

ANÁLISIS DE CLUSTERS

Sesiones de estadística DAP-Cat. 2022-2023

22 de Noviembre 2022



Albert Roso-Llorach

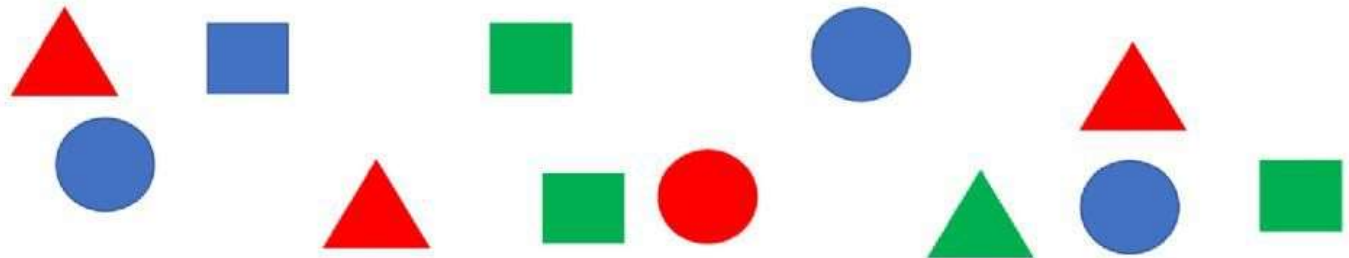
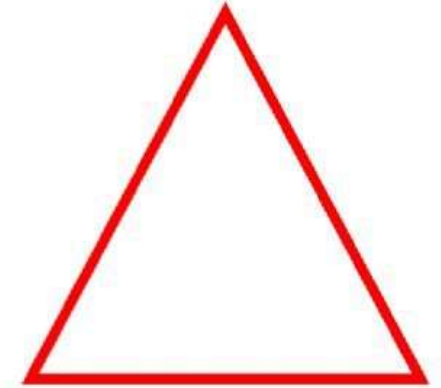
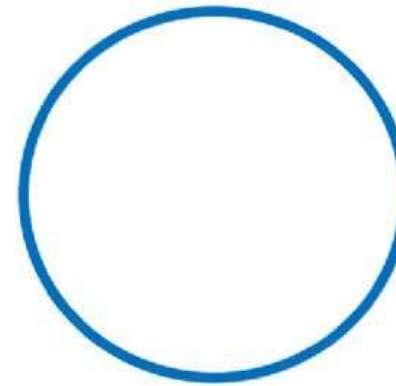
*Fundació Institut Universitari per a la recerca a
l'Atenció Primària de Salut Jordi Gol i Gurina
(IDIAPJGol), Barcelona
aroso@idiapjgol.org*



Introducción

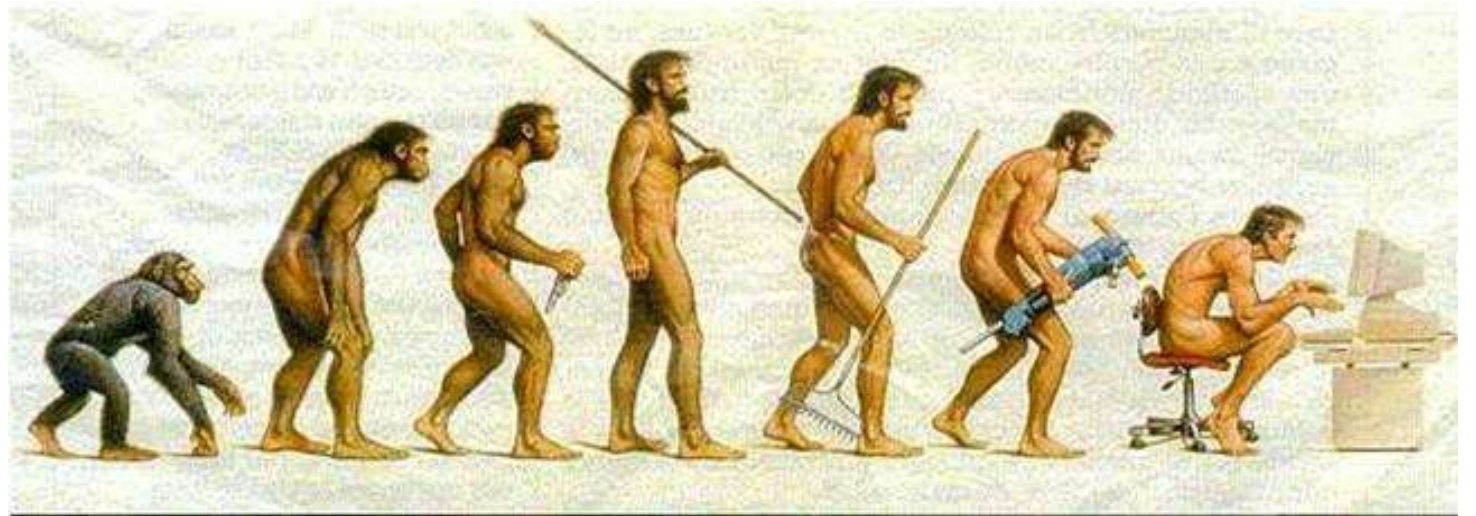
- Una de las habilidades más básicas de los seres vivos consiste en agrupar objetos para producir una clasificación

Clasifica las figuras de acuerdo a la forma y el color



Introducción

- La idea de clasificar cosas similares en categorías ya viene de lejos!!!!
- Los humanos primitivos necesitaban saber que cosas son:
 - Comestibles
 - Venenosas
 - Peligrosas
 - etc.



Clasificación

- ¿Qué podemos clasificar?
- Personas
- Animales
- Elementos químicos
- Estrellas
- Etc.

TABLA PERIÓDICA DE LOS ELEMENTOS

Number of elements: 118

Symbol: B

Atomic number: 5

Atomic mass: 10.811

Group: 13 (IIIA)

Period: 2

Classification: Metalloid

Legend:

- metals alcalinos
- metales alcalinotérreos
- metales
- metales de transición
- semimetales
- metaloideos
- no metales
- halógenos
- gases nobles
- actínidos

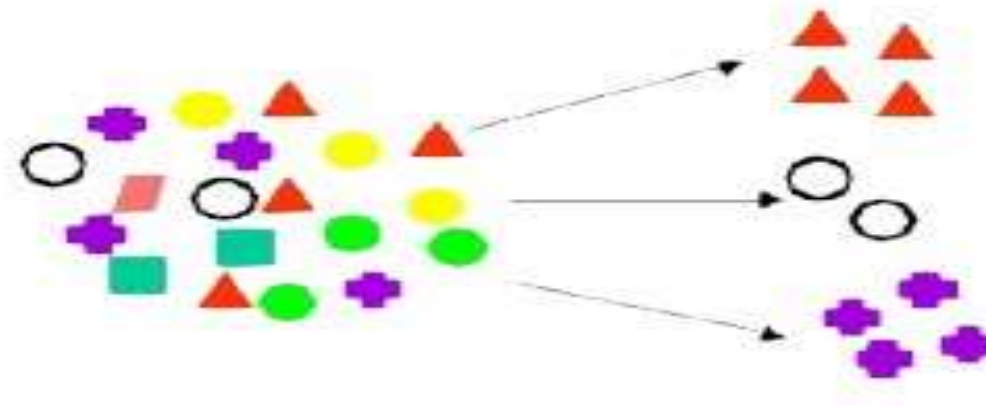
¿Porqué clasificar?

- Para simplificar grandes cantidades de información

Datos



Grupos de
objetos



- Descripción concisa de patrones de similitudes y diferencias en los datos

Métodos de clasificación

- Técnicas numéricas -> Biología y Zoología
- Más objetivas y menos subjetivas!!
- Clasificaciones:
 - Reproducibles
 - Estables

Métodos de clasificación



Iris Versicolor

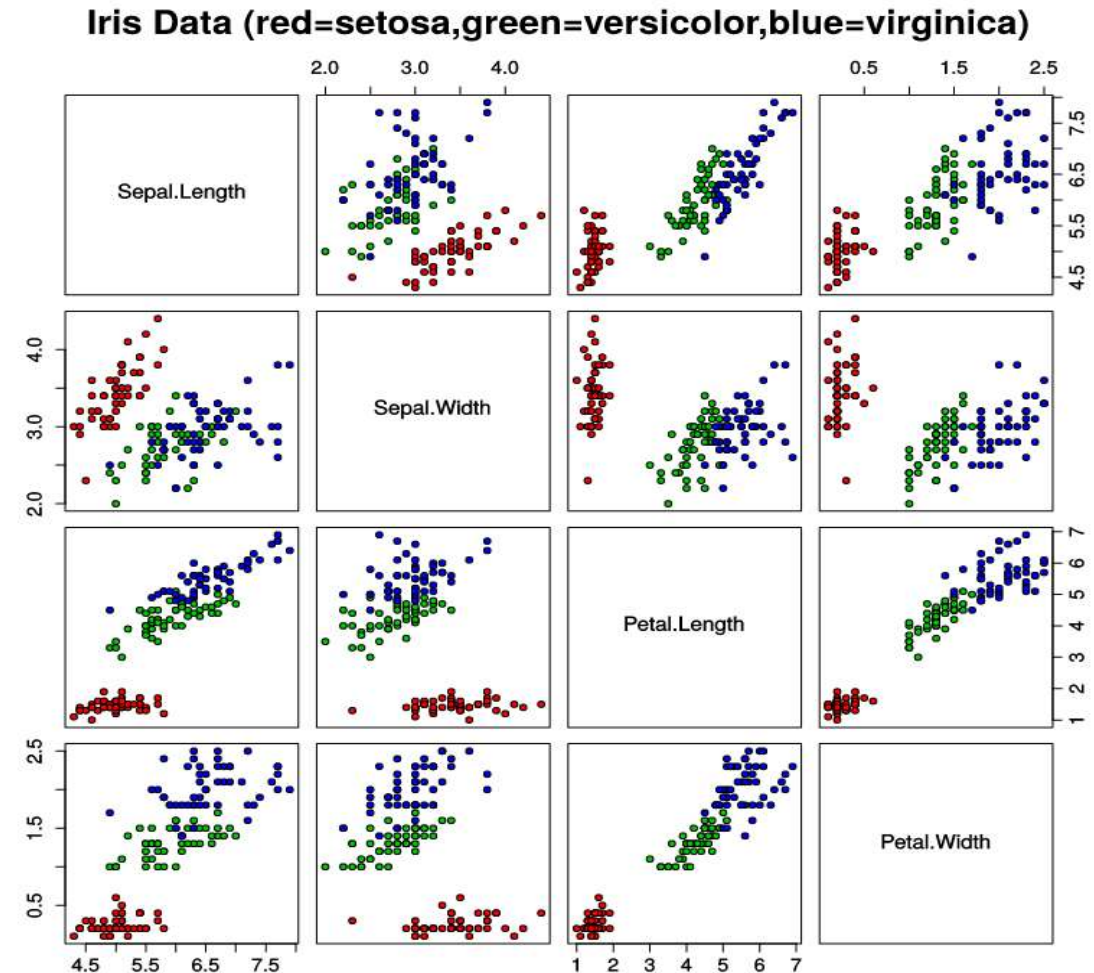


Iris Setosa



Iris Virginica

[Ronald Fisher](#) (1936)



¿Qué es un clúster?

- Cluster, grupo, clase, conglomerados??
- Definición Bonner (1964):
 - “clúster”: tiene una respuesta de valor para el investigador
- Definición Cormack (1971) y Gordon (1999):
 - cohesión interna (homogeneidad) y externa (separación)



Análisis de clusters

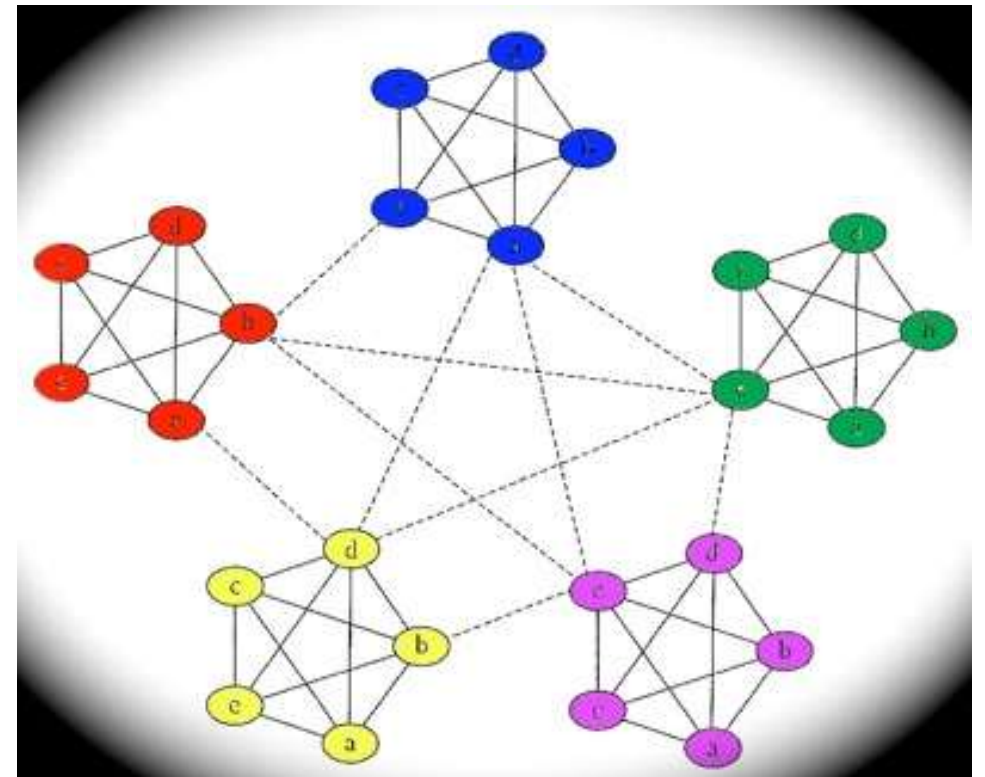
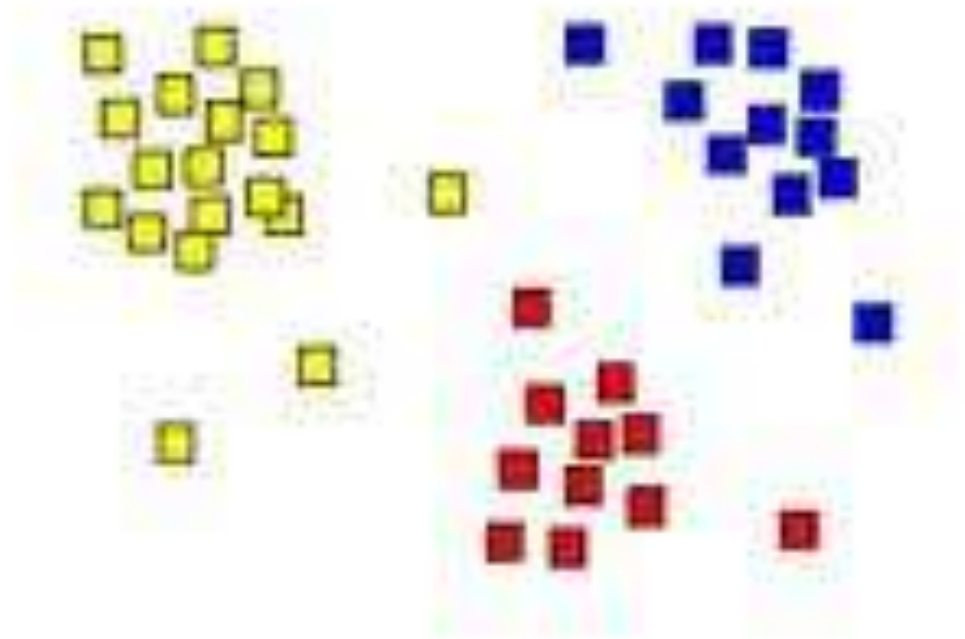
Objetivo: Dado un conjunto de objetos (variables o casos) agruparlos de manera que:

1. Objetos pertenecientes a un mismo grupo sean lo más parecidos entre sí, manteniendo cohesión interna.
2. Objetos pertenecientes a grupos diferentes tengan un comportamiento diferente con respecto a las variables analizadas, es decir, que cada grupo esté aislado externamente de los otros grupos.

Análisis de clusters

Técnica exploratoria:

No se utiliza ningún tipo de modelo estadístico para llevar a cabo el proceso de clasificación -> técnica de aprendizaje no supervisado





Campos de aplicación

- Estadística
- Reconocimiento de patrones
- Análisis de imágenes
- Búsqueda y recuperación de información
- Bioinformática
- Compresión de datos
- Computación Gráfica
- Aprendizaje automático (Machine Learning)

Etapas análisis de clústers

Selección de objetos (variables / individuos) a analizar

Selección de la medida de asociación (distancia / similitud / proximidad)

Selección y aplicación del criterio de agrupación

Determinación de la estructura correcta (elección del número de grupos)

Etapas análisis de clústers

Selección de objetos (variables / individuos) a analizar

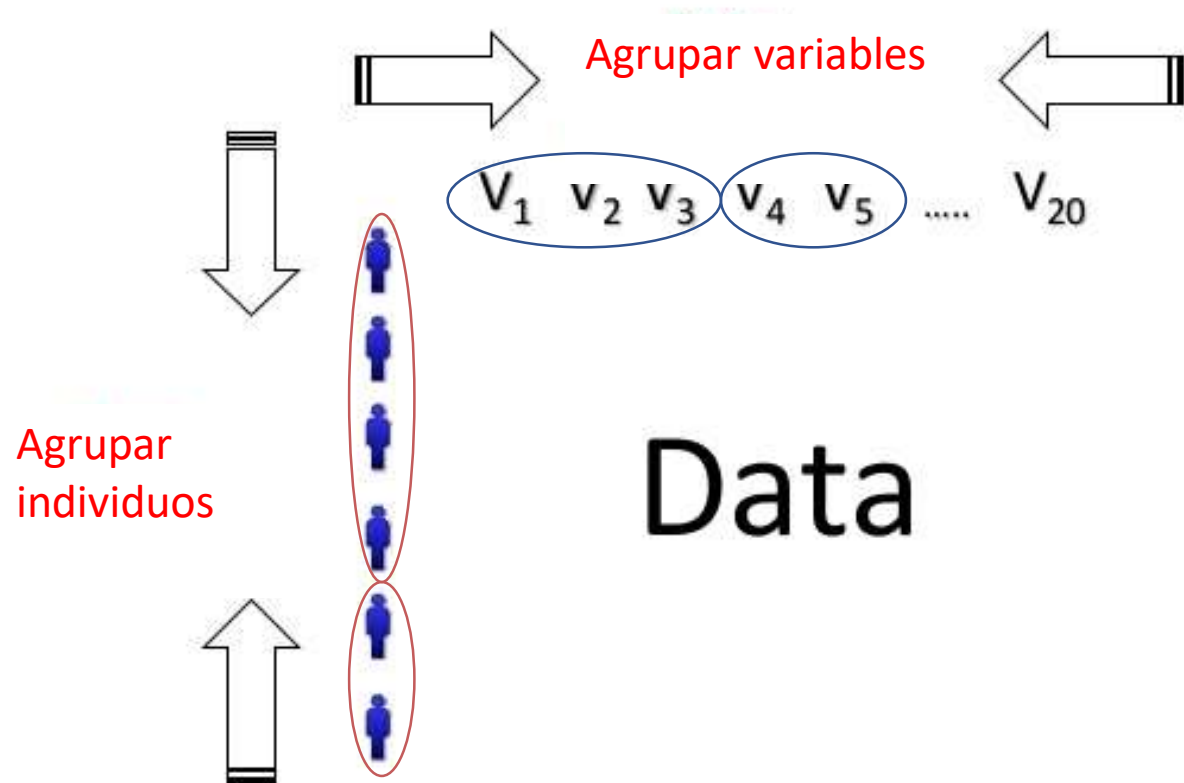
Selección de la medida de asociación
(distancia / similitud / proximidad)

Selección y aplicación del criterio de
agrupación

Determinación de la estructura correcta
(elección del número de grupos)

Clusters de variables o individuos?

Diferencias entre clasificar individuos y variables



Etapas análisis de clústers

Selección de objetos (variables / individuos) a analizar

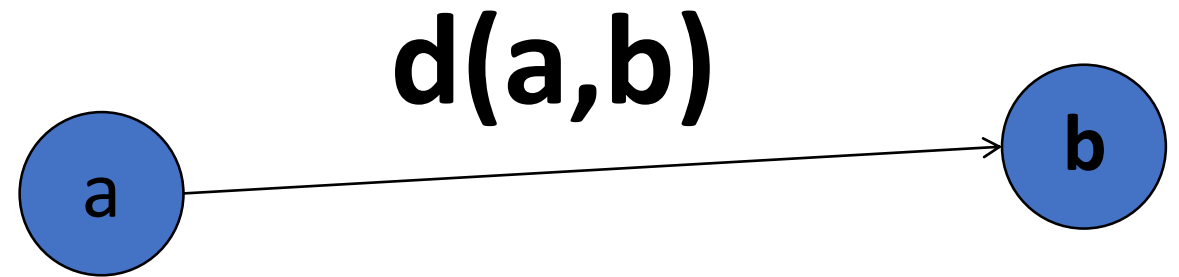
Selección de la medida de asociación (distancia / similitud / proximidad)

Selección y aplicación del criterio de agrupación

Determinación de la estructura correcta (elección del número de grupos)

Medidas de distancia/similitud

- Una vez establecidas las variables, individuos y/o objetos a clasificar se debe establecer una medida de proximidad o de distancia entre ellos que cuantifique el grado de similitud entre cada par de objetos
- Dado dos objetos, "a", "b", podemos obtener una medida de la distancia/similitud, $d(a,b)$ entre ambas



Ejemplos..



- Distancia entre ciudades
- Medida en millas

Matriz de distancias

Atlanta	Chicago	Denver	Houston	Los Angeles	Miami	New York	San Francisco	Seattle	Washington D.C.	City
0	Atlanta
587	0	Chicago
1212	920	0	Denver
701	940	879	0	Houston
1936	1745	831	1374	0	Los Angeles
604	1188	1726	968	2339	0	Miami
748	713	1631	1420	2451	1092	0	.	.	.	New York
2139	1858	949	1645	347	2594	2571	0	.	.	San Francisco
2182	1737	1021	1891	959	2734	2408	678	0	.	Seattle
543	597	1494	1220	2300	923	205	2442	2329	0	Washington D.C.

Datos numéricos continuos

Measure	Formula
D1: Euclidean distance	$d_{ij} = \left[\sum_{k=1}^p w_k^2 (x_{ik} - x_{jk})^2 \right]^{1/2}$
D2: City block distance	$d_{ij} = \sum_{k=1}^p w_k x_{ik} - x_{jk} $
D3: Minkowski distance	$d_{ij} = \left(\sum_{k=1}^p w_k^r x_{ik} - x_{jk} ^r \right)^{1/r} \quad (r \geq 1)$
D4: Canberra distance (Lance and Williams, 1966)	$d_{ij} = \begin{cases} 0 & \text{for } x_{ik} = x_{jk} = 0 \\ \sum_{k=1}^p w_k x_{ik} - x_{jk} / (x_{ik} + x_{jk}) & \text{for } x_{ik} \neq 0 \text{ or } x_{jk} \neq 0 \end{cases}$
D5: Pearson correlation	$\delta_{ij} = (1 - \phi_{ij}) / 2 \text{ with}$ $\phi_{ij} = \frac{\sum_{k=1}^p w_k (x_{ik} - \bar{x}_{i\cdot})(x_{jk} - \bar{x}_{j\cdot})}{\left[\sum_{k=1}^p w_k (x_{ik} - \bar{x}_{i\cdot})^2 \sum_{k=1}^p w_k (x_{jk} - \bar{x}_{j\cdot})^2 \right]^{1/2}}$ $\text{where } \bar{x}_{i\cdot} = \frac{\sum_{k=1}^p w_k x_{ik}}{\sum_{k=1}^p w_k}$
D6: Angular separation	$\delta_{ij} = (1 - \phi_{ij}) / 2 \text{ with}$ $\phi_{ij} = \frac{\sum_{k=1}^p w_k x_{ik} x_{jk}}{\left(\sum_{k=1}^p w_k x_{ik}^2 \sum_{k=1}^p w_k x_{jk}^2 \right)^{1/2}}$

Table 3.1 (b) Dissimilarity matrix calculated from Table 3.1 (a).

	AT	BO	CH	DA	DE	DT	HA	HO	HS	KC	LA	NO	NY	PO	TU	WA
AT	0.00	4.24	2.78	2.79	3.85	3.84	3.29	3.58	2.30	3.21	5.51	3.24	4.87	3.09	2.42	3.58
BO	4.24	0.00	3.59	5.31	4.36	4.78	3.29	3.22	4.04	4.10	6.27	3.98	5.05	4.40	3.40	4.42
CH	2.78	3.59	0.00	3.61	4.39	3.69	3.59	4.66	2.75	3.19	5.56	2.48	4.54	4.22	2.97	3.05
DA	2.79	5.31	3.61	0.00	3.44	2.85	5.09	4.87	1.84	2.27	3.61	2.94	3.94	3.74	3.80	2.90
DE	3.85	4.36	4.39	3.44	0.00	2.48	4.79	3.55	3.37	1.90	2.66	2.47	3.13	2.58	3.69	3.12
DT	3.84	4.78	3.69	2.85	2.48	0.00	5.39	4.62	2.33	1.85	2.88	2.43	1.92	3.58	4.34	1.09
HA	3.29	3.29	3.59	5.09	4.79	5.39	0.00	2.53	4.31	4.65	6.88	4.56	5.69	3.10	1.53	4.86
HO	3.58	3.22	4.66	4.87	3.55	4.62	2.53	0.00	4.02	4.11	5.92	4.55	4.77	2.18	2.52	4.45
HS	2.30	4.04	2.75	1.84	3.37	2.33	4.31	4.02	0.00	2.07	4.31	2.77	3.52	3.51	3.27	1.98
KC	3.21	4.10	3.19	2.27	1.90	1.85	4.65	4.11	2.07	0.00	2.80	1.65	3.25	3.24	3.34	2.19
LA	5.51	6.27	5.56	3.61	2.66	2.88	6.88	5.92	4.31	2.80	0.00	3.40	3.34	4.62	5.62	3.73
NO	3.24	3.98	2.48	2.94	2.47	2.43	4.56	4.55	2.77	1.65	3.40	0.00	3.43	3.63	3.48	2.58
NY	4.87	5.05	4.54	3.94	3.13	1.92	5.69	4.77	3.52	3.25	3.34	3.43	0.00	3.81	4.97	2.07
PO	3.09	4.40	4.22	3.74	2.58	3.58	3.10	2.18	3.51	3.24	4.62	3.63	3.81	0.00	2.32	3.55
TU	2.42	3.40	2.97	3.80	3.69	4.34	1.53	2.52	3.27	3.34	5.62	3.48	4.97	2.32	0.00	3.95
WA	3.58	4.42	3.05	2.90	3.12	1.09	4.86	4.45	1.98	2.19	3.73	2.58	2.07	3.55	3.95	0.00

Datos categóricos binarios

Table 3.3 Similarity measures for binary data.

Measure	Formula
S1: Matching coefficient	$s_{ij} = (a + d) / (a + b + c + d)$
S2: Jaccard coefficient (Jaccard, 1908)	$s_{ij} = a / (a + b + c)$
S3: Rogers and Tanimoto (1960)	$s_{ij} = (a + d) / [a + 2(b + c) + d]$
S4: Sneath and Sokal (1973)	$s_{ij} = a / [a + 2(b + c)]$
S5: Gower and Legendre (1986)	$s_{ij} = (a + d) / \left[a + \frac{1}{2}(b + c) + d \right]$
S6: Gower and Legendre (1986)	$s_{ij} = a / \left[a + \frac{1}{2}(b + c) \right]$

Table 3.2 Counts of binary outcomes for two individuals.

		Individual i		
	Outcome	1	0	Total
Individual j	1	a	b	$a + b$
	0	c	d	$c + d$
	Total	$a + c$	$b + d$	$p = a + b + c + d$

Etapas análisis de clústers

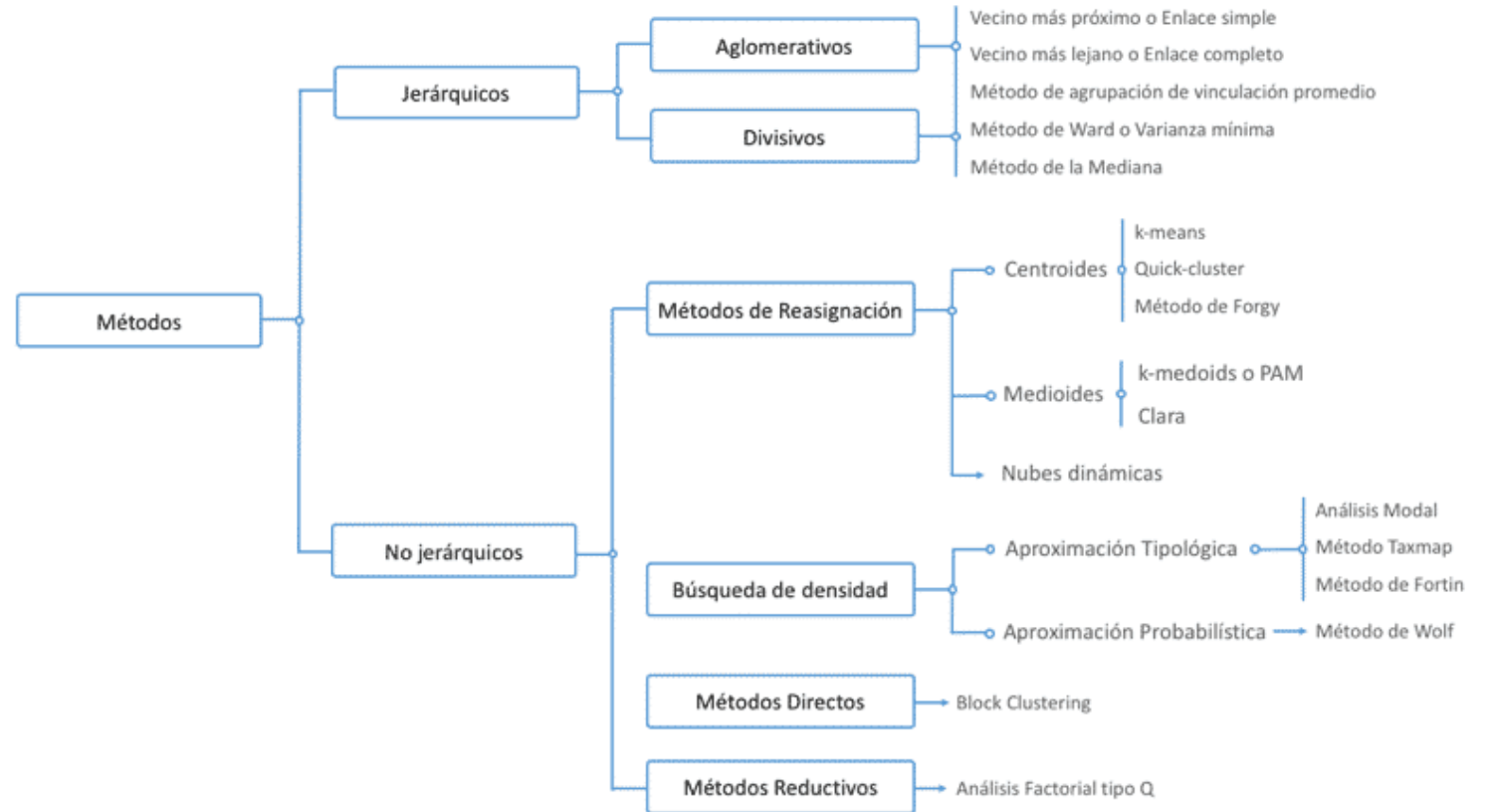
Selección de objetos (variables / individuos) a analizar

Selección de la medida de asociación (distancia / similitud / proximidad)

Selección y aplicación del criterio de agrupación

Determinación de la estructura correcta (elección del número de grupos)

Métodos de clustering

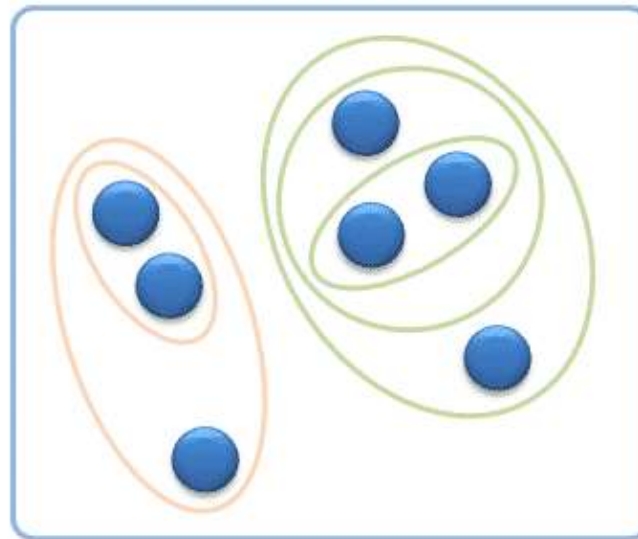


Métodos jerárquicos vs no jerárquicos

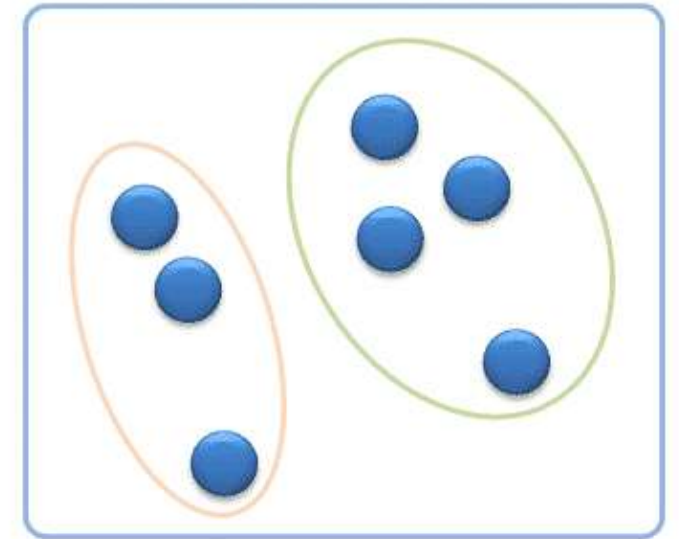
- **MÉTODOS JERÁRQUICOS:** Agrupar clústeres para formar uno nuevo o separar alguno ya existente para dar origen a otros dos de manera que se maximice una medida de similitud o se minimice alguna distancia.
- Cada paso sólo un objeto cambia de grupo y los grupos están anidados en los de pasos anteriores. Si un objeto ha sido asignado a un grupo ya no cambia más de grupo.
- **MÉTODOS NO JERÁRQUICOS:** Están diseñados para la clasificación de individuos (no de variables) en K grupos. El procedimiento es elegir una partición de los individuos en K grupos e intercambiar a los miembros de los clústeres para tener una partición mejor.

Métodos jerárquicos vs no jerárquicos

CLUSTER JERÁRQUICOS
HIERARCHICAL CLUSTER

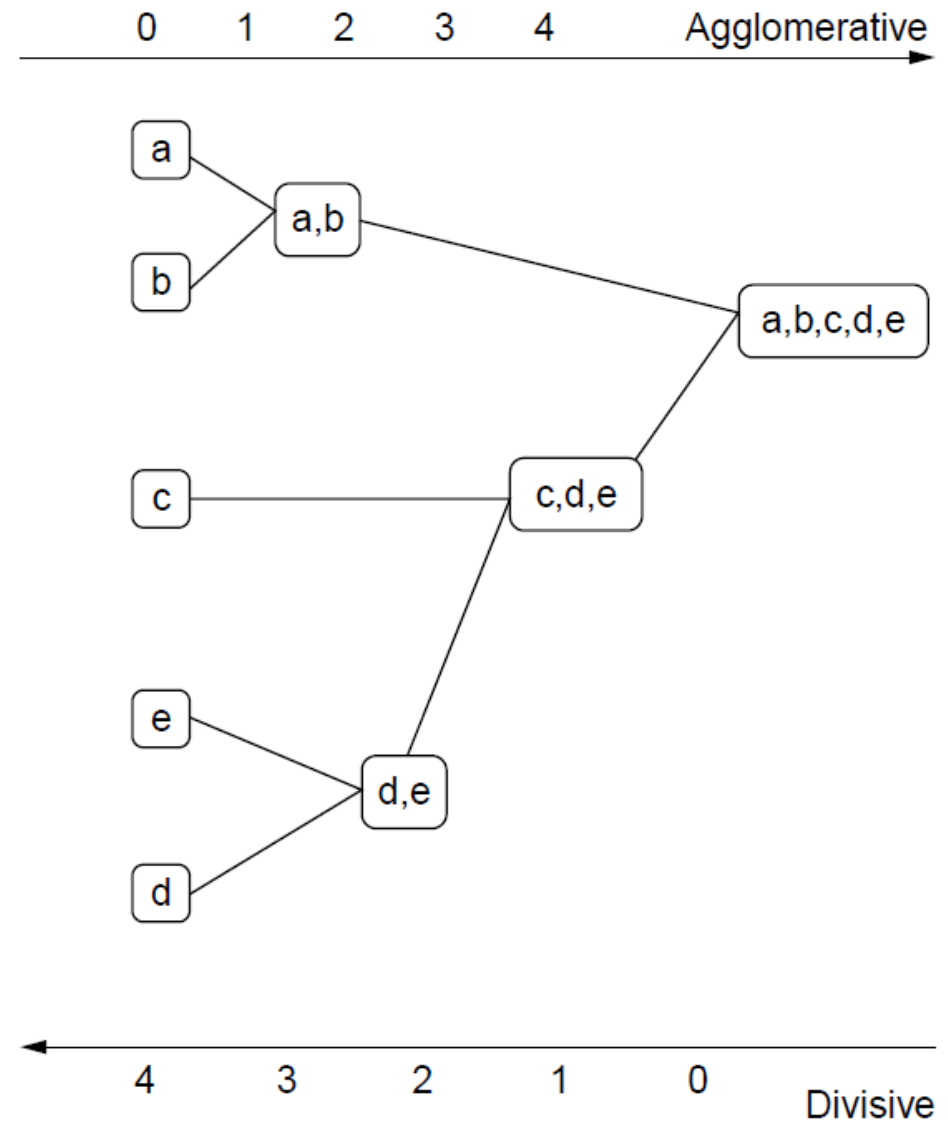


CLUSTER NO JERÁRQUICOS
PARTITIONING CLUSTER

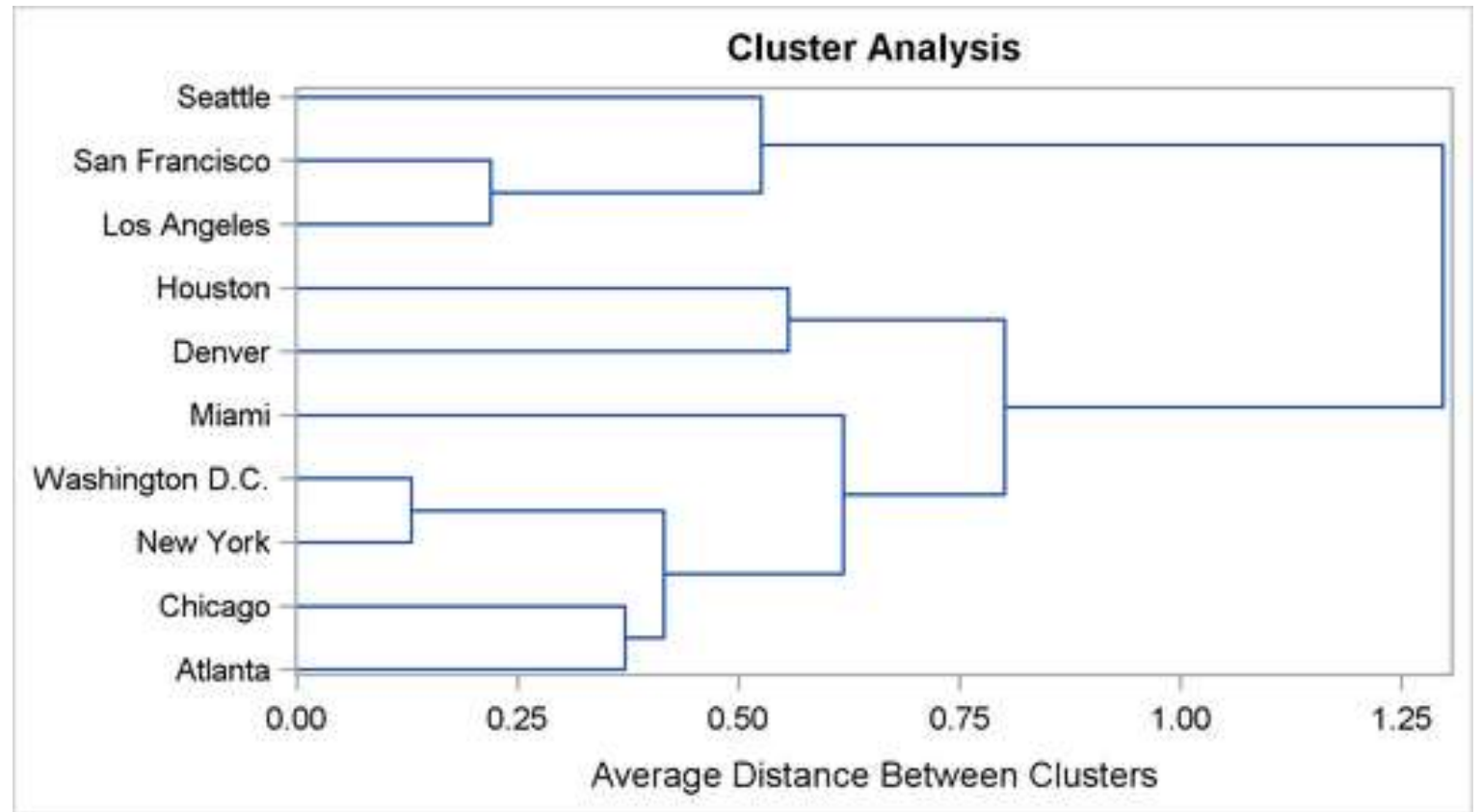


Métodos jerárquicos

- Aglomerativos: Se parte de tantos grupos como individuos hay en el estudio y se van agrupando hasta llegar a tener todos los casos en un mismo grupo.
- Divisivos: Se parte de un solo grupo que contiene todos los casos y a través de sucesivas divisiones se forman grupos cada vez más pequeños.



Dendrogram



Métodos de clúster jerárquicos aglomerativos

Single linkage, vecino más cercano (Sneath, 1957)

Complete linkage, vecino más lejano (Sorensen, 1948)

Average linkage (Sokal and Michener, 1958)

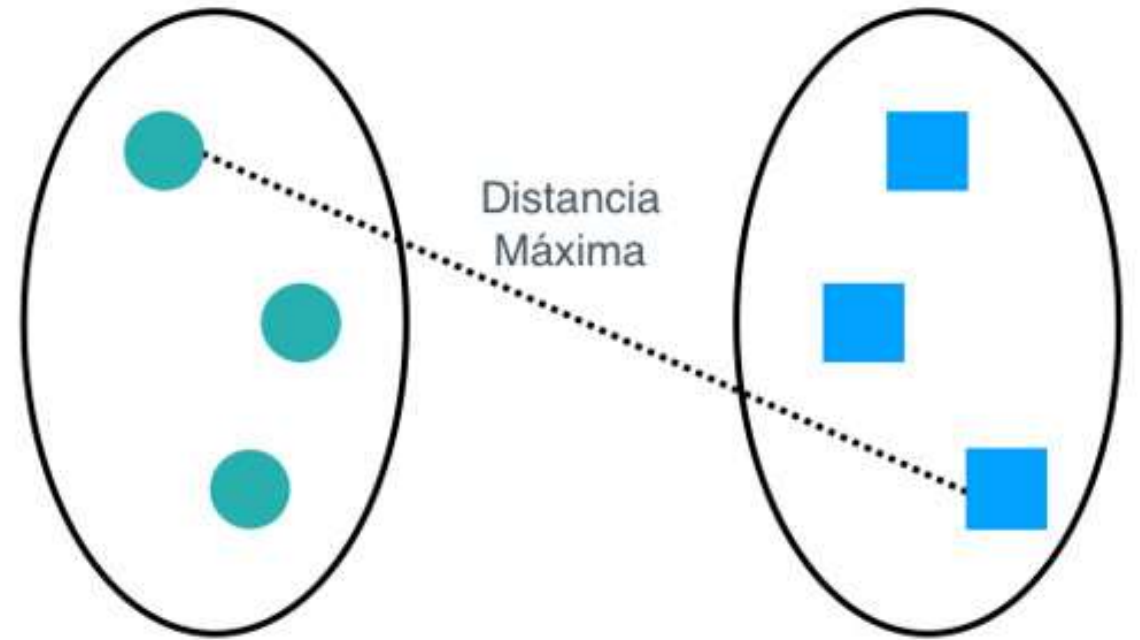
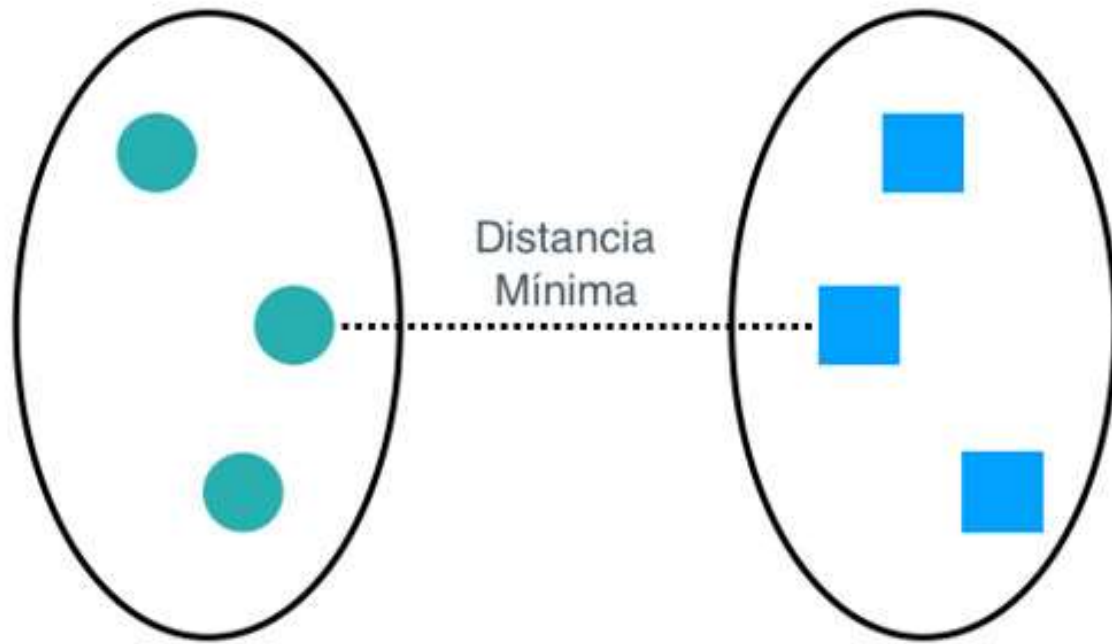
Centroid linkage (Sokal and Michener, 1958)

Weighted average linkage (McQuitty, 1966)

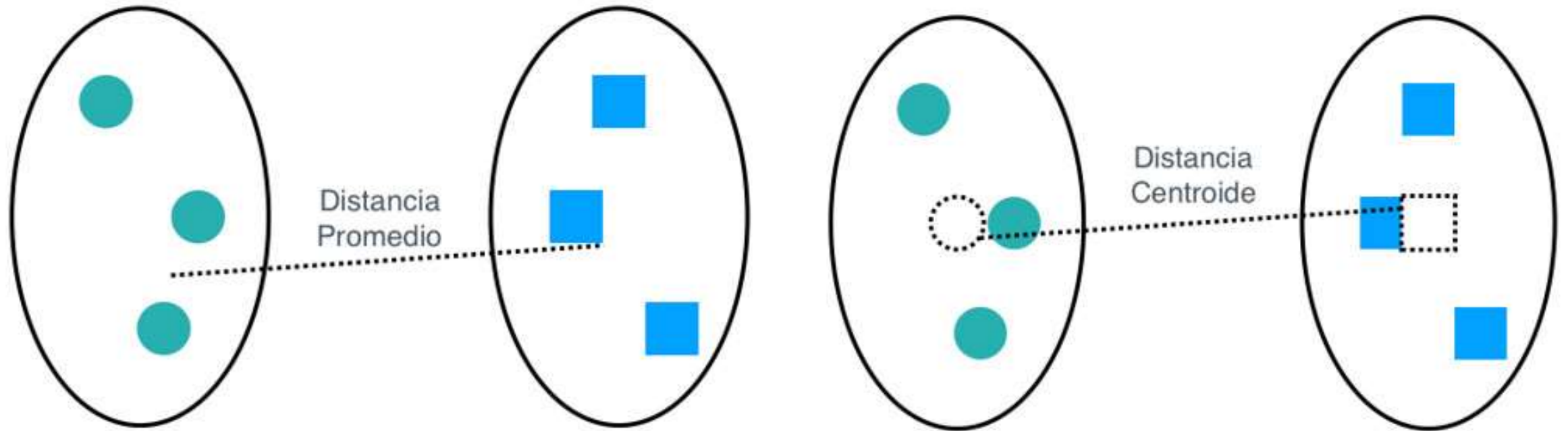
Median linkage (Gower, 1967)

Ward's method (Ward, 1963)

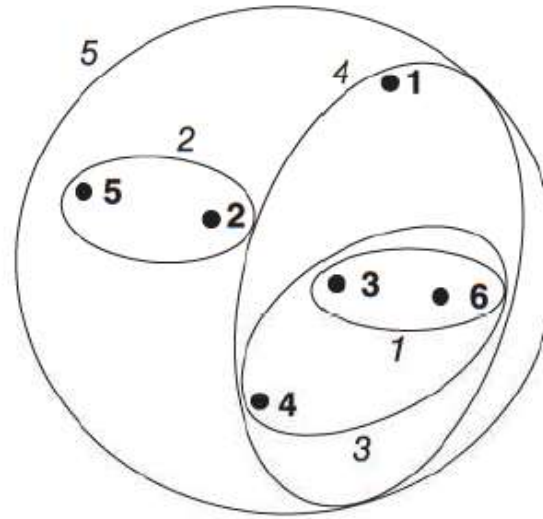
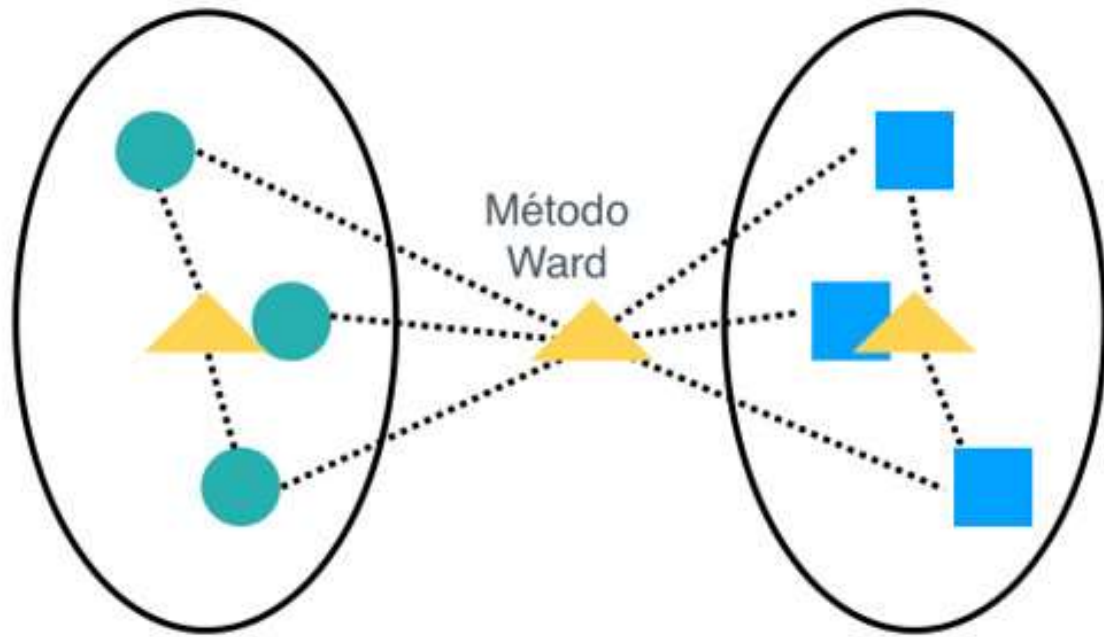
Métodos de clúster jerárquicos



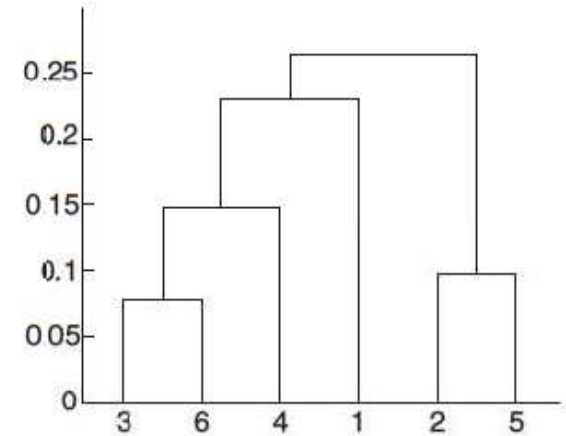
Métodos de clúster jerárquicos



Métodos de clúster jerárquicos

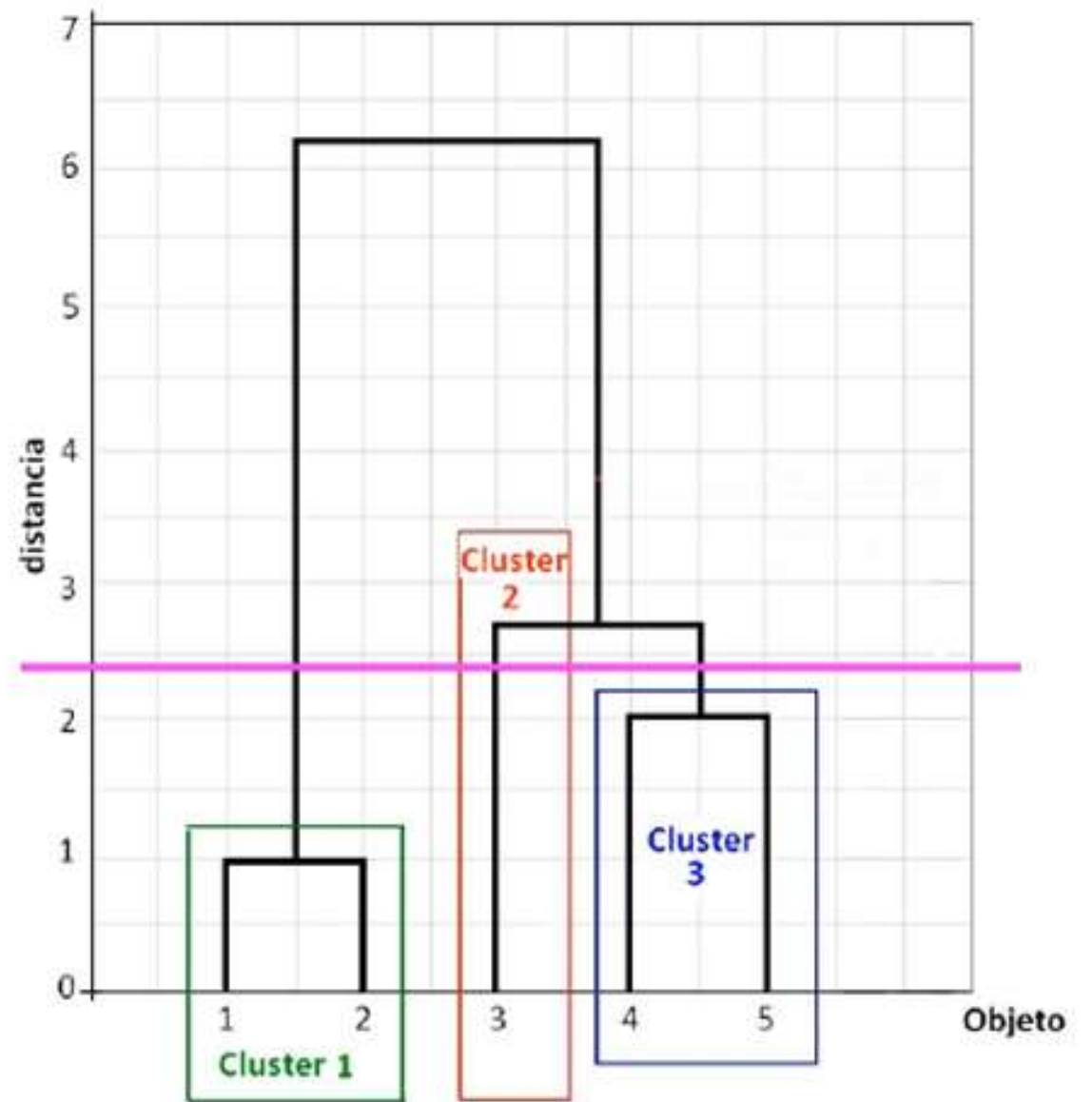


(a) Ward's clustering.

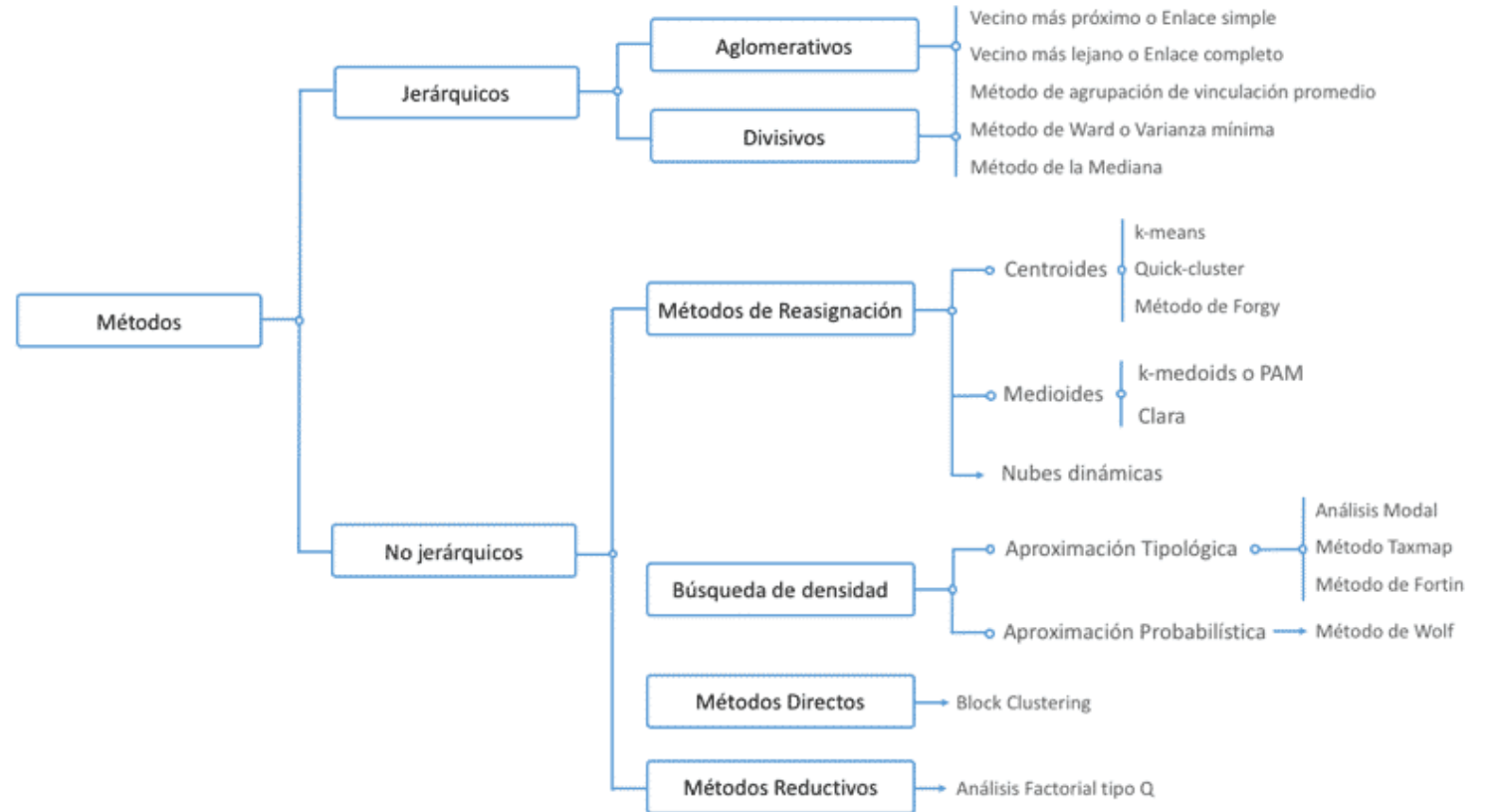


(b) Ward's dendrogram.

Clasificación jerárquica

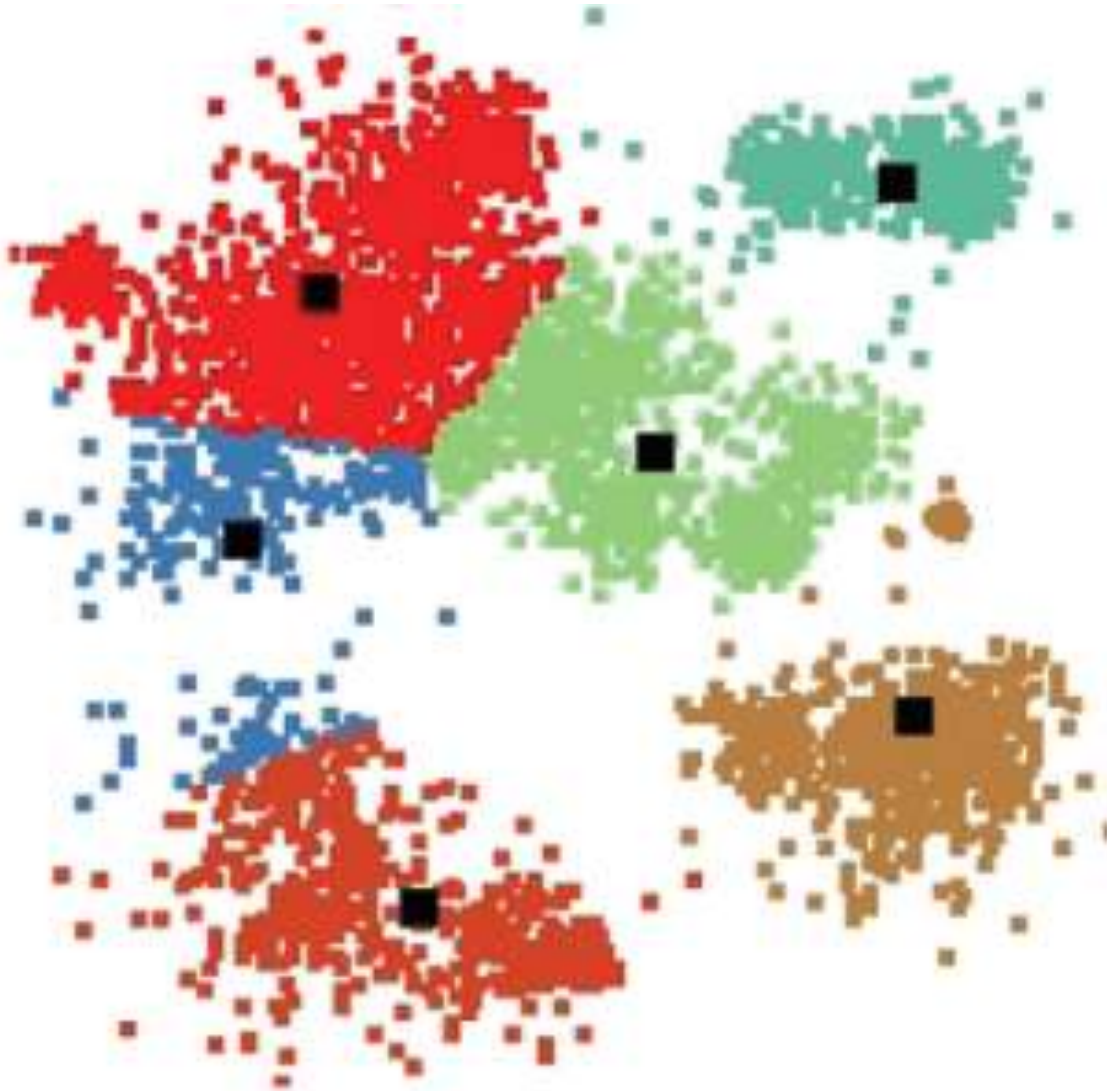


Métodos de clustering



Métodos no jerárquicos

- La idea central de la mayoría de estos procedimientos es elegir alguna partición inicial de individuos y después intercambiar los miembros de estos clusters para obtener una partición mejor.



Métodos de reasignación

k-means (centroides)

k-medoids o PAM (medioides)

Clara (medioides)

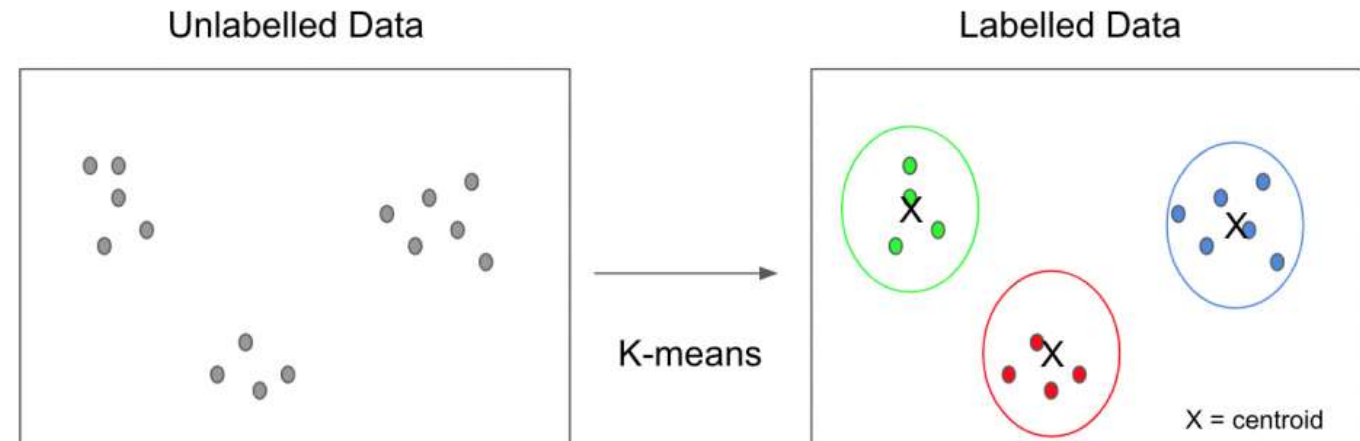
Quick-cluster (centroides)

Método de Forgy (centroides)

Nubes dinámicas

K-medias o k-means

- El término "*k-medias*" fue utilizado por primera vez por James MacQueen en 1967
- El algoritmo de K-Means es una técnica popular de aprendizaje no supervisado para agrupar observaciones



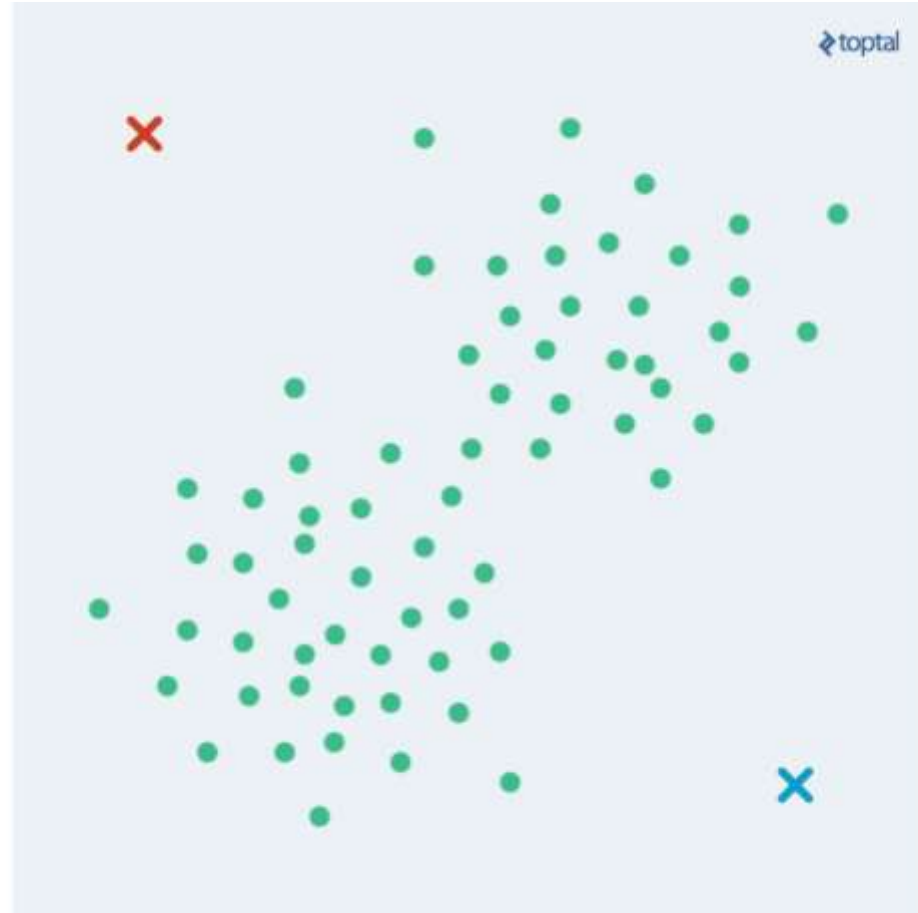
Algoritmo k-means

Definir el número de clústers K :

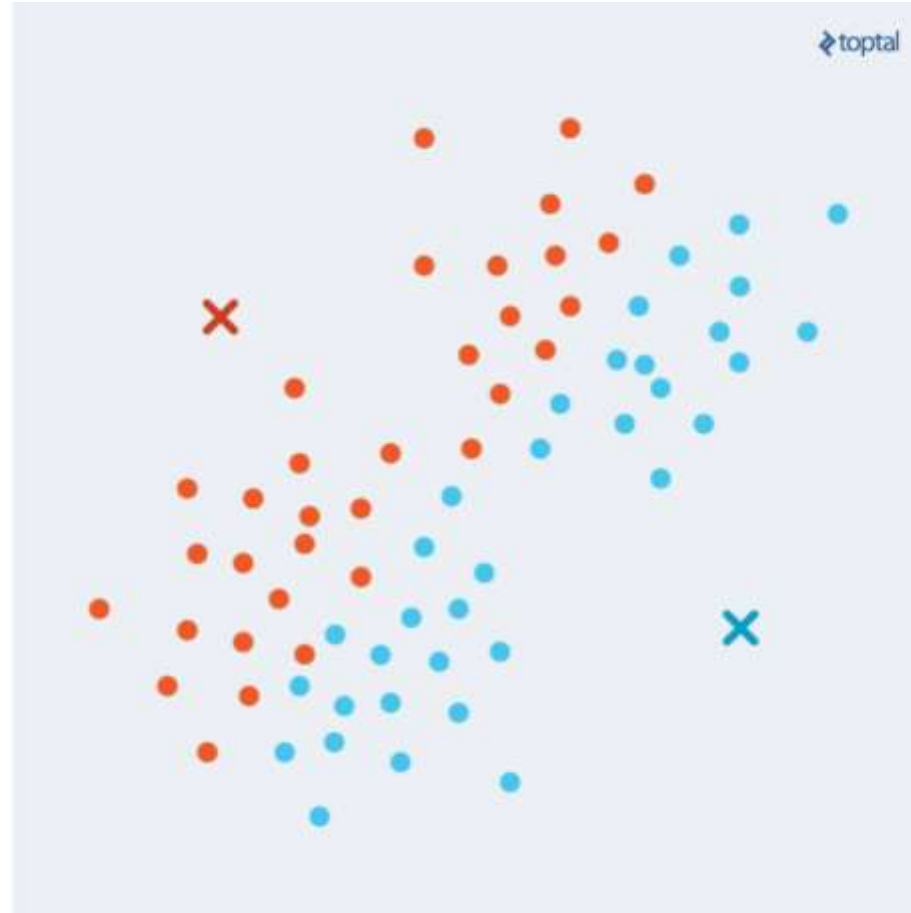
- 1) Coloca K puntos en el espacio representado por los pacientes que están siendo agrupados. Estos puntos representan los centroides del grupo inicial.
- 2) Asigna cada paciente al grupo que tenga el centroide más cercano.
- 3) Cuando todos los pacientes hayan sido asignados, vuelve a calcular las posiciones de los centroides K .

Repite los pasos 2 y 3 hasta que los centroides ya no se muevan. Esto produce una separación de los pacientes en grupos homogéneos al tiempo que maximiza la heterogeneidad entre los grupos.

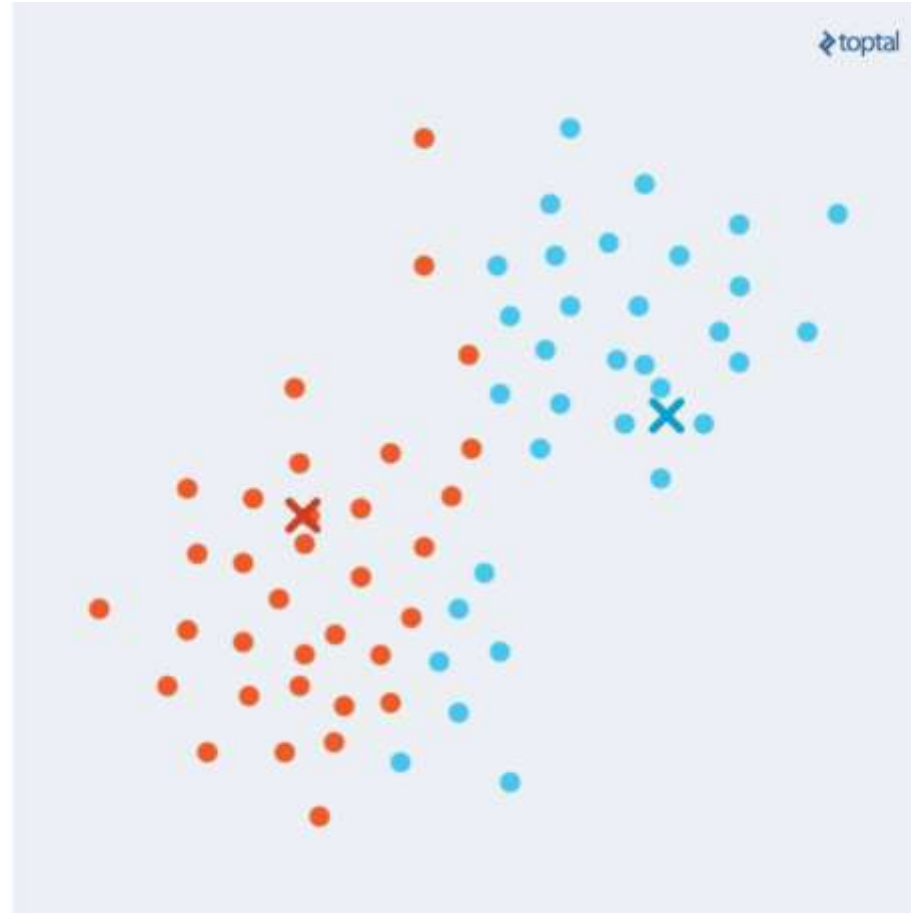
Algoritmo k-means



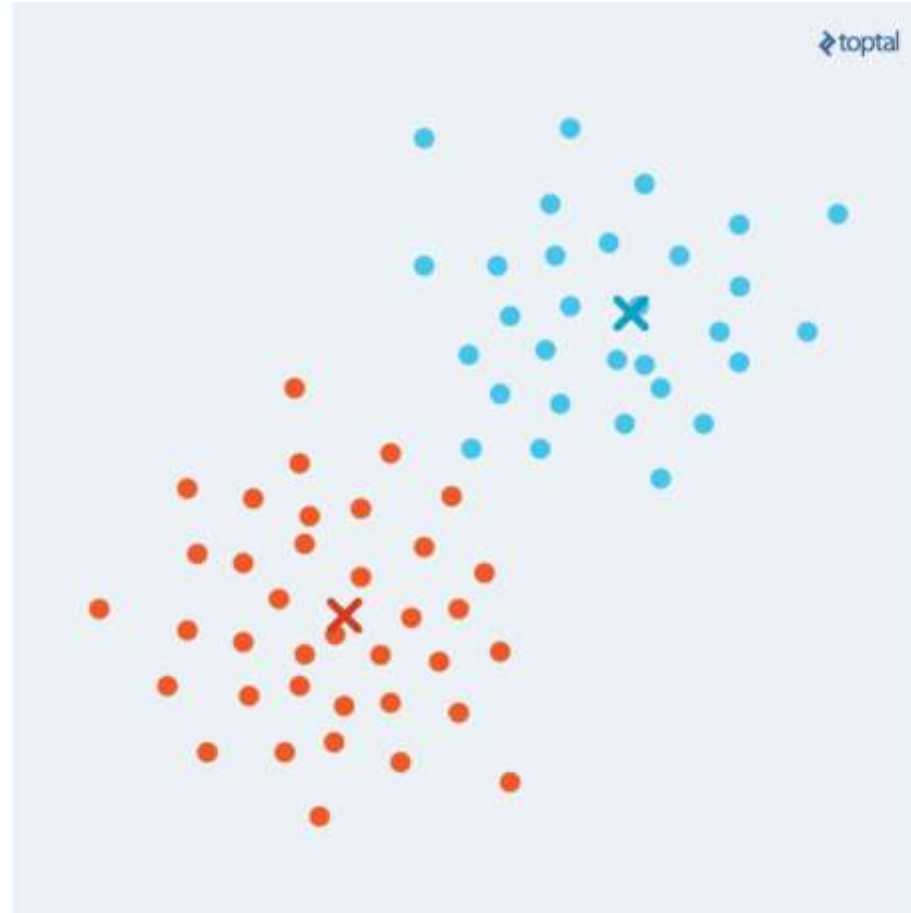
Algoritmo k-means



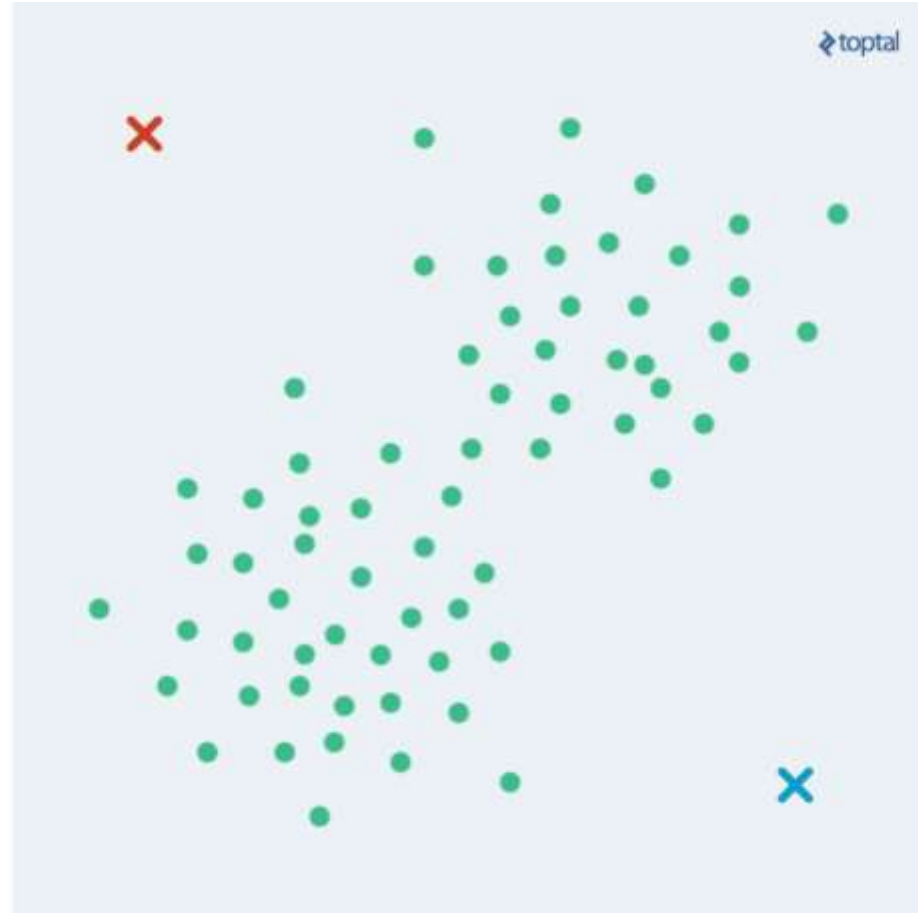
Algoritmo k-means



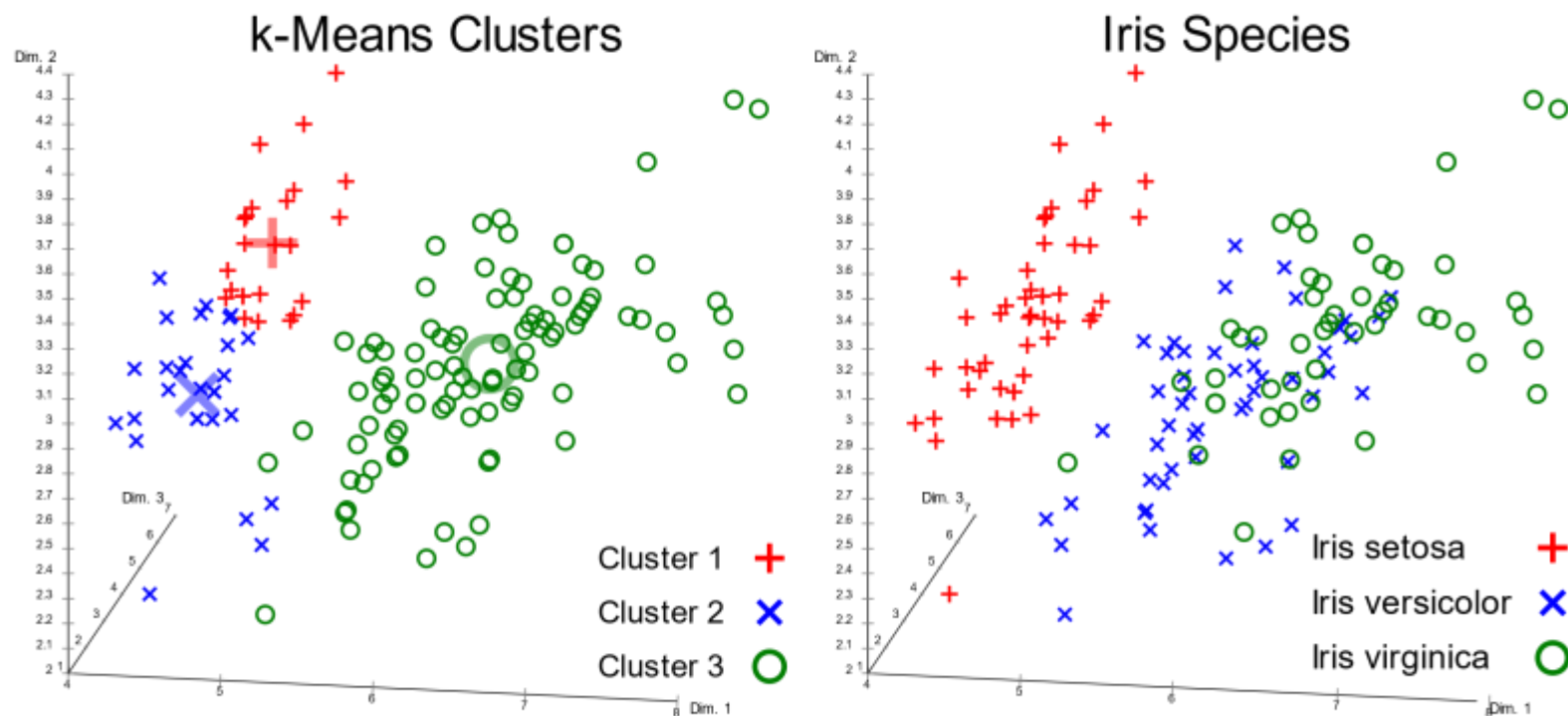
Algoritmo k-means



Algoritmo k-means



Aplicación k-means



Hard vs soft clustering

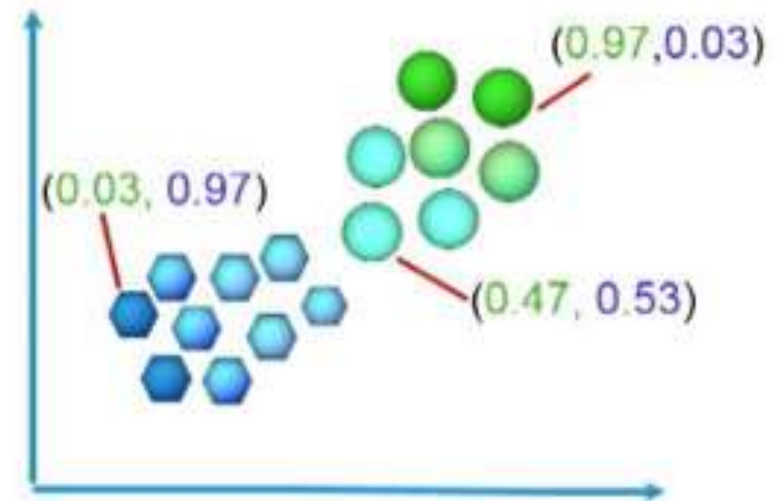


Hard vs soft clustering

Hard Clustering

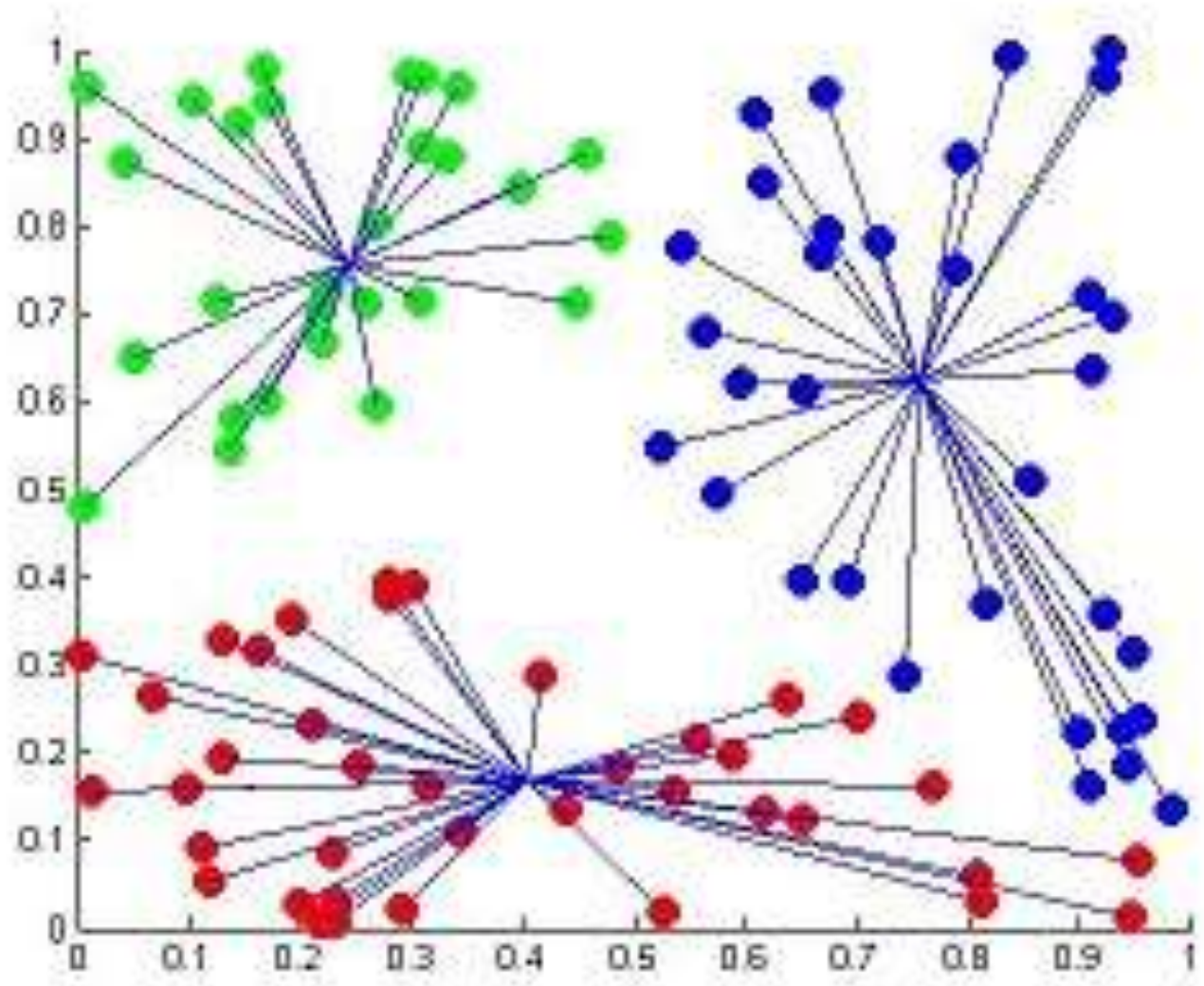


Soft Clustering



Fuzzy c-means

- Un algoritmo de soft clustering propuesto por Bezdek (1974; 1981)
- Los individuos se distribuyen en grupos, formando parte de más de uno
- Centros de clúster (similares a k-medias) pero con fuzziness (borrosidad) para que ese punto pueda pertenecer a más de un clúster



Parámetros fuzzy c-means

m (parámetro de “fuzzication”)

m $\uparrow \uparrow$ individuos distribuidos de manera más equitativa

m $\longrightarrow 1$ Distribución desigual (k-means)

Membership matrix

Distribución de individuos entre conglomerados

	<u>Cluster 1</u>	<u>Cluster 2</u>	<u>Cluster 3</u>
Individual 1	0.8	0.1	0.1
Individual 2	0.02	0.9	0.08
Individual 3	0.3	0.3	0.4
Individual 4	0.03	0.22	0.75

Etapas análisis de clústers

Selección de objetos (variables / individuos) a analizar

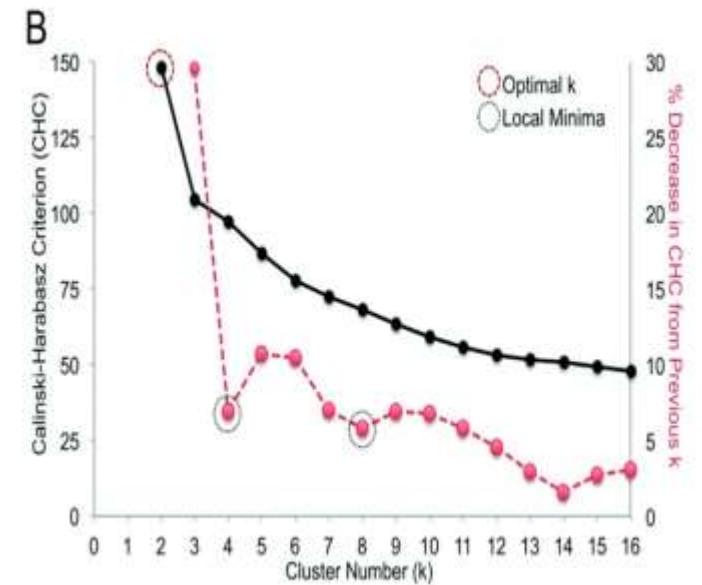
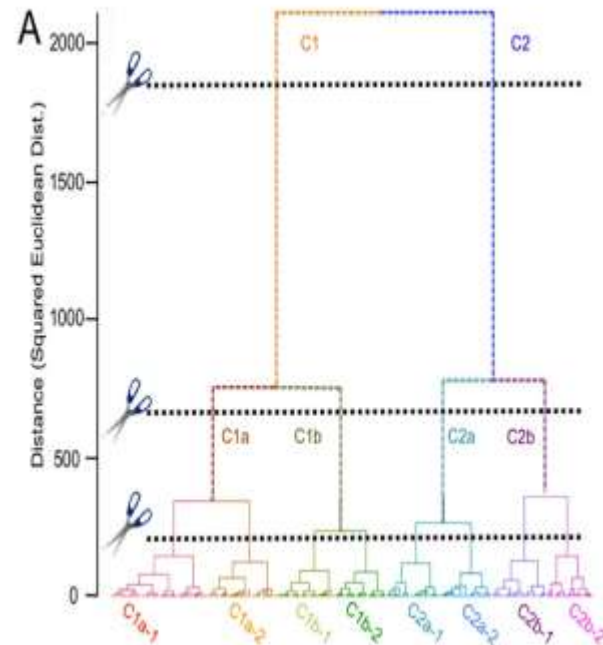
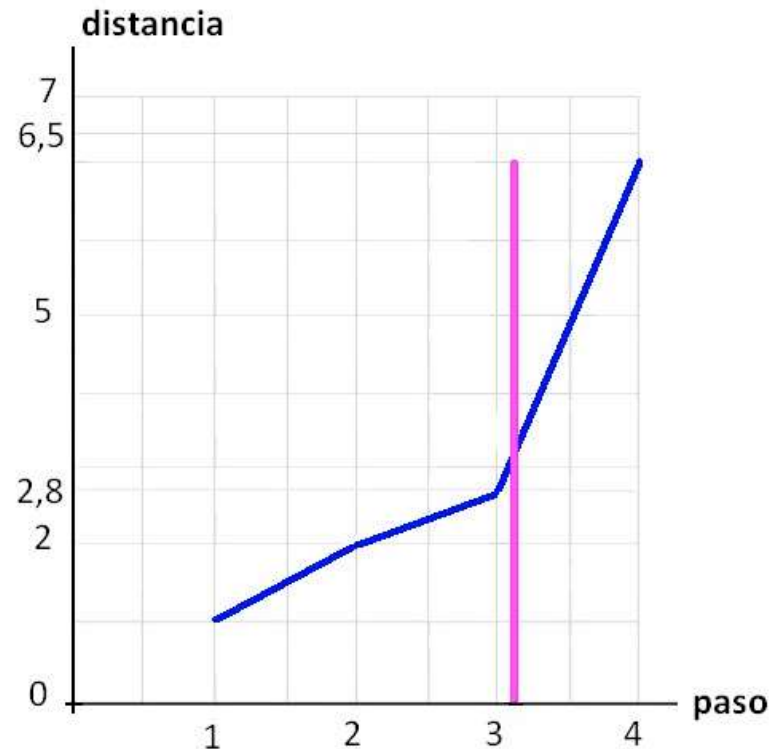
Selección de la medida de asociación (distancia / similitud / proximidad)

Selección y aplicación del criterio de agrupación

Determinación de la estructura correcta (elección del número de grupos)

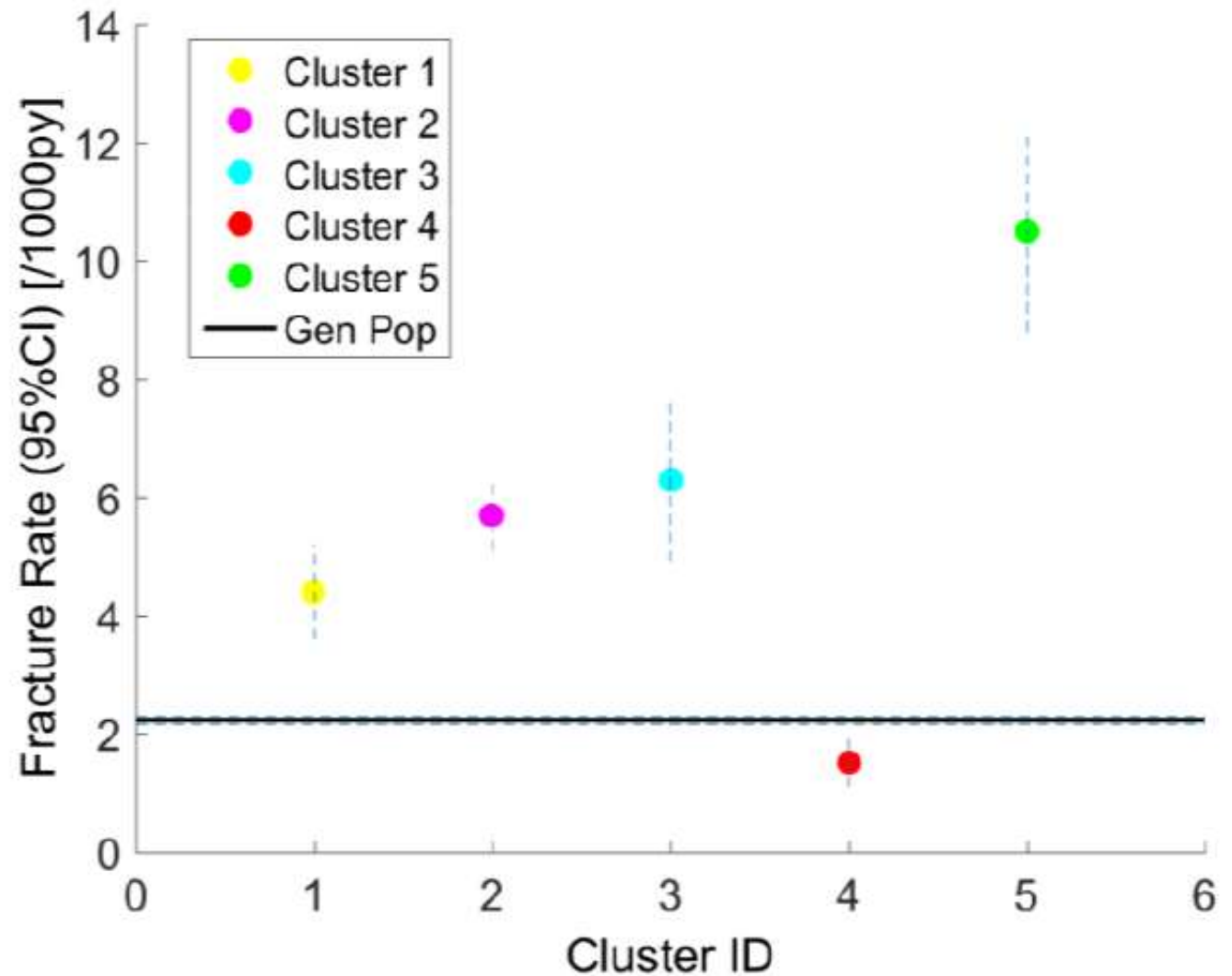
Elección del número de grupos

- Gráficamente
- Índices de validación:
 - Estadístico T^2
 - Pseudo F
 - Calinski Harabasz
 - Fukuyama, ...,



Validación interna

- Validación cruzada
- Bootstrap
- Outcome
- Asesoramiento Experto



Validación externa

- Muestra diferente
- Comparar resultados según criterio de expertos independientes
- Analizar los grupos en función de otras variables no utilizadas

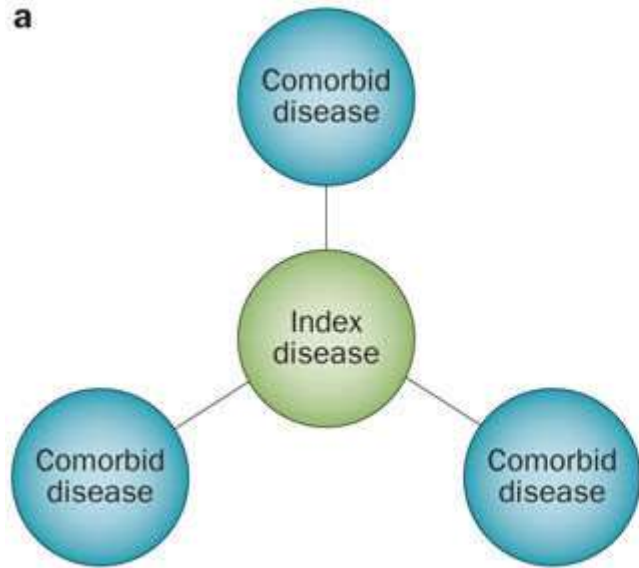


Estudio de los patrones de Multimorbilidad



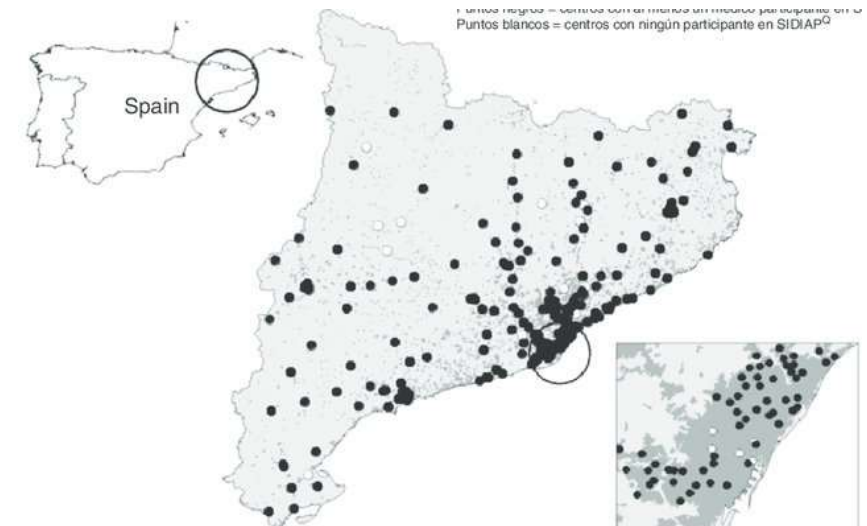
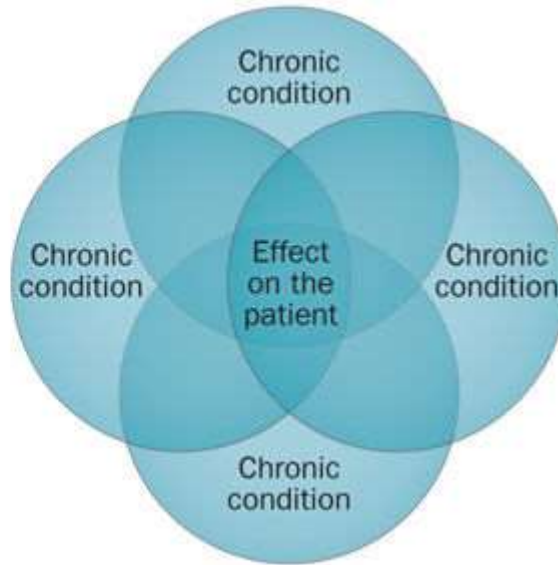
Comorbilidad

a



Multimorbilidad

b




Cluster jerárquicos

PLOS ONE

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Multimorbidity Patterns in Elderly Primary Health Care Patients in a South Mediterranean European Region: A Cluster Analysis

Quintí Foguet-Boreu , Concepción Violán, Teresa Rodríguez-Blanco, Albert Roso-Llorach, Mariona Pons-Vigués, Enriqueta Pujol-Ribera, Yolima Cossio Gil, Jose M. Valderas

Published: November 2, 2015 • <https://doi.org/10.1371/journal.pone.0141155>

Cluster jerárquicos

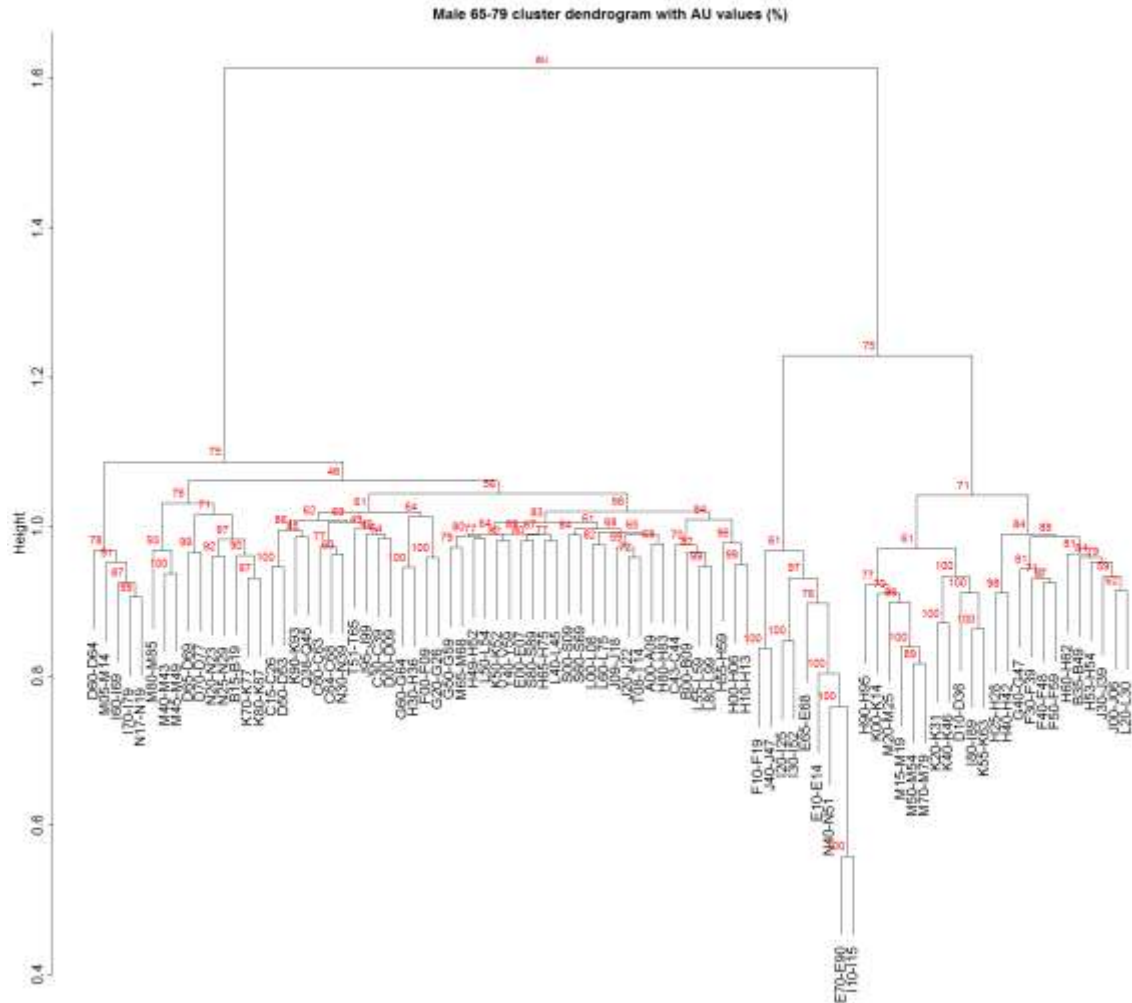


Table 4. Four most prevalent clusters of diagnoses: Prevalence and composition of clusters in men aged 65–79 years (n = 103,512).

Cluster rank	Number of patients	Diagnosis prevalence in stratum (%)		Diagnoses	Prevalence (%)	
		≥ 1 diagnosis	≥ 2 diagnosis		In stratum	In cluster
1	92.419	89.3	63.2	Hypertensive diseases	60.6	67.9
				Metabolic disorders	52.6	58.9
				Diseases of male genital organs	36.1	40.4
				Diabetes mellitus	27.8	31.1
				Obesity and other hyperalimentation	16.6	18.6
2	63.964	61.8	28.1	Other dorsopathies	25.7	41.6
				Arthrosis	22.7	36.7
				Other soft tissue disorders	18.1	29.3
				Diseases of oral cavity, salivary glands and jaws	13.0	21.0
				Other joint disorders	11.9	19.3
				Other disorders of ear	11.6	18.8
3	35.334	34.1	5.6	Chronic lower respiratory diseases	21.5	62.9
				Mental and behavioural disorders due to psychoactive substance use	18.2	53.4
4	31.144	30.1	4.6	Other forms of heart disease	21.1	70.1
				Ischaemic heart diseases	13.6	45.2

Cluster no jerárquicos

Violán et al. *BMC Family Practice* (2018) 19:108
<https://doi.org/10.1186/s12875-018-0790-x>


BMC Family Practice

RESEARCH ARTICLE

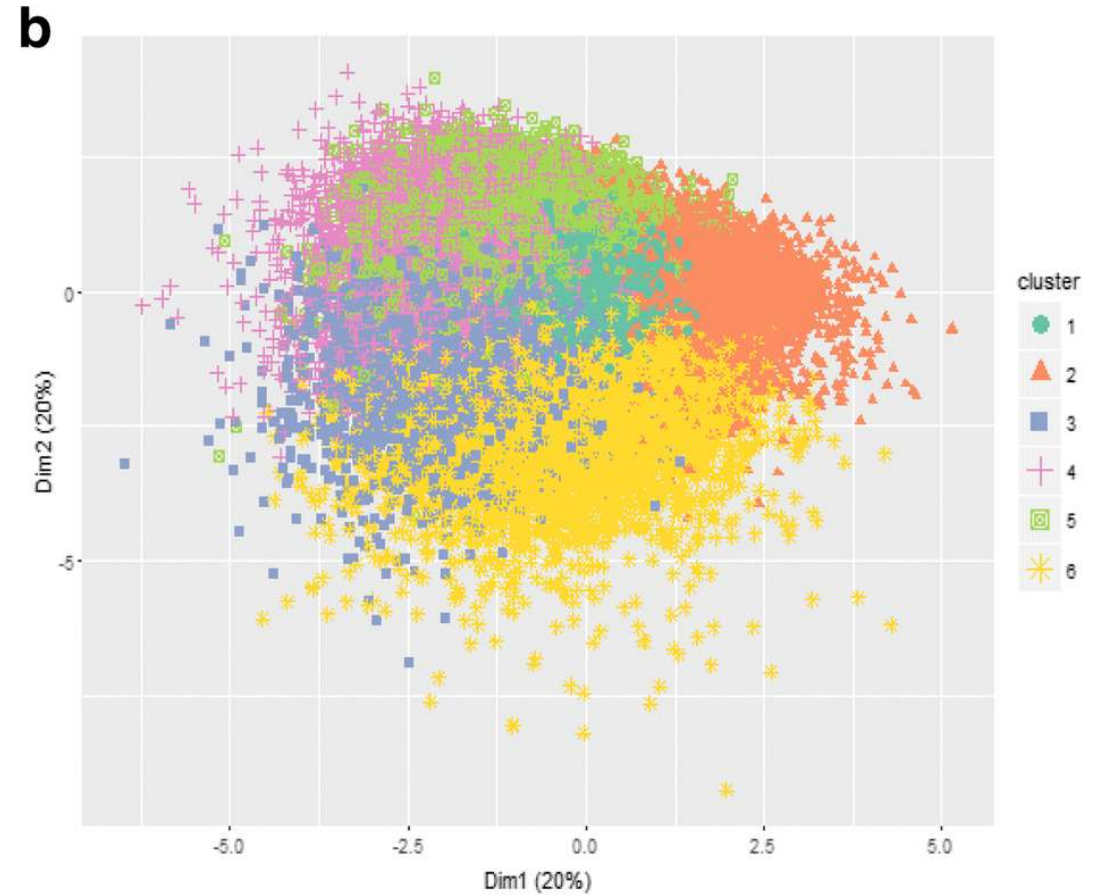
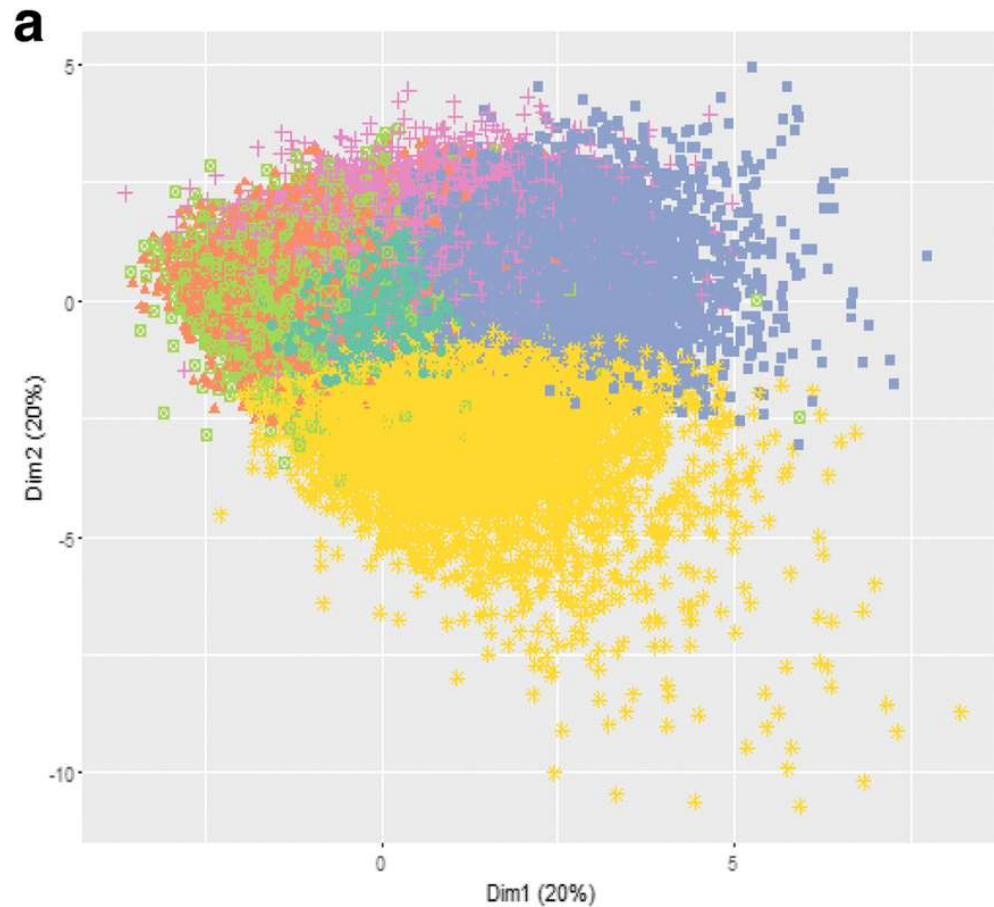
Open Access

Multimorbidity patterns with K-means nonhierarchical cluster analysis



Concepción Violán^{1,2*} , Albert Roso-Llorach^{1,2}, Quintí Foguet-Boreu^{1,2,3}, Marina Guisado-Clavero^{1,2}, Mariona Pons-Vigués^{1,2,4}, Enriqueta Pujol-Ribera^{1,2,4} and Jose M. Valderas⁵

Cluster no jerárquicos



Cluster no jerárquicos

Cluster n (%) ^a	Blocks of diagnoses	Prevalence in cluster (%) ^b	Prevalence in men (%) ^c	O/E ratio ^d	Exclusivity (%)	Centrality	Mean Age	Median number of diagnoses
1 73,979 (38.7)	E70-E90:Metabolic disorders	38.4	42.2	0.91	35.3	0.8	53.3	3
	I10-I15:Hypertensive diseases	28.1	32.5	0.86	33.4			
	F10-F19:Mental and behavioural disorders due to psychoactive substance use	25.4	33.6	0.76	29.2			
	M50-M54:Other dorsopathies	20.8	27.8	0.75	28.9			
	M70-M79:Other soft tissue disorders	10.7	16.9	0.63	24.6			
	E65-E68:Obesity and other hyperalimentation	10.6	14.6	0.73	28.2			
2 28,951 (15.1)	F10-F19:Mental and behavioural disorders due to psychoactive substance use	77.3	33.6	2.30	34.9	1.5	52.6	4
	E70-E90:Metabolic disorders	26.4	42.2	0.63	9.5			
	F40-F48:Neurotic, stress-related and somatoform disorders	25.1	13.5	1.86	28.1			
	M50-M54:Other dorsopathies	23.7	27.8	0.85	12.9			
	K00-K14:Diseases of oral cavity, Salivary glands and jaws	23.2	12.0	1.93	29.2			
	J40-J47:Chronic lower respiratory diseases	19.4	9.3	2.09	31.6			
	F30-F39:Mood [affective] disorders	17.0	6.3	2.72	41.2			
	B15-B19:Viral hepatitis	16.6	3.2	5.13	77.6			
	I10-I15:Hypertensive diseases	14.2	32.5	0.44	6.6			
	K70-K77:Diseases of liver	12.5	5.2	2.38	36.1			
	K20-K31:Diseases of oesophagus, Stomach and duodenum	12.3	11.5	1.06	16.1			
	M70-M79:Other soft tissue disorders	10.4	16.9	0.62	9.4			
3 22,458 (11.8)	E70-E90:Metabolic disorders	43.4	42.2	1.03	12.1	1.9	55.2	6
	K20-K31:Diseases of oesophagus, Stomach and duodenum	40.0	11.5	3.47	40.7			
	K40-K46:Hernia	31.3	8.8	3.57	41.9			
	N40-N51:Diseases of male genital organs	30.9	12.1	2.54	29.9			
	I10-I15:Hypertensive diseases	30.3	32.5	0.93	10.9			

Soft clustering



BMJ Open Soft clustering using real-world data for the identification of multimorbidity patterns in an elderly population: cross-sectional study in a Mediterranean population

Concepción Violán,^{1,2} Quintí Foguet-Boreu,^{1,3,4} Sergio Fernández-Bertolín,^{1,2} Marina Guisado-Clavero,^{1,2} Margarita Cabrera-Bean,⁵ Francesc Formiga,⁶ Jose Maria Valderas,^{1,7} Albert Roso-Llorach^{1,2}

Soft clustering

Table 3 Most frequent 15 diseases found in multimorbidity patterns in individuals aged 65–94 years (n=916619, Catalonia, 2012)

Pattern	Disease	O	O/E ratio	EX	Pattern	Disease	O	O/E ratio	EX
1 Nervous and digestive (n=40 037)	Parkinson and parkinsonism	38.7	17	74.3	2 Respiratory, circulatory and nervous (n=50 639)	Asthma	34.5	7.2	40
	Other neurological diseases	49.5	15.9	69.4		Peripheral vascular disease	13.9	4.2	22.9
	Chronic liver diseases	13.2	5.4	23.4		Parkinson and parkinsonism	8.5	3.8	20.8
	Chronic pancreas, biliary tract and gall bladder diseases	7.9	2.7	11.6		Other neurological diseases	11.7	3.7	20.7
	Dementia	14.7	2.3	9.9		COPD, emphysema, chronic bronchitis	31	2.6	14.3
	Other digestive diseases	4.8	2	8.7		Allergy	10.8	2.4	13.5
	Cerebrovascular disease	16.9	1.9	8.4		Heart failure	16.6	2	11.3
	Colitis and related diseases	24.1	1.7	7.3		Ischaemic heart disease	21.1	2	11.2
	Other metabolic diseases	3.4	1.7	7.2		Other eye diseases	14	1.9	10.3

Soft clustering

Table 4 Variables characterising each cluster in baseline study for 2% prevalence cut-off point (n=916619)

	1 Nervous and digestive	2 Respiratory, circulatory and nervous	3 Circulatory and digestive	4 Mental, nervous and digestive, female dominant	5 Mental, digestive and blood, female oldest-old dominant	6 Nervous, musculoskeletal and circulatory, female dominant	7 Genitourinary, mental and musculoskeletal, male dominant	8 Non- specified, youngest-old dominant	All
Number of people	40 037	50 639	67 492	94 453	106 845	145 074	173 746	238 333	916 619
Multimorbidity, n (%)	39 776 (99.3)	50 513 (99.8)	67 443 (99.9)	94 442 (100.0)	106 696 (99.9)	144 869 (99.9)	171 983 (99.0)	177 363 (74.4)	853 085 (93.1)
Polypharmacy, n (%)	28 484 (71.1)	38 869 (76.8)	54 658 (81.0)	64 154 (67.9)	71 830 (67.2)	86 317 (59.5)	90 603 (52.1)	52 588 (22.1)	487 502 (53.1)
Women, n (%)	22 628 (56.5)	26 690 (52.7)	38 023 (56.3)	78 922 (83.6)	85 735 (80.2)	113 629 (78.3)	15 730 (9.1)	147 773 (62.0)	529 131 (57.7)
Men, n (%)	17 409 (43.5)	23 949 (47.3)	29 469 (43.7)	15 531 (16.4)	21 110 (19.8)	31 445 (21.7)	158 016 (90.9)	90 560 (38.0)	387 488 (42.3)
Age (categories), n (%)									
(65, 70)	7188 (18.0)	10 400 (20.5)	7233 (10.7)	28 305 (30.0)	12 036 (11.3)	38 829 (26.8)	52 003 (29.9)	96 184 (40.4)	252 178 (27.5)
(70, 80)	17 804 (44.5)	22 743 (44.9)	24 724 (36.6)	40 577 (43.0)	33 624 (31.5)	70 643 (48.7)	84 037 (48.4)	100 435 (42.1)	394 586 (43.0)
(80, 90)	13 460 (33.6)	15 568 (30.7)	29 908 (44.3)	22 638 (24.0)	48 453 (45.3)	32 714 (22.6)	34 785 (20.0)	37 217 (15.6)	234 744 (25.6)
(90, 99)	1587 (4.0)	1927 (3.8)	5628 (8.3)	2934 (3.1)	12 732 (11.9)	2888 (2.0)	2920 (1.7)	4497 (1.9)	35 111 (3.8)

Discusión

Síntesis de las diferencias entre los clusters jerárquicos y no jerárquicos:

JERÁRQUICO	NO JERÁRQUICO
<ul style="list-style-type: none">▪ <i>No exigen una definición previa del número de conglomerados.</i>▪ <i>Llevan a cabo un proceso iterativo, de abajo hacia arriba con $(n-1)$ pasos, partiendo de n grupos para terminar en 1 (aglomerativos).</i>▪ <i>Permite obtener distintos tipos de resultados gráficos y numéricos que facilitan la interpretación de los resultados.</i>▪ <i>Precisan una gran cantidad de cálculos, que en ocasiones limita la posibilidad de aplicación con muestras muy grandes.</i>▪ <i>Pueden aplicarse sobre los casos y sobre las variables.</i>	<ul style="list-style-type: none">▪ <i>Exigen definir previamente el número de clusters.</i>▪ <i>Poseen algunos índices que indican el número óptimo de conglomerados.</i>▪ <i>Proporcionan los valores de los centroides de los grupos, lo que facilita la interpretación.</i>▪ <i>Ofrecen resultados adicionales que permiten seleccionar las variables para la interpretación de los conglomerados.</i>▪ <i>Sólo pueden aplicarse sobre casos. Dan soluciones de tipo óptimo.</i>

Discusión

	Ventajas	Desventajas
Hierarchical CA	<ul style="list-style-type: none">• Ofrece una descripción simple pero completa de las soluciones de clustering• Las medidas de similitud permiten aplicar este análisis a casi cualquier tipo de pregunta de investigación• Genera un conjunto completo de soluciones de clustering de manera conveniente	<ul style="list-style-type: none">• Susceptible al impacto de valores atípicos en los datos• No es susceptible de analizar muestras grandes
K-means CA	<ul style="list-style-type: none">• Resultados menos susceptibles a valores atípicos en los datos, influencia de la medida de distancia elegida o la inclusión de variables inapropiadas o irrelevantes• Puede analizar conjuntos de datos extremadamente grandes	<ul style="list-style-type: none">• Soluciones diferentes para cada conjunto de puntos semilla y sin garantía de agrupación óptima de observaciones• No es eficiente cuando se debe considerar un gran número de posibles soluciones de clúster

Discusión



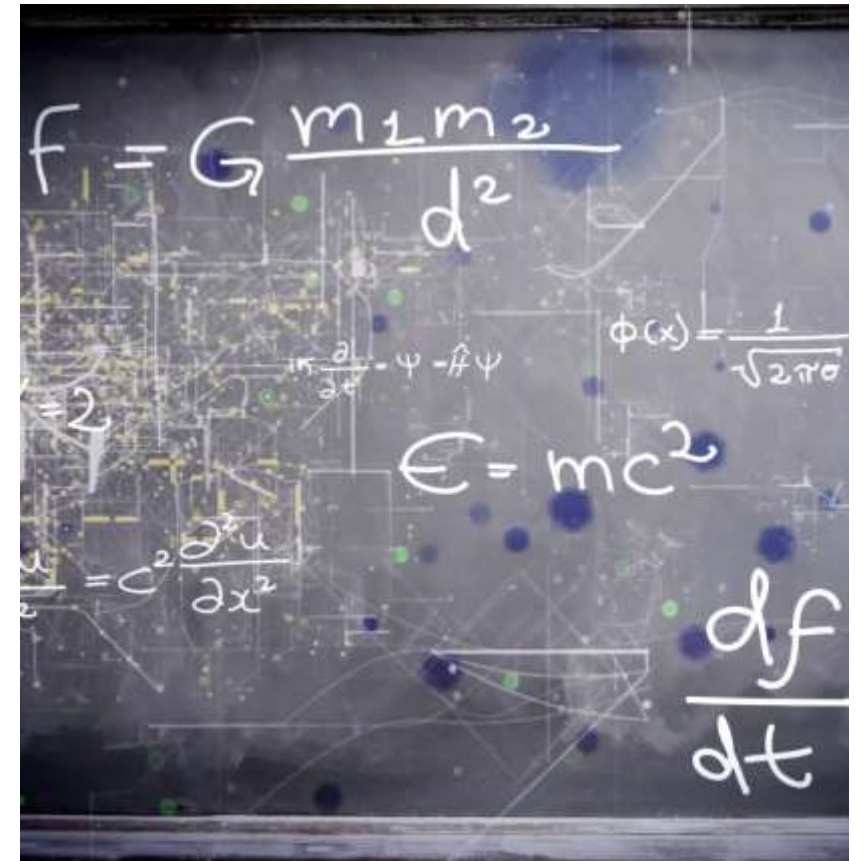
K-means	Fuzzy c-means
<i>Cómo funciona</i>	<i>Cómo funciona</i>
Particiona los datos en k número de clústers mutuamente excluyentes. Qué tan bien encaja un punto en un clúster está determinado por la distancia desde ese punto hasta el centro del clúster.	Agrupación en clústeres basada en particiones mutuamente excluyentes. Qué tan bien encaja un punto en un clúster está determinado por la distancia desde ese punto hasta el centro del clúster.
<i>Mejor utilizado..</i>	<i>Mejor utilizado..</i>
<ul style="list-style-type: none">• Cuando se conoce el número de clústeres• Para una agrupación rápida de grandes conjuntos de datos	<ul style="list-style-type: none">• Cuando se conoce el número de clústeres• Para el reconocimiento de patrones• Cuando los clústeres se superponen

Pasos recomendados en Análisis de clusters

1. Descriptiva exploratoria para detectar posibles errores o la influencia de valores atípicos
2. Seleccionar / determinar numero finito de variables o casos
3. Estandarizar / homogeneizar las medidas en caso necesario
4. Decidir la medida de la distancia / similitud
5. Elección de método
6. Selección de número de clústers
7. Validación

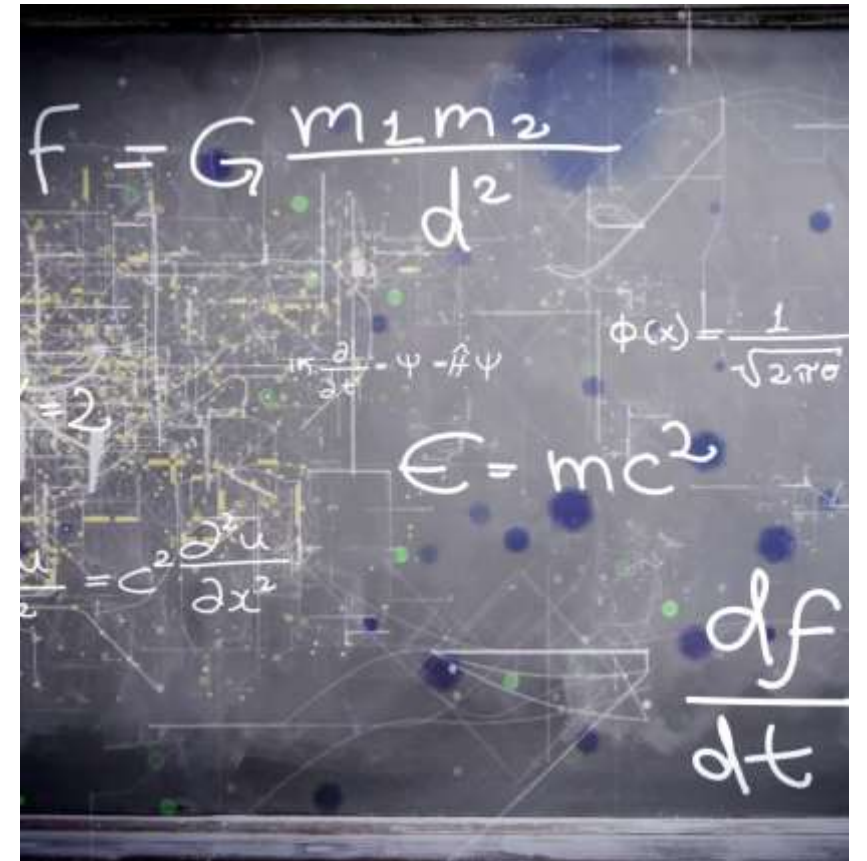
Conclusiones

- Método exploratorio
 - No es un método inferencial (No modelo subyacente, no inferencia)
 - Técnica adecuada para extraer información sin imponer restricciones previas en forma de modelos estadísticos.
 - Reduce la matriz de datos inicial
 - Identificar patrones
- Sensible a la distancia escogida, outliers
- Sensible al tamaño de la muestra, prevalencia



Conclusiones

- No existe un único método de agrupamiento óptimo
- No hay criterio único para determinar el número de conglomerados
- La solución final debe corroborarse/ consensuar según criterios de experto



Análisis de Clusters

Estadísticos

Investigadores

Bibliografia

- Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis: Fifth edition. Cluster Analysis: Fifth Edition. 2011.
- Bezdek JC. Pattern Recognition with Fuzzy Objective Function Algorithms. Pattern Recognit with Fuzzy Object Funct Algorithms. 1981
- Liao M, Li Y, Kianifard F, Obi E, Arcona S. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. BMC Nephrol. 2016;17:25.
- <https://www.diegocalvo.es/cluster-jerarquicos-y-no-jerarquicos/>
- <https://aprendeia.com/algoritmo-agrupamiento-jerarquico-teoria/>

Y
recordad..

