Estadística IV

30/10/2024

Segundo Semestre 2024

Profesor: Kevin Carrasco

Daniela Olivares

Ayudante: María Fernanda Núñez

Universidad Alberto Hurtado

Departamento de Sociología

Análisis de Conglomerados
(Análisis de tipologías o clusters)

*****Bibliografía recomendada*****

- Cea D' Ancona, M. A. 2002. *Análisis multivariable: teoría y práctica.* Madrid: Ed. Síntesis (Cap 3, Análisis de Conglomerados)

- Hair J, F. et al (2004): *Análisis multivariante.* Madrid: Pearson (Cap. "Análisis de conglomerados")

I. Antecedentes y objetivos

- Sentido genérico: técnica enmarcada en el “Análisis de tipologías” (cluster analysis, Taxonomía Numérica, Análisis Q, etc.)

- Reúne un conjunto de procedimientos orientados a **agrupar *casos*** (individuos u objetos) en función de una serie de variables de clasificación

- **Objetivo:**
  - agrupar objetos o individuos en grupos o conglomerados (clusters) de modo tal de que los objetos en el mismo grupo se parezcan más entre sí que otros pertenecientes a otro grupo

I. Antecedentes y objetivos

- *Idea clave:* maximizar la homogeneidad de los objetos dentro de un grupo y maximizar la heterogeneidad entre los grupos.

- Con ello se busca identificar un de patrón de agrupamiento subyacente a los casos.

- Esta técnica es especialmente valiosa porque no necesita que el/la investigador/a defina previamente los grupos, sino que permite que los datos mismos revelen agrupamientos naturales en la población.

II. Utilidad y limitaciones

- Utilidad*:*

  - *Taxonomía* (uso exploratorio)*:* identificación exploratoria, proceso de clasificar y agrupar casos en categorías significativas, estructurar la diversidad dentro de los datos

  - *Reducción de datos*: generación de perfiles de sujetos (u objetos). Ej. perfiles de consumidores; perfiles de países, perfiles de votantes, perfiles de trabajadores independientes, etc. ≠ análisis factorial

  - *Identificación de relaciones:* al armar grupos, es posible analizar relaciones no observables a primera vista

  - *Contrastación de hipótesis sobre conglomerados* (uso confirmatorio): Una vez generadas los grupos, es posible contrastar hipótesis relativas al carácter agrupado de los datos

II. Utilidad y limitaciones

- Ejemplos de preguntas de investigación:

  - ¿Es posible diferenciar un conjunto de instituciones (universidades, hospitales, etc.) según el tipo de relación que tienen con el Estado (ej. su financiamiento estatal, la venta de servicios a instituciones estatales, etc.)?

  - ¿Es posible dividir un conjunto determinado de países según el tipo de relaciones laborales imperante en cada uno de ellos (ej. niveles de sindicalización, niveles de conflictividad laboral, tipo de leyes que regulan las relaciones colectivas entre empresarios y trabajadores, etc.)?

  - ¿Es posible clasificar a los/as chilenos/as según sus opiniones a temas valóricos como la legalización del aborto, la legalización del consumo de drogas, la extensión de derechos a grupos con distintas identidades de género, etc.?

II. Utilidad y limitaciones

- Ejemplo PNUD, 2015. *Los tiempos de la politización*. Cap. 17, Los diversos modos de estar involucrado (pp. 150-155):
  - Generación de una tipología de ***modos de involucramiento con lo político*** a partir de variables como éstas (más otras no mostradas acá):

91. ¿Con qué frecuencia conversa con su familia, compañeros de trabajo o amigos sobre temas políticos? (porcentaje)

	Con mucha frecuencia	Con bastante frecuencia	Con poca frecuencia	Con ninguna frecuencia	NS-NR
a. Con su familia	6,0	18,8	48,8	26,2	0,2
b. Con sus compañeros de trabajo o estudio	3,4	14,5	37,8	36,2	8,1
c. Con sus amigos	4,5	15,6	40,0	38,7	1,2
d. Con personas que no conoce	1,7	5,3	26,9	64,6	1,5

100. ¿Y cuánto diría usted que la política influye en su vida? (porcentaje)

Mucho	Bastante	Poco	Nada	NS-NR
4,7	17,1	38,8	36,6	2,8

101. ¿Votó en las últimas elecciones municipales? (porcentaje)

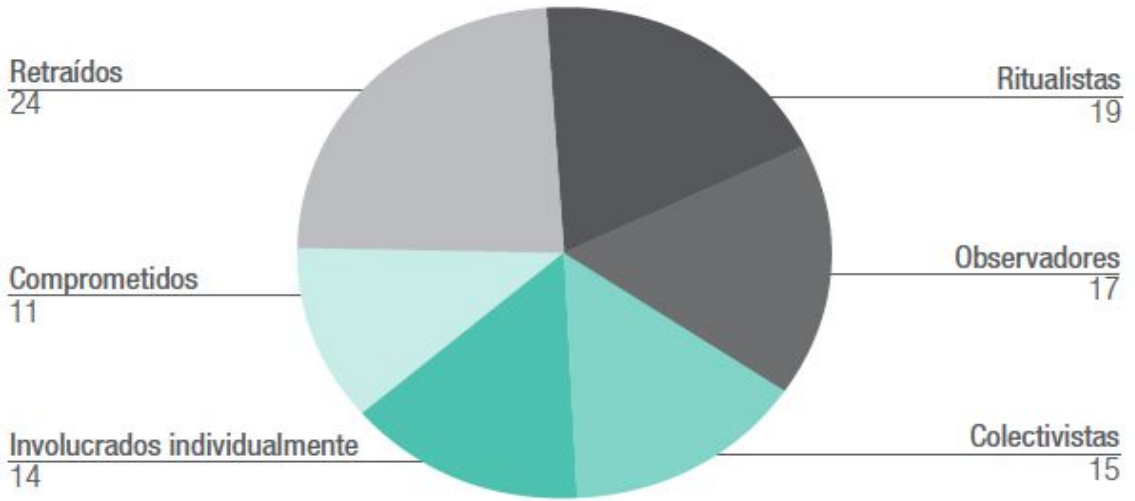
Sí	No	NS-NR
61,5	37,2	1,3

II. Utilidad y limitaciones

- Tipología resultante:

GRÁFICO 17

Modos de involucramiento con lo político* (porcentaje)



* n=1.805

Fuente: Elaboración propia sobre la base de la Encuesta de Desarrollo Humano, PNUD 2013.

II. Utilidad y limitaciones

- *Limitaciones:*
  - Muchas veces funciona la técnica funciona “muy bien” (Hair et al, 2014: p. 419),
  - Es una técnica descriptiva, ateórica y no permite hacer inferencias relativas a la población; resultados no necesariamente generalizables debido a dependencia “excesiva” de la unidad de medida de las variables

III. Funcionamiento

- El análisis de conglomerados requiere 6 tareas fundamentales*:*

1. Resolución de aspectos previos: diseño de investigación, revisión de supuestos y elección de las variables y su forma de medición
2. **Medición de la similitud** entre los objetos o sujetos (casos): hay diferentes formas que llevan a resultados diferentes
3. **Formación de los grupos: dos grandes procedimientos**
   - Procedimientos jerárquicos
   - Procedimientos no-jerárquicos
4. Toma de decisión: ¿con cuántos grupos quedarse?
5. Caracterización de los grupos
6. Validación de los grupos

III. Funcionamiento

- El análisis de conglomerados requiere 6 tareas fundamentales*:*

1. **Resolución de aspectos previos: diseño de investigación, revisión de supuestos y elección de las variables y su forma de medición**

1. Diseño de investigación, supuestos y variables

- Importancia del diseño de investigación; capacidad de responder las siguientes preguntas

a) ¿Qué variables incluir?

- Importancia de la teoría
- Variables pueden ser ordinales o métricas
- Mientras más variables, mejor (si hay un n grande)
- Unidad de medida
- <u>NO debe haber multicolinealidad entre las variables</u>
  - Regresiones: multicolinealidad impide discernir el “impacto real” de variables X multicolineales sobre Y
  - A. Factorial: multicolinealidad no es un problema
  - A. Conglomerados: variables multicolineales tendrán más peso porque su *medida de similitud será sistemáticamente mayor que variables no multicolineales*
    - Ej. concreto: no es recomendable incluir nivel de estudios y años de estudio en el mismo análisis

1. Diseño de investigación, supuestos y variables

b) ¿Es adecuada la muestra?

- Idealmente debe ser de un tamaño representativo; aunque aquí *no* se trata de estimar parámetros
- Un n grande ayuda a que todos los grupos de la población estén igualmente representados (problema de outliers)
- Un n muy grande dificulta la comprensión de los métodos jerárquicos

c) ¿Existen outliers que afecten la formación de grupos?

- Primero, identificar outliers (especialmente importante para n chicas)
- Si existen, buscar motivos de peso para eliminarlos (en caso que se quieran eliminar)

1. Diseño de investigación, supuestos y variables

d) ¿Se deben estandarizar las variables?

- La forma más común es transformación a puntaje Z
- La estandarización es buena porque facilita la comparación entre variables con distintas unidades de medida y elimina el potencial efecto de la unidad de medida
- Sin embargo, es posible que diferentes escalas de medición representen fenómenos que en realidad tienen diferentes ”escalas” (ej. probablemente 1$ de salario ≠ 1 punto en “escala de felicidad” que va de 1 a 10)

- Para procedimientos no jerárquicos (K – medias), la estandarización es un procedimiento muy común

III. Funcionamiento

- El análisis de conglomerados requiere 6 tareas fundamentales*:*

1. …
2. **Medición de la similitud** entre los objetos o sujetos (casos): hay diferentes procedimientos que llevan a resultados diferentes

2. Medidas de similitud entre los objetos

- Dependiendo del tipo de investigación y variables, existen tres tipos de medidas de similitud

1. Medidas de asociación como Chi-cuadrado: rara vez se utilizan porque los softwares estadísticos enfatizan el tratamiento “numérico” de variables categóricas para desarrollar Análisis de Conglomerados (Vila-Baños et al, 2014)

2. Correlación: se usa, aunque no mucho porque *no* mide distancia sino que más bien la pauta de variación conjunta entre variables

3. **Medidas de distancia**: son las más usadas para medir similitud. Existen varias medidas de distancia.

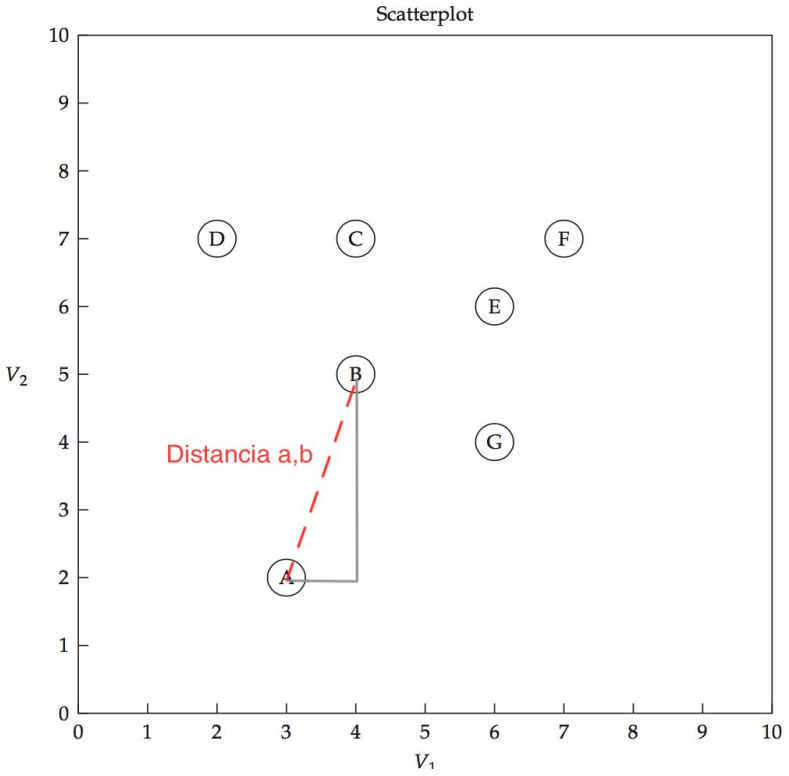
2. Medidas de similitud entre los objetos

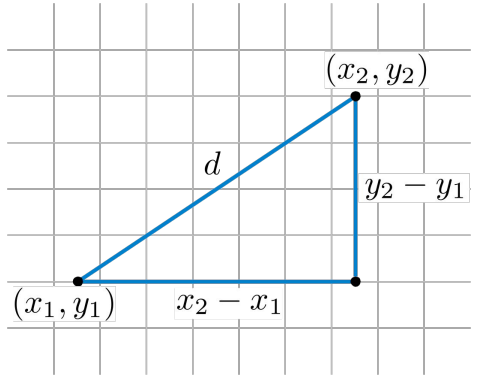
- **Medidas de distancia (similitud)**
  1. **Distancia euclídea**: una de las más usadas en la investigación empírica, define similitud a través de una línea recta (cercanía en un plano)

2. Medidas de similitud entre los objetos

- **Medidas de distancia (similitud)**

1. **Distancia euclídea**: una de las más usadas en la investigación empírica, define similitud a través de una línea recta (cercanía en un plano)





$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$Distancia\ (a, b) = \sqrt{\sum (a_i - b_i)^2}$$

2. Medidas de similitud entre los objetos

- **Medidas de distancia (similitud)**
  1. **Distancia euclídea…**

  2. **Distancia euclídea al cuadrado**: se desarrolló para hacer las cosas más rápidas; estadísticamente es lo mismo que la anterior

  3. Distancia de Manhattan: se calcula como la suma de todas las diferencias absolutas entre las puntuaciones de los casos (la suma de los lados del triángulo, no sólo de la hipotenusa)

  4. Distancia de Mahalanobis: mide distancia estandarizando las variables y considerando la existencia de correlación entre ellas (ajusta la varianza de cada grupo por la correlación entre sus miembros).

III. Funcionamiento

- El análisis de conglomerados requiere 6 tareas fundamentales*:*

1. …
2. **…**
3. **Formación de los grupos: dos grandes procedimientos**
   - Procedimientos jerárquicos
   - Procedimientos no-jerárquicos

3. Formación de grupos: métodos de aglomeración

- Decisión clave para aglomerar grupos:
  - métodos de aglomeración jerárquicos: método (casi) enteramente exploratorio
  - métodos no jerárquicos: número de grupos predefinidos

3. Métodos de aglomeración jerárquicos

- Clasifican y agrupan los casos por etapas mediante un proceso en forma de "árbol" (dendrograma)
  - Originalmente, hechos para variables métricas, también se usan con variables ordinales
  - Ventajas: permiten observar gráficos de aglomeración (dendrogramas) y matrices de distancia especialmente útiles con n pequeños

  - Existen 2 tipos de procedimientos jerárquicos:

1. **Aglomerativos (más usado)**
2. Divisivos

3. Métodos de aglomeración jerárquicos: ejemplo

- **Datos ficticios:** ¿Cómo opera cada paso? Considerar este ejemplo (ficticio) planteado por Hair, Joseph F Jr., et. al. 2014. *Multivariate data analysis*. London: Pearson. Cap. 8

- Problema de investigación: caracterizar un grupo de 7 consumidores según dos variables:
  1. V1: fidelidad a tienda donde compran
  2. V2: fidelidad a la marca consumida

- Los sujetos respondieron una encuesta donde cada variable se midió en una escala de 0 a 10

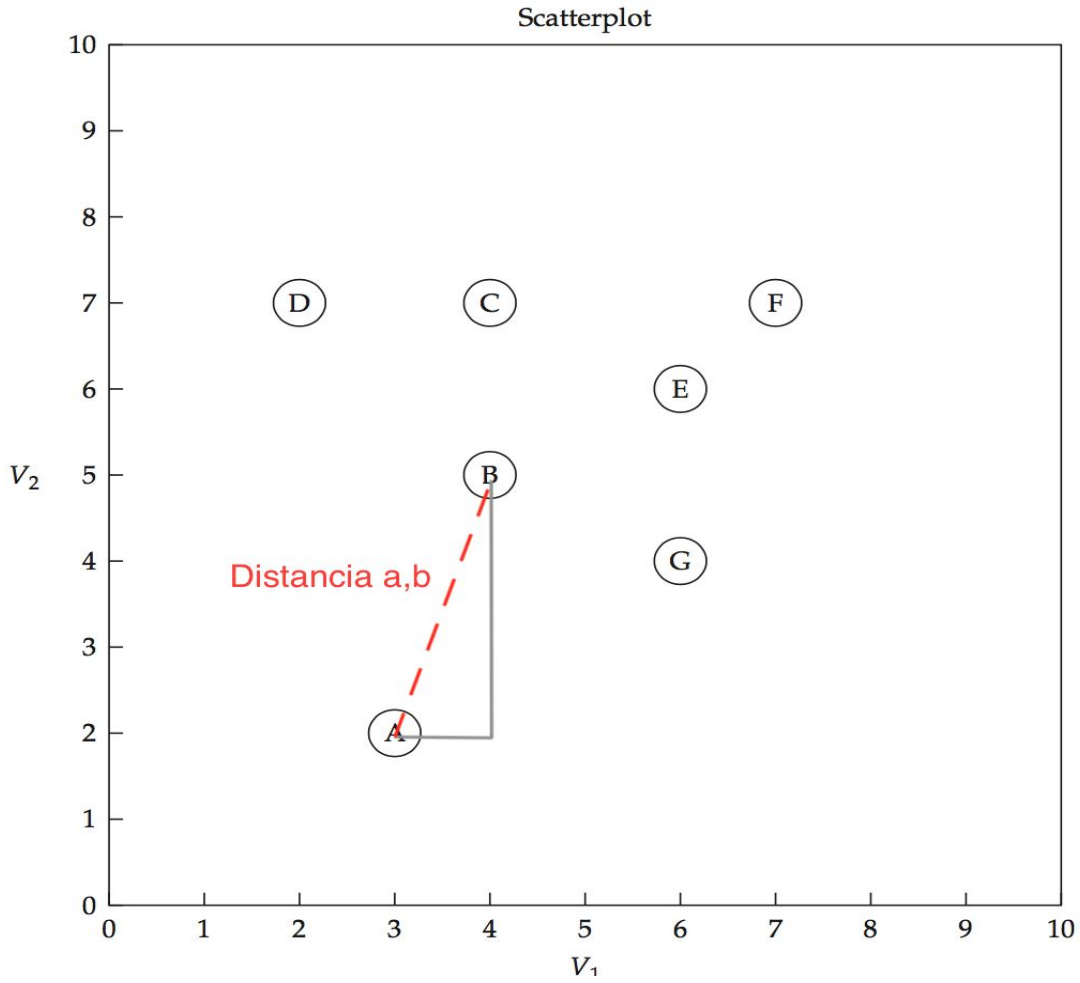
Ejemplo

- Resultados de la encuesta:

	Sujetos						
Variables	**A**	**B**	**C**	**D**	**E**	**F**	**G**
V_1 Fidelidad a la tienda	3	4	4	2	6	7	6
V_2 Fidelidad a la marca	2	5	7	7	6	7	4

Ejemplo

- **Cálculo distancia euclídea**



Ejemplo

- Si calculamos todas las distancias euclídeas entre los pares de casos, se tiene la siguiente matriz de proximidad

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	2,000	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

Ejemplo

- ¿Qué nos dice esta matriz?

- Valores más bajos = casos más cercanos = casos más similares entre sí = miembros potenciales del mismo grupo

- ¿Cómo formar grupos a partir de esto?

- Idea comúnmente utilizada: *identificar las dos observaciones más cercanas (similares) que no se encuentren en el mismo grupo y unirlas.*
  - Esto implicaría comenzar un **proceso formación de grupos:**
    - **jerárquico** (de “abajo hacia arriba”) y
    - **aglomerativo,** a través del procedimiento del “vecino más cercano”.

- Esto se presenta a continuación

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	2,000	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
0	Ninguna	7 grupos (grupo = caso individual)

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	2,000	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	**1,414**	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
1	E – F	6 grupos: (E-F) + 5 casos individuales restantes (A), (B), (C), (D), (G)

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	2,000	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	**2,000**	3,162	—

- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
2	Hay varias opciones: B-C, C-D y **E-G**. Sin embargo, E ya es parte del primer grupo. Por eso, G se suma a ese grupo	5 grupos: (E-F-G) + 4 casos individuales restantes (A), (B), (C), (D)

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	**2,000**	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
3	Se combinan los pares restantes cercanos: B-C y C-D. Se parte con C-D	4 grupos: (E-F-G), (C-D) + 2 casos individuales restantes (A), (B)

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	**2,000**	—				
D	5,099	2,828	2,000	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
4	Se sigue con B-C. ¿Qué pasa? C ya es parte de C-D. Así B se une al grupo	3 grupos: (E-F-G), (B-C-D) + 1 caso individual restante (A)

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	2,000	—			
E	5,000	**2,236**	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
5	Se debería integrar A. **Problema:** hay una distancia menor aún sin agrupar (distancia B-E = 2,236). Por eso, los grupos donde están B y E se unen	2 grupos: (B-C-D-E-F-G) + 1 caso individual restante (A)

Formación de grupos

Matriz de proximidad (distancias euclídeas)							
	A	B	C	D	E	F	G
A	—						
B	3,162	—					
C	5,099	2,000	—				
D	5,099	2,828	2,000	—			
E	5,000	2,236	2,236	4,123	—		
F	6,403	3,606	3,000	5,000	1,414	—	
G	3,606	2,236	3,606	5,000	2,000	3,162	—

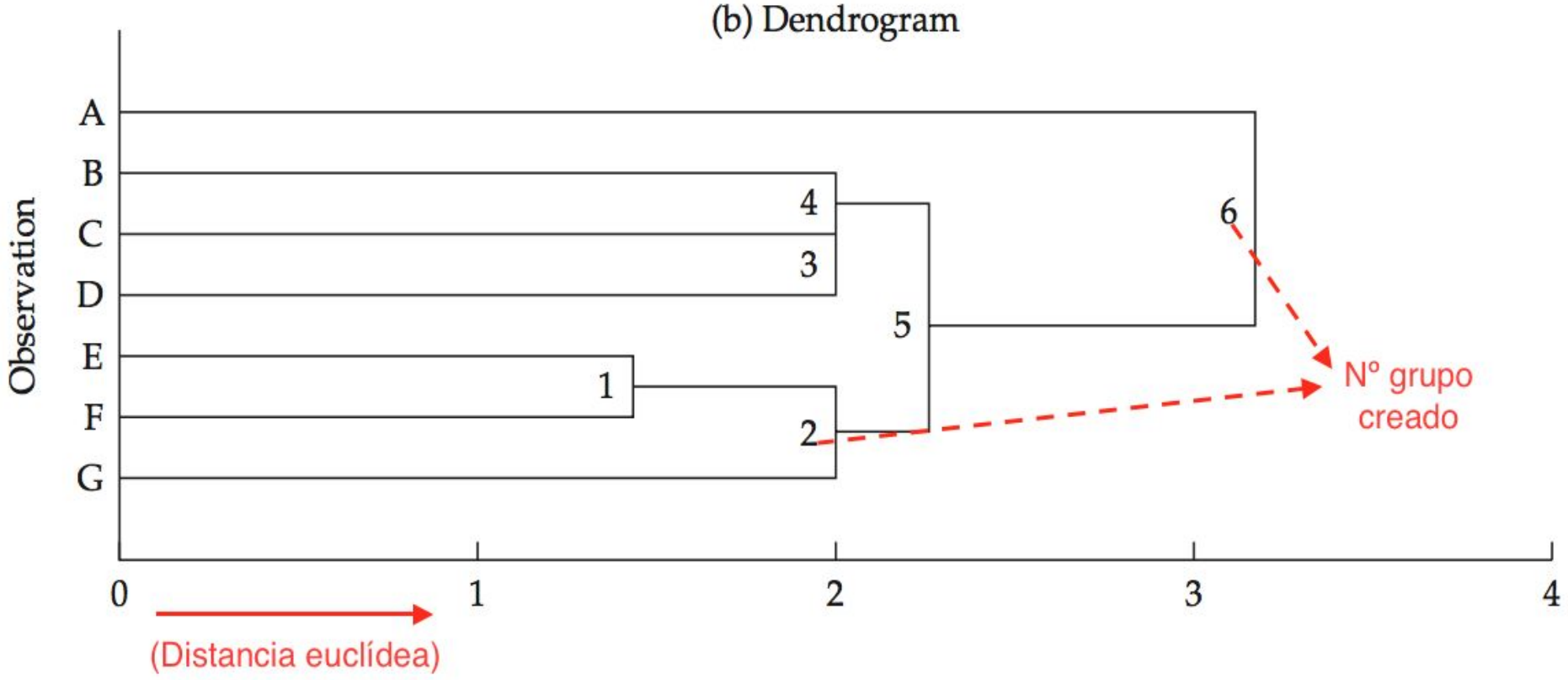
- Proceso de aglomeración

Paso	Observaciones combinadas	Resultado
6	Sólo falta integrar A. La menor distancia es con B (3,162). Como B ya es parte de un grupo, A se integra a ese grupo	1 gran grupo que incluye a todos los casos (A-B-C-D-E-F-G)

Formación de grupos

Tabla resumen proceso de agrupamiento jerárquico – aglomerativo					
	Proceso de aglomeración			Solución	
Paso	Distancia mínima (distancia euclídea) entre observaciones	Par de observaciones		Miembros del grupo o conglomerado	Número de grupos o conglomerados
0 (solución inicial)				(A) (B) (C) (D) (E) (F) (G)	7
1	1, 414	E-F		(A) (B) (C) (D) **(E-F)** (G)	6
2	2,000	E-G		(A) (B) (C) (D) **(E-F-G)**	5
3	2,000	C-D		(A) (B) **(C-D) (E-F-G)**	4
4	2,000	B-C		(A) **(B-C-D) (E-F-G)**	3
5	2,236	B-E		(A) **(B-C-D-E-F-G)**	2
6	3,162	A-B		**(A-B-C-D-E-F-G)**	1

Formación de grupos



En resumen, en este ejemplo se agrupó a los individuos

- **De forma jerárquica (de "abajo hacia arriba")**
- Los casos se unieron a través de un **proceso aglomerativo** y del procedimiento (algoritmo) del **"vecino más cercano"**

3. Métodos de aglomeración jerárquicos: algoritmos

- **Dentro de los métodos de aglomeración jerárquicos,** existen diversos ***algoritmos de agrupamiento***
  - Procedimientos que definen cuándo los grupos se van a combinar entre sí
  - Ellos, por tanto, entregan información sobre cuántos grupos escoger como "solución óptima"

- Existen al menos 6 algoritmos de agrupamiento; cada uno con su método particular de funcionamiento

- Estos algoritmos ***no*** dependen necesariamente del tipo de distancia utilizada

3. Métodos de aglomeración jerárquicos: algoritmos

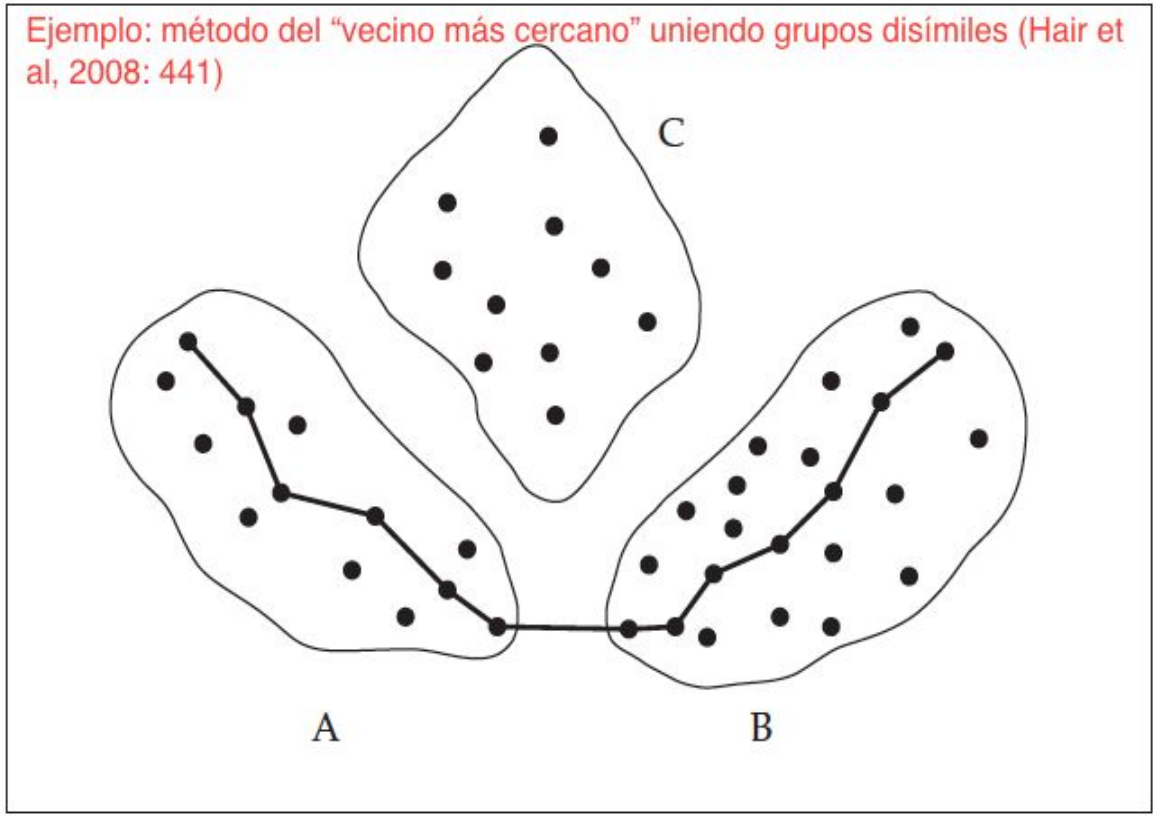
1. Método del “Vecino más cercano” (single-linkage)

- Similitud entre grupos = distancia más corta entre el objeto de uno y otro grupo (ver ejemplo)

- Problema: riesgo de encadenamiento

3. Métodos de aglomeración jerárquicos: algoritmos

- Problema: riesgo de encadenamiento

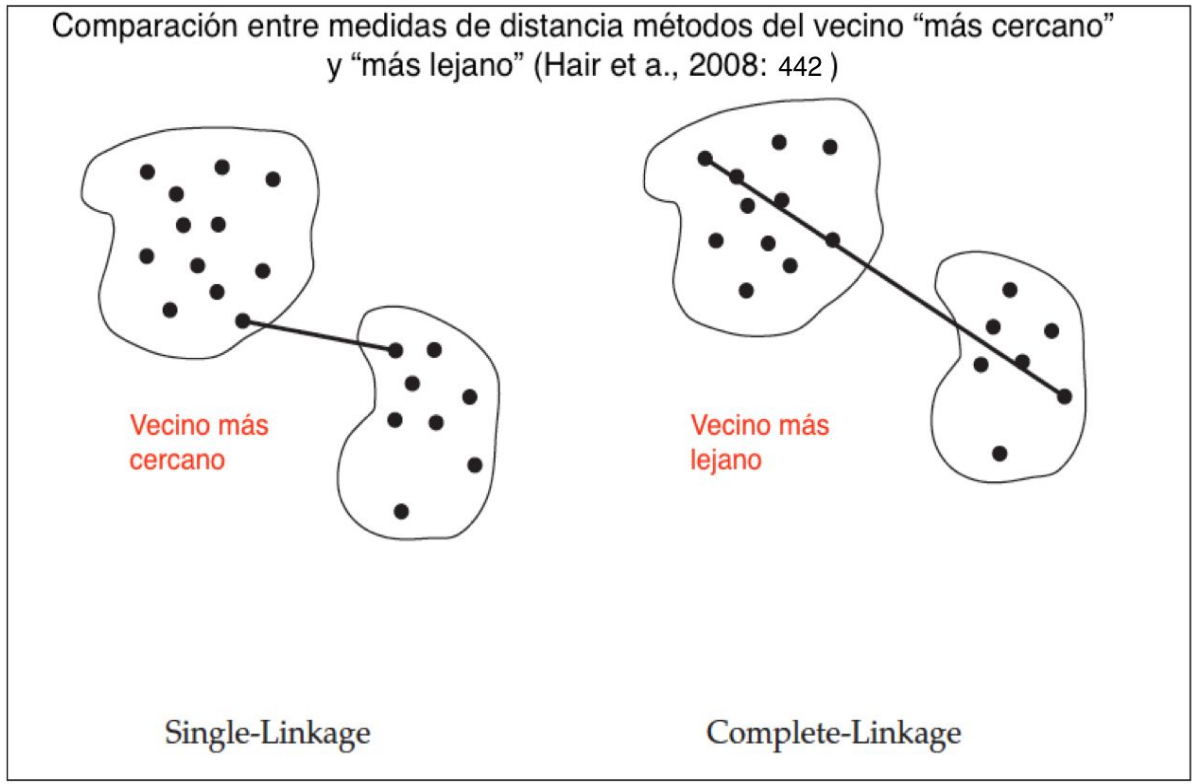


3. Métodos de aglomeración jerárquicos: algoritmos

2. Método del “Vecino más lejano” (complete-linkage)

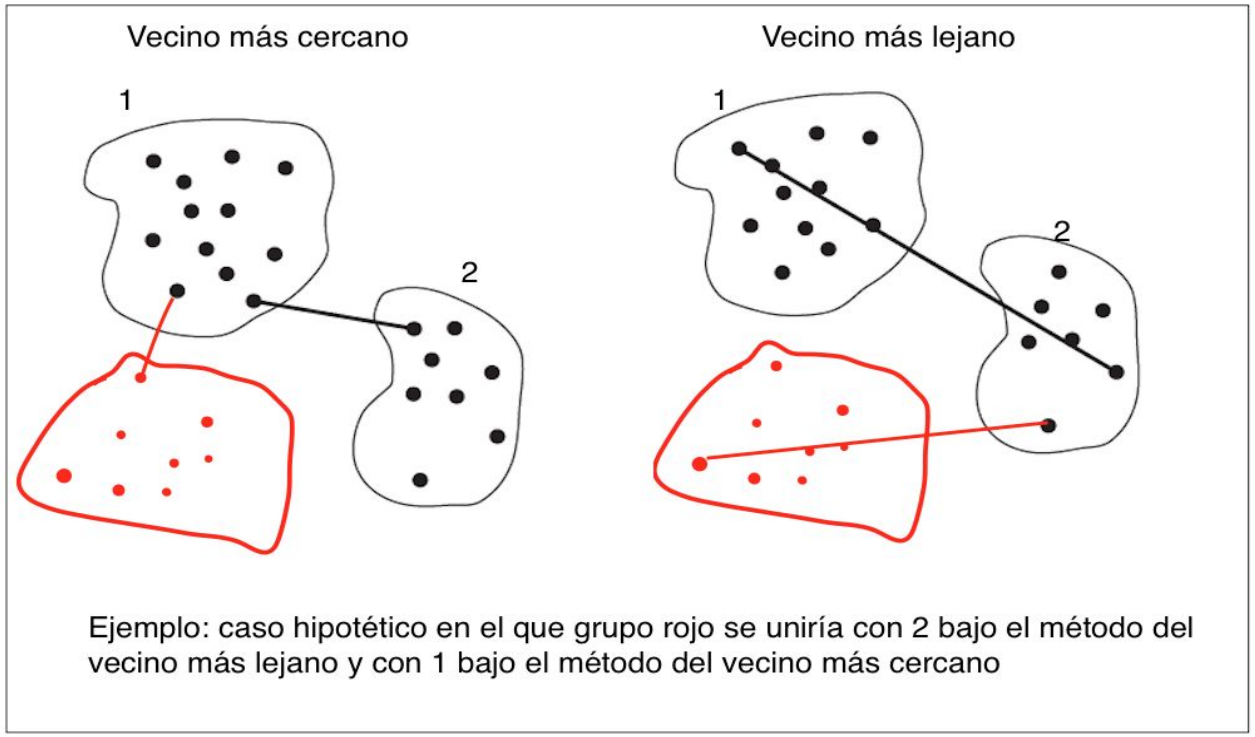
- Similitud entre grupos:
  - Análisis las distancias máximas entre las observaciones de cada grupo
  - Se unen grupos cuyas distancias “más lejanas” son las menores

3. Métodos de aglomeración jerárquicos: algoritmos



3. Métodos de aglomeración jerárquicos: algoritmos

Diferencias entre métodos del ”Vecino más cercano” y “Vecino más lejano”

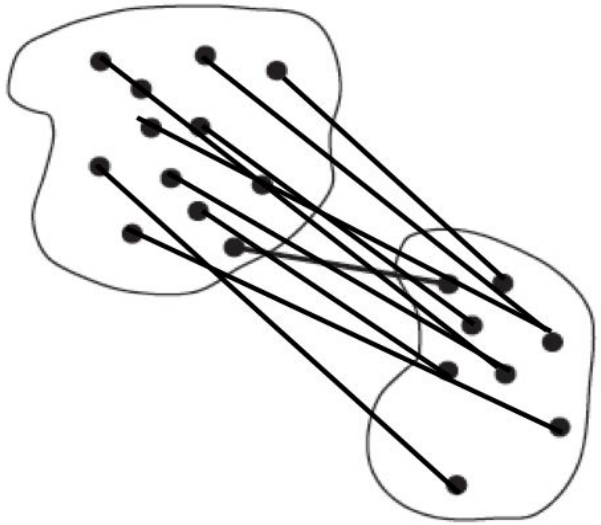


Ejemplo: caso hipotético en el que grupo rojo se uniría con 2 bajo el método del vecino más lejano y con 1 bajo el método del vecino más cercano

3. Métodos de aglomeración jerárquicos: algoritmos

3. Promedio de las distancias entre grupos (average linkage between groups)

- Se calcula la distancia promedio de *todos* los integrantes de un grupo respecto de los integrantes de otro grupo; dos grupos se combinan cuando el promedio distancia es la menor posible (algoritmo definido por defecto en algunas versiones de SPSS)



Promedio de las distancias entre grupos

3. Métodos de aglomeración jerárquicos: algoritmos

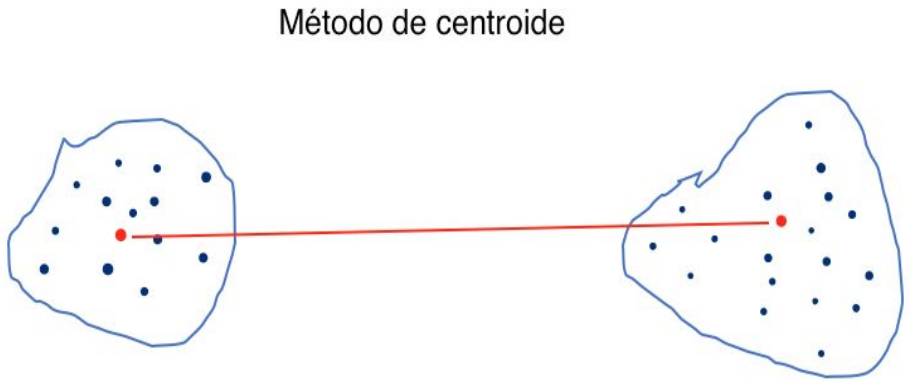
4. Promedio intragrupal

- Variante del algoritmo anterior
- Se parte combinando de dos en dos los grupos, de manera sucesiva.
- Luego se calcula la distancia promedio entre los miembros de esos grupos.
- Se mantienen sólo los grupos cuya distancia promedio intragrupal es la menor

3. Métodos de aglomeración jerárquicos: algoritmos

5. Método del centroide

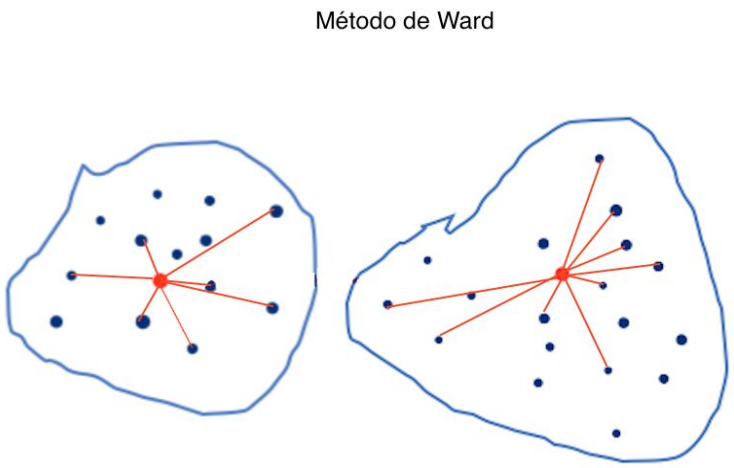
- Similitud = distancia entre centroides (valores promedios) de los grupos
- Cálculo: cada vez que los casos son reagrupados, un nuevo centroide se reclacula



3. Métodos de aglomeración jerárquicos: algoritmos

6. Método de Ward

- Se calculan las distancias al cuadrado de cada caso a la media de su grupo al cuadrado
- Luego se calcula la *suma total* de dichas distancias
- Dos grupos se unen cuando esa suma total sea la menor posible, es decir, cuando se cumpla el criterio de minimización de “varianza” (suma de distancias al cuadrado), asumiendo que *siempre* el valor de las distancias al cuadrado va a aumentar



3. Métodos de aglomeración jerárquicos: Ejemplo

Ejemplo

- Problema de investigación: ¿Es posible agrupar países según sus modalidades de relaciones laborales?
- Casos: 45 países
- Base de datos: Institutional Characteristics of Trade Unions, Wage Setting, State Intervention and Social Pacts (ICTWWS) de Jelle Visser

Variables y medición

1. Nivel de centralización de las relaciones laborales: valor teórico 0 - 100
2. Derecho a sindicalización (índice): valor teórico 0 - 100
3. Derecho a negociación colectiva (índice): valor teórico 0 - 100
4. Derecho a huelga (índice): valor teórico 0 - 100
5. % Cobertura negociación colectiva: valor teórico 0 - 100

Variable relevante excluida: intervención del estado en relaciones laborales (multicolinealidad)

3. Métodos de aglomeración jerárquicos: ejemplo

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
1. Centralization	45	,00	100,00	30,0000	25,89314
2. Right of association	45	33,33	100,00	88,5185	15,00094
3. Right of collective bargaining	45	33,33	100,00	74,4444	19,00292
4. Right to strike	45	33,33	100,00	64,8148	19,85497
5. Bargaining coverage rate	45	6,00	98,00	44,5647	29,06477
Valid N (listwise)	45				

3. Métodos de aglomeración jerárquicos: ejemplo

- Análisis de conglomerado jerárquico
  - Medida de distancia: distancia euclídea al cuadrado
  - Algoritmo de agrupamiento*:* Ward
  - Especificación rango de soluciones: 2 a 4

3. Métodos de aglomeración jerárquicos: ejemplo

Matriz de proximidad (extracto)

Proximity Matrix

Case	Squared Euclidean Distance									
	1:Argentina	2:Australia	3:Austria	4:Belgium	5:Brazil	6:Bulgaria	7:Canada	8:Chile	11:Cyprus	12:Czech Republic
1:Argentina	,000	3691,744	3855,214	8286,682	1669,395	4306,568	5347,462	5656,947	4289,090	4164,789
2:Australia	3691,744	,000	3403,598	9696,686	2038,857	1518,358	2975,993	3212,773	3275,835	1985,078
3:Austria	3855,214	3403,598	,000	3632,472	2076,877	6167,250	9499,031	10561,164	6695,439	6737,685
4:Belgium	8286,682	9696,686	3632,472	,000	4292,730	12336,222	15320,826	18561,709	8484,716	10584,003
5:Brazil	1669,395	2038,857	2076,877	4292,730	,000	2756,196	3797,053	5253,556	1574,127	1998,419
6:Bulgaria	4306,568	1518,358	6167,250	12336,222	2756,196	,000	1458,333	1020,652	2762,026	2157,616
7:Canada	5347,462	2975,993	9499,031	15320,826	3797,053	1458,333	,000	673,674	1442,218	768,316
8:Chile	5656,947	3212,773	10561,164	18561,709	5253,556	1020,652	673,674	,000	3578,611	2394,964
11:Cyprus	4289,090	3275,835	6695,439	8484,716	1574,127	2762,026	1442,218	3578,611	,000	438,470
12:Czech Republic	4164,789	1985,078	6737,685	10584,003	1998,419	2157,616	768,316	2394,964	438,470	,000
13:Denmark	6692,795	5591,967	3807,111	1828,028	2182,934	8042,361	8596,680	12132,280	3468,237	4680,606
14:Estonia	8852,278	4773,929	11753,472	15329,000	5099,507	2883,222	869,468	2523,590	1395,051	1302,042
15:Finland	8395,064	8604,444	4261,111	443,028	3855,469	11613,361	13417,477	17148,706	6747,867	8547,300
16:France	4688,547	4653,598	3125,000	4188,028	1799,099	7417,250	6721,254	10005,609	2945,439	3473,796
17:Germany	3245,144	2523,261	2187,716	3436,366	489,171	2779,766	5001,345	6296,436	2393,660	3161,753

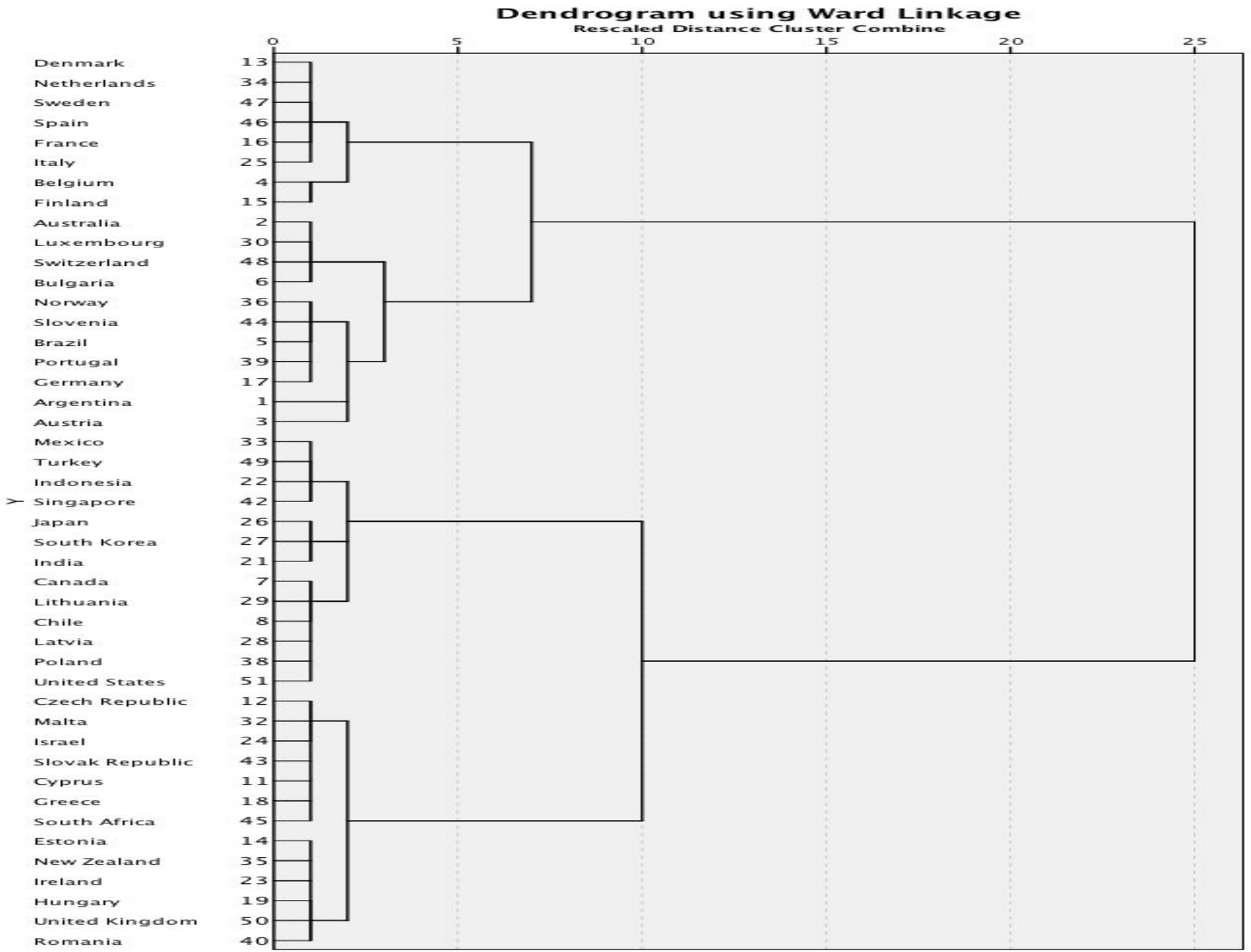
Coef. algoritmo de Ward
(suma al cuadrado de distancias)

Historial de aglomeración (extracto)

Agglomeration Schedule

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	13	34	,350	0	0	3
2	8	28	5,289	0	0	23
3	13	47	19,283	1	0	7
4	14	35	48,928	0	0	5
5	14	23	165,970	4	0	8
6	11	18	309,938	0	0	22
7	13	46	479,634	3	0	30
8	14	19	688,211	5	0	26
9	24	43	905,950	0	0	27
10	5	39	1124,913	0	0	24
11	36	44	1343,927	0	0	32
12	4	15	1565,441	0	0	37
13	26	27	1796,864	0	0	20
14	33	49	2030,158	0	0	19
15	16	25	2270,283	0	0	30

Dendrograma



4. Definición número de grupos: ejemplo

Caracterización/Validación de grupos: ¿tiene sentido formar 2, 3 o 4 grupos? ¿A qué grupo pertenece cada país?

Cluster Membership

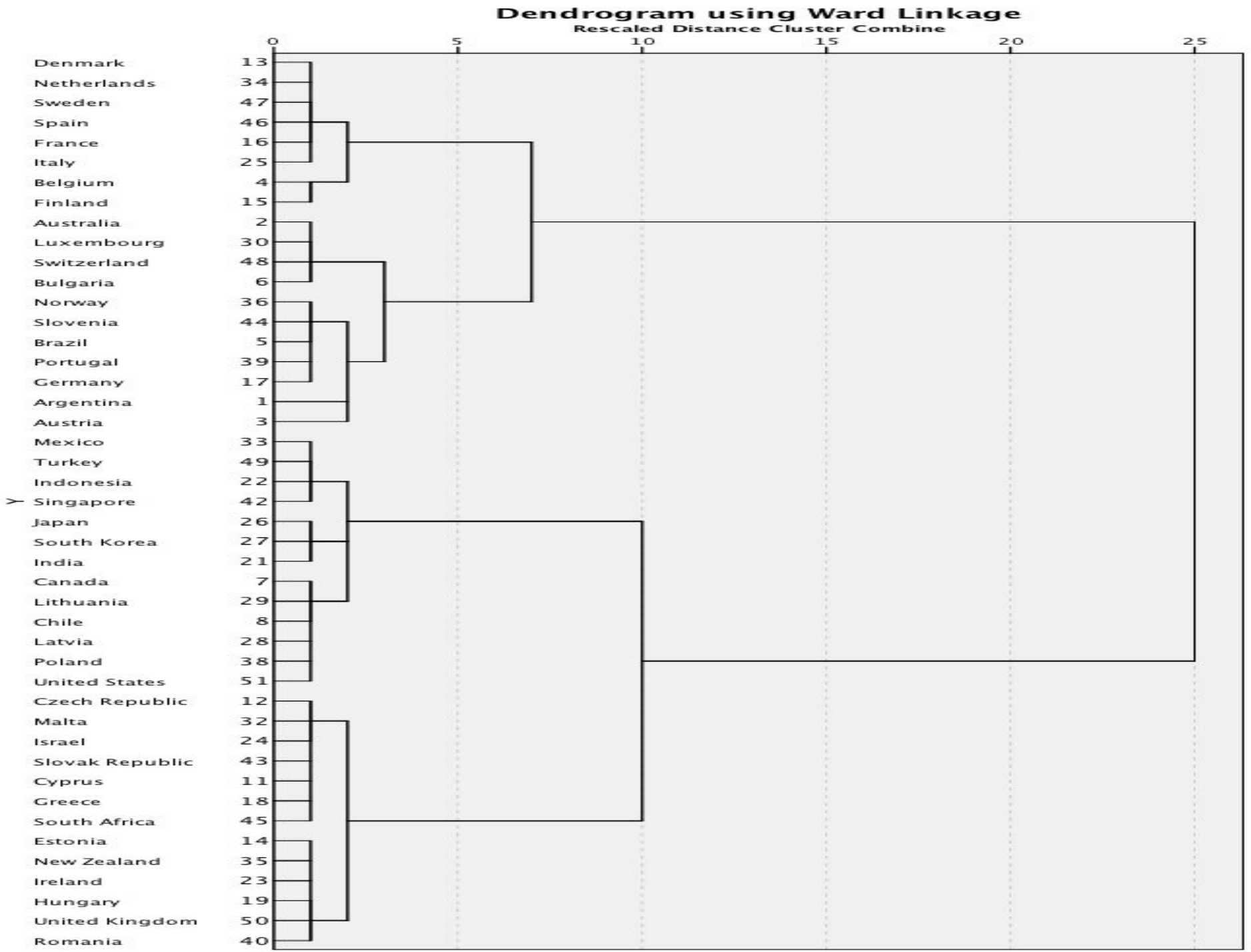
Case	4 Clusters	3 Clusters	2 Clusters
1:Argentina	1	1	1
2:Australia	1	1	1
3:Austria	1	1	1
4:Belgium	2	1	1
5:Brazil	1	1	1
6:Bulgaria	1	1	1
7:Canada	3	2	2
8:Chile	3	2	2
11:Cyprus	4	3	2
12:Czech Republic	4	3	2
13:Denmark	2	1	1
14:Estonia	4	3	2
15:Finland	2	1	1
16:France	2	1	1
17:Germany	1	1	1
18:Greece	4	3	2
19:Hungary	4	3	2
21:India	3	2	2

Grupo 1: Argentina, Australia, Austria, Bélgica, Alemania, Suecia, etc. relaciones "corporativistas"

Grupo 2: Canadá, Chile, Corea de Sur, Singapur, USA, etc. relaciones liberales

Grupo 3: Chipre, República Checa, Estonia, Grecia, Hungría, etc. intermedios

Dendrograma



4. Selección del número de conglomerados

- ¿Qué se puede concluir?

- En general, la heterogeneidad tiende a aumentar en la medida en que se tenga menos grupos (más casos dentro de un grupo = mayores distancias entre pares de observaciones = grupos más heterogéneos)

- Solución ideal: paso en donde la heterogeneidad no incrementó “mucho” □ necesidad de respaldo teórico o empírico

- En este caso, la solución apropiada sería la con 3, ya que las medidas de heterogeneidad no aumentan sustancialmente.

2 pasos restantes

5. Caracterización de los conglomerados
   - ¿Cómo definir los grupos formados?
   - ¿Qué características tienen sus miembros?

6. Validación de los conglomerados
   - Idea clave: si los grupos formados son “reales”, entonces ellos deberían generar diferencias significativas en otras variables no utilizadas para generarlo (ej. ingresos, orientaciones políticas, etc.)

5-6. Caracterización/validación de grupos: ejemplo

Caracterización/Validación de grupos:

Ejemplo: solución de 3 grupos.

- Cruce con variable "taza de sindicalización"

		Union density rate		
		Count	Mean	Standard Deviation
Ward Method	1	19	32,33	19,81
	2	13	13,64	6,09
	3	13	24,14	13,49
	Total	45	24,57	16,78

ANOVA

Union density rate, net union membership as a proportion of wage earners in employment (Num*100/WSEE)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	2698,845	2	1349,423	5,846	,006
Within Groups	9695,372	42	230,842		
Total	12394,217	44			

Grupo 1: relaciones "corporativistas"

Grupo 2: relaciones liberales

Grupo 3: países intermedios