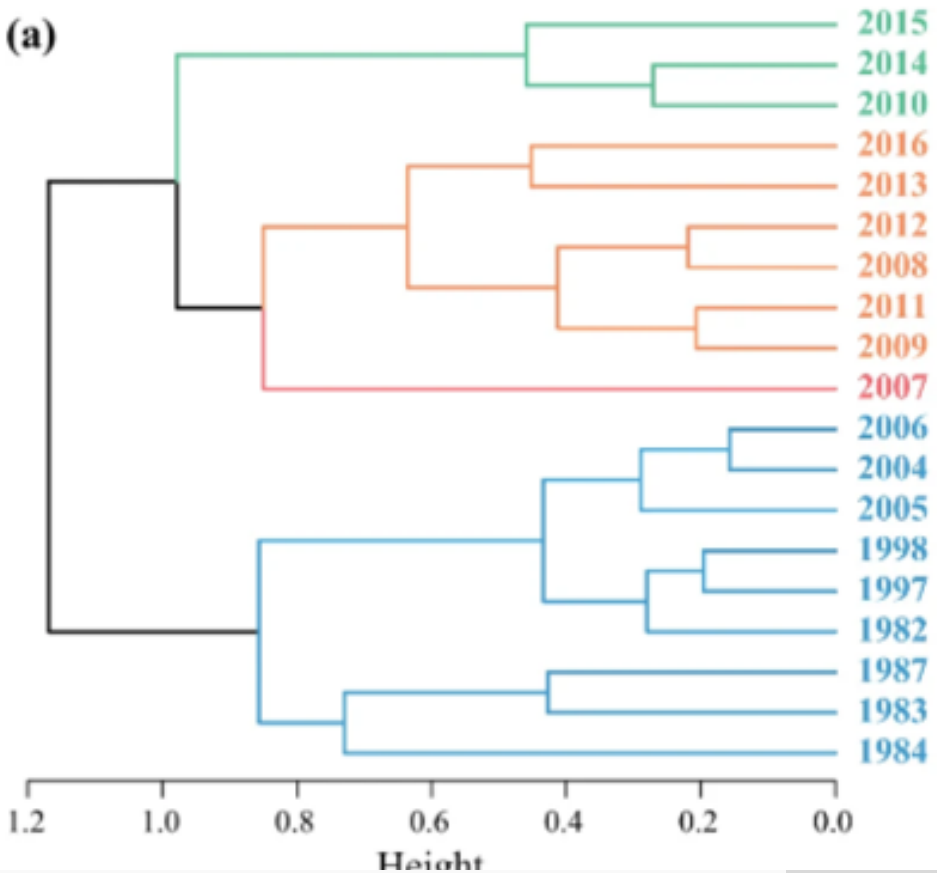
## 2). Práctica con R y Rstudio cloud.

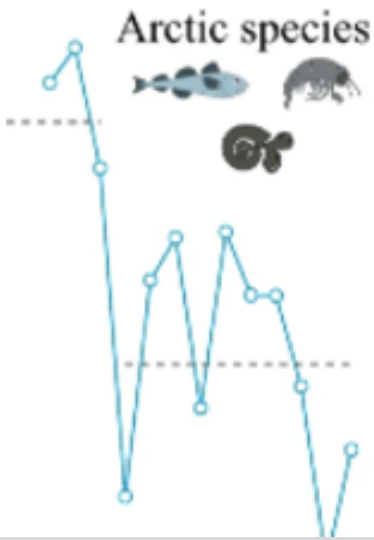
- Elaborar análisis de cluster jerárquico y no jerárquico con R.

# ANÁLISIS DE CLUSTER

**¿Qué son los análisis de cluster?**

Son herramientas de exploración de datos que permiten agrupar y visualizar datos multivariados con base a su similitud (matriz de distancia).

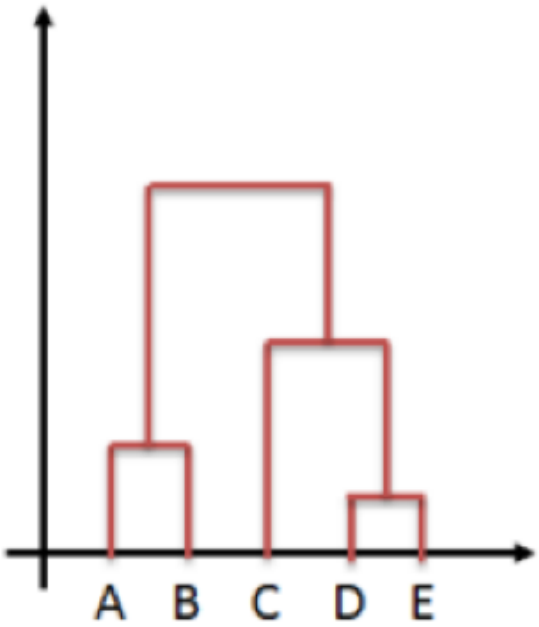


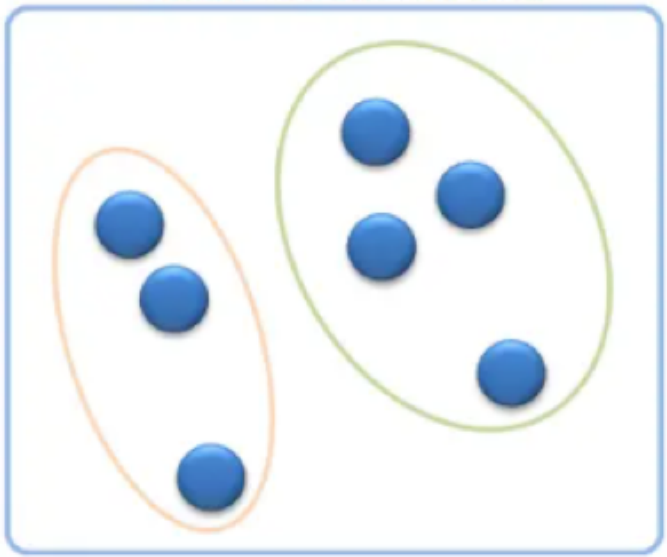



# ANALISIS DE CLUSTER: CLASIFICACION

**Jerárquico**: Los grupos se fusionan sucesivamente siguiendo una jerarquía de homogeneidad, la cual decrece a medida que se agregan más elementos al grupo.

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



# ANALISIS JERÁRQUICO: MÉTODO

**¿Qué hace el algoritmo estándar?**

1. Agrupa dos elementos por su similitud. 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?**

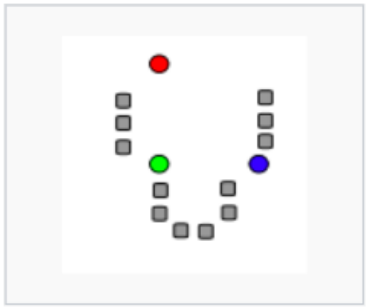
1. Método de distancia máxima (vecino más lejano). 2. Método de distancia mínima (vecino más próximo). 3. Método UPGMA (unweighted Pair-group arithmetic averages).

# ANALISIS NO JERÁRQUICO: MÉTODO

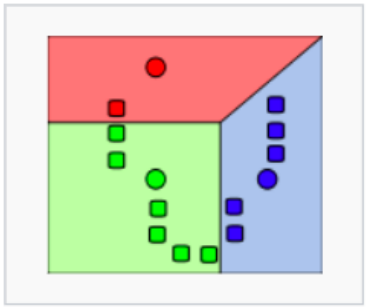
**Métodos K-MEANS**

**¿Qué hace el algoritmo k-means?**
1. Se crea un conjunto inicial de k centroides (lo define el investigador). 2. Asigna cada elemento al grupo con la media más cercana. 3. Calcula un nuevo centroide para cada grupo. 4. Finaliza cuando las asignaciones no cambian

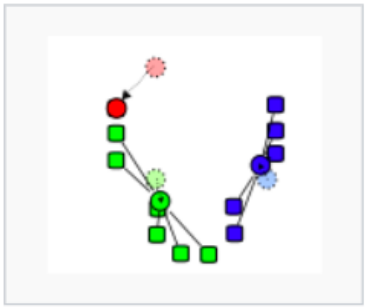
# MÉTODO K-MEANS



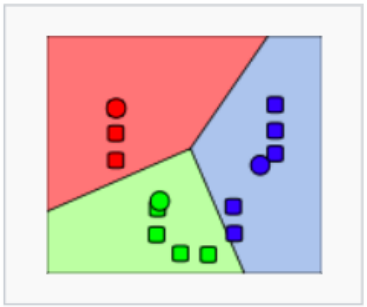
1) *k* centroides iniciales (en este caso *k*=3) son generados aleatoriamente dentro de un conjunto de datos (mostrados en color).



2) *k* grupos son generados asociándole el punto con la media más cercana. La partición aquí representa el diagrama de Voronoi generado por los centroides.



3) EL centroide de cada uno de los *k* grupos se recalcula.



4) Pasos 2 y 3 se repiten hasta que se logre la convergencia.

Fuente: https://es.wikipedia.org/wiki/K-medias

# VENTAJAS Y DESVENTAJAS

Jerárquico Ventajas: No requiere especificar Nº de grupos al inicio.
Desventajas: Difícil decidir que grupos son relevantes y cuales no.
Difícil de interpretar cuando existen muchos elementos.

No jerárquico: Ventajas: Útil cuando existen muchos elementos.
Desventajas: El número de cluster que se define al inicio, podría no ser el adecuado.

# EJEMPLO ESTUDIO DIVERSIDAD ESPECIES

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

| Samples | Species | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *sp1* | *sp2* | *sp3* | *sp4* | *sp5* | *sp6* | *sp7* | *sp8* | *sp9* | *sp10* |
| 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 es muy utilizado en Oceanografía para expresar el grado en el que dos muestras son semejantes por las especies presentes en ellas.

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

| | | Sample 2: 1 | Sample 2: 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 | sp1 | sp2 | sp3 | sp4 | sp5 | sp6 | sp7 | sp8 | sp9 | sp10 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| B | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

| Sitio | sp1 | sp2 | sp3 | sp4 | sp5 | sp6 | sp7 | sp8 | sp9 | sp10 |
|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 |
| F | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

| Sitio | sp1 | sp2 | sp3 | sp4 | sp5 | sp6 | sp7 | sp8 | sp9 | sp10 |
|---|---|---|---|---|---|---|---|---|---|---|
| B | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| F | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 1 |

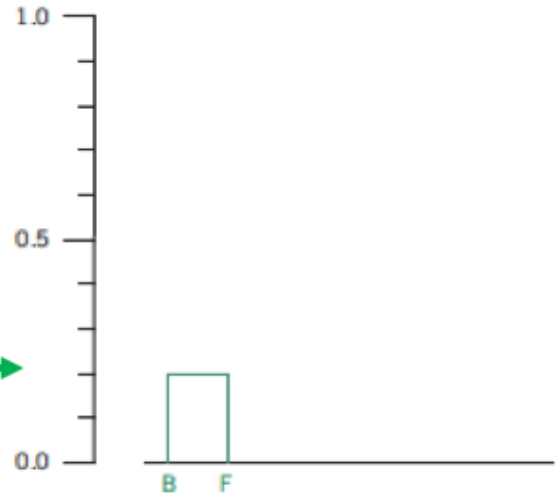
# MATRIZ DE SIMILARIDAD DE JACARD

| 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

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



# 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.**

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

| 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 |

# 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 |

| 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 |

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

**(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 |

| 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 |

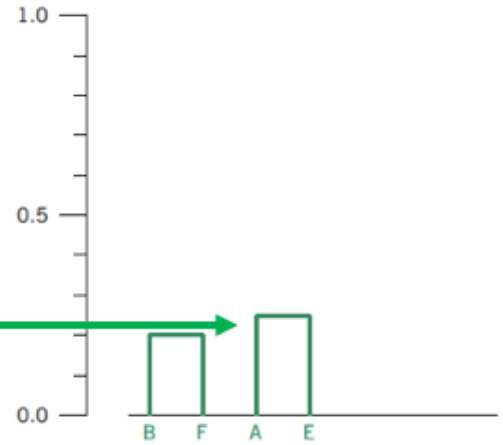
**Construcción nueva matriz usando método de distancia máxima.**

**(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**

| 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 |



# 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 |

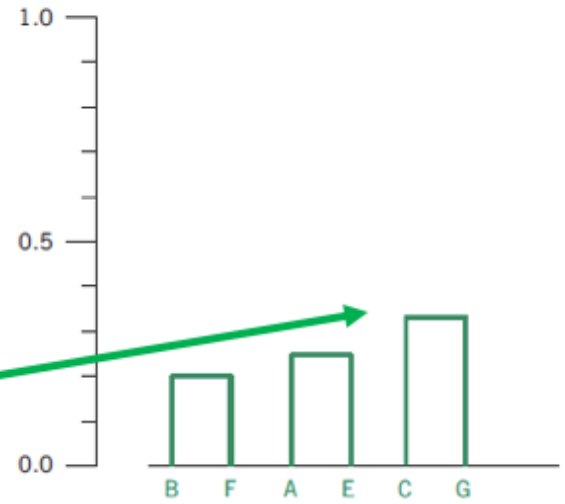
**Construcción nueva matriz usando método de distancia máxima.**

**(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

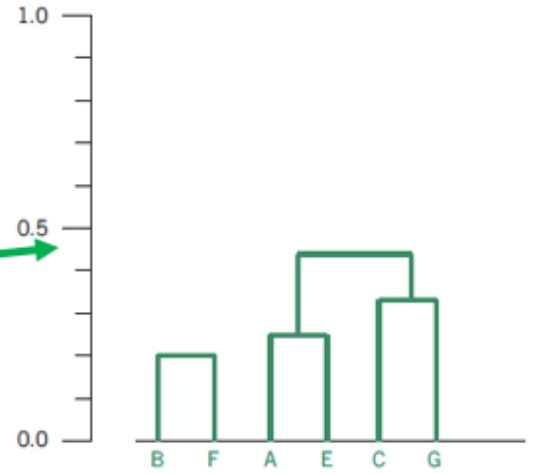
| Samples | (A,E) | (B,F) | C | D | G |
|---|---|---|---|---|---|
| (A,E) | 0.000 | 0.778 | 0.429 | 1.000 | 0.375 |
| (B,F) | 0.778 | 0.000 | 0.714 | 0.833 | 0.778 |
| C | 0.429 | 0.714 | 0.000 | 1.000 | 0.333 |
| D | 1.000 | 0.833 | 1.000 | 0.000 | 0.857 |
| G | 0.375 | 0.778 | 0.333 | 0.857 | 0.000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



# AGRUPAMIENTO JERARQUICO: PASO 6

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

| SAMPLES | (A,E) | (B,F) | (C,G) | D |
|---|---|---|---|---|
| (A,E) | 0.000 | 0.778 | 0.429 | 1.000 |
| (B,F) | 0.778 | 0.000 | 0.778 | 0.833 |
| (C,G) | 0.429 | 0.778 | 0.000 | 1.000 |
| D | 1.000 | 0.833 | 1.000 | 0.000 |

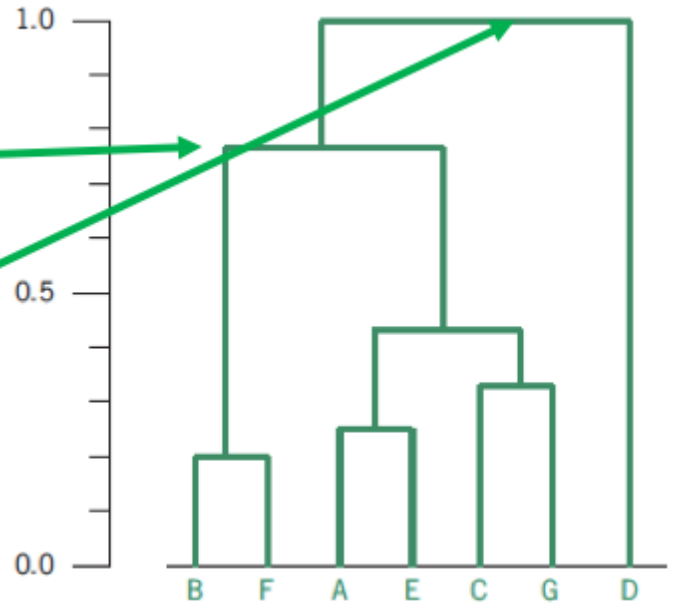


# AGRUPAMIENTO JERARQUICO: PASO 7

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

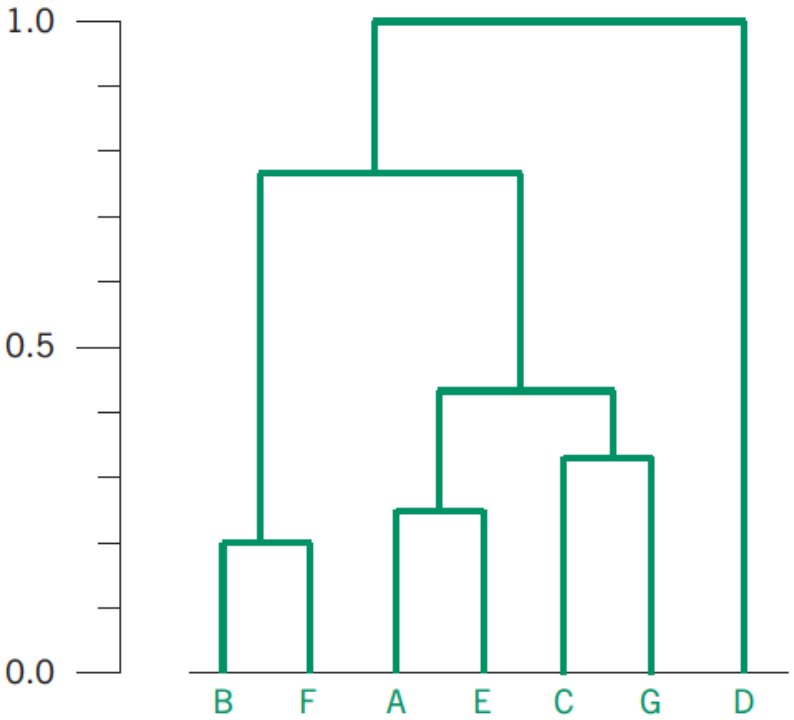
| Samples | (A,E,C,G) | (B,F) | D |
|---|---|---|---|
| (A,E,C,G) | 0.000 | 0.778 | 1.000 |
| (B,F) | 0.778 | 0.000 | 0.833 |
| D | 1.000 | 0.833 | 0.000 |

| 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 de cluster?.
- Analisis de cluster jerarquico (dendograma).
- Analisis de cluster no jerarquico.
- Indice de Jacard para datos de conteo (diversidad de especies).