CIENCIA DE DATOS PARA DECISIONES DE NEGOCIO

MÓDULO 10

Clustering jerárquico · Dendrogramas y linkages

De la distancia entre observaciones a una jerarquía interpretable y operativa.

Clases conceptuales · Matemáticas · Evidencia visual · Decisión de negocio



Sergio Gevatschnaider · Material de cátedra

Por qué construir jerarquías, no solo grupos

IDEAS CLAVE

- Investigar estructura anidada sin fijar K desde el inicio.
- Observar el orden y la distancia a los que se fusionan grupos.
- Seleccionar resoluciones apropiadas para distintas decisiones.
- Distinguir jerarquía matemática de una jerarquía causal o empresarial.

PREGUNTAS DE DECISIÓN

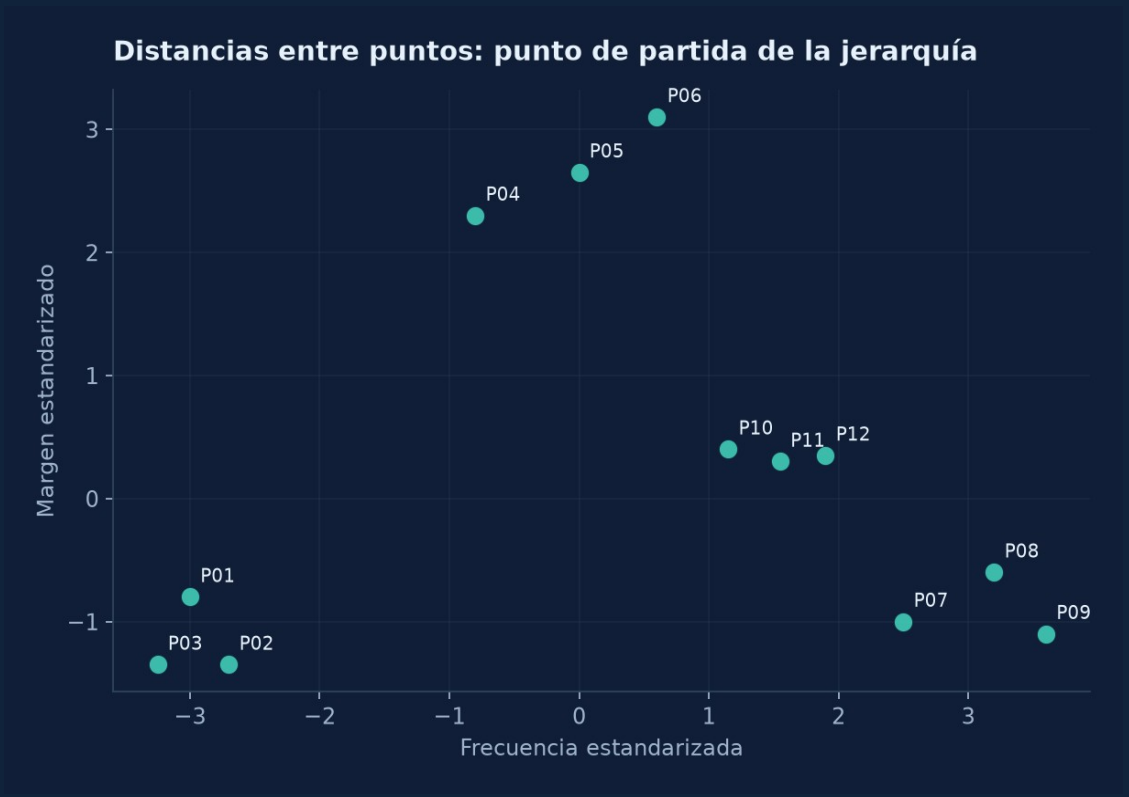
1. ¿Cuántos niveles de segmentación interesan?
2. ¿Qué observaciones se parecen primero?
3. ¿Qué significa fusionar grupos?
4. ¿Qué punto de corte es útil?

IDEA QUE DEBE QUEDAR **Un dendrograma muestra una historia de fusiones, no “categorías naturales” definitivas.**

El mismo caso de 12 clientes facilita seguir fusiones

IDEAS CLAVE

- Cada observación tiene frecuencia y margen estandarizados.
- Primero se calcula una matriz de distancias entre pares.
- Al principio cada cliente forma su propio grupo.
- La estrategia de linkage decide qué grupos se fusionan.

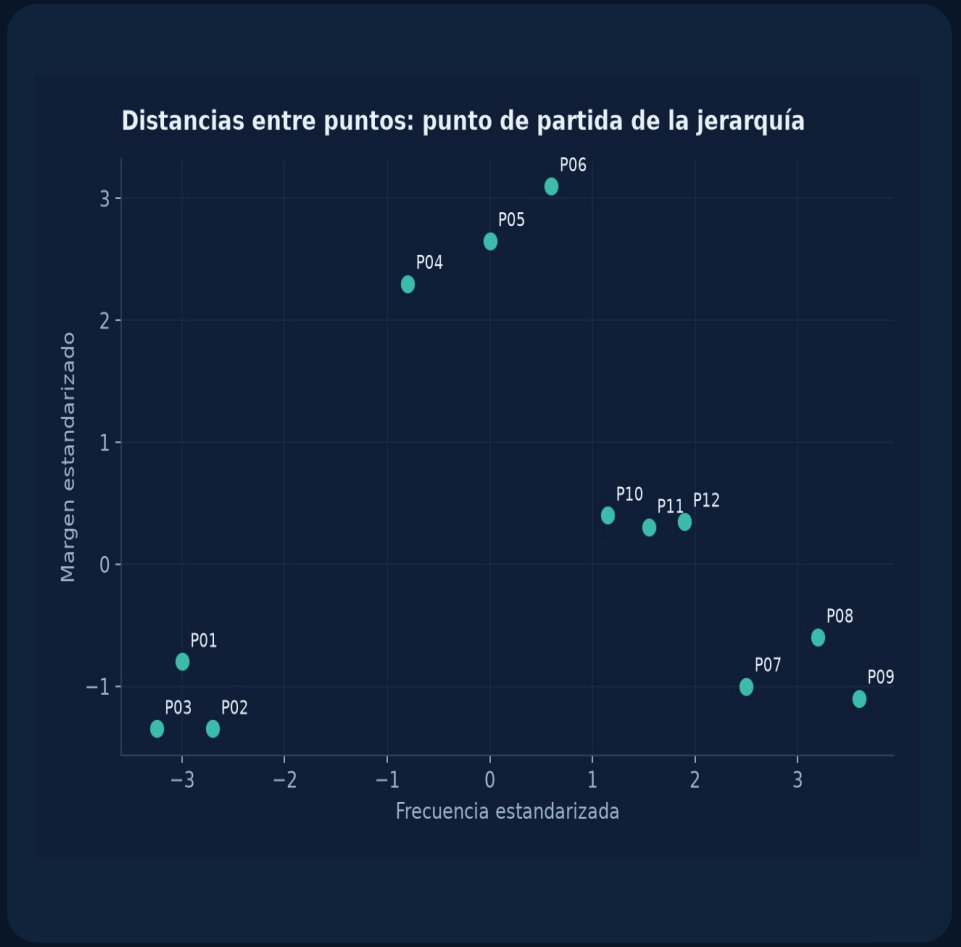


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Con pocos puntos podemos verificar a mano cada paso del algoritmo.**

01 · Cómo nace la jerarquía

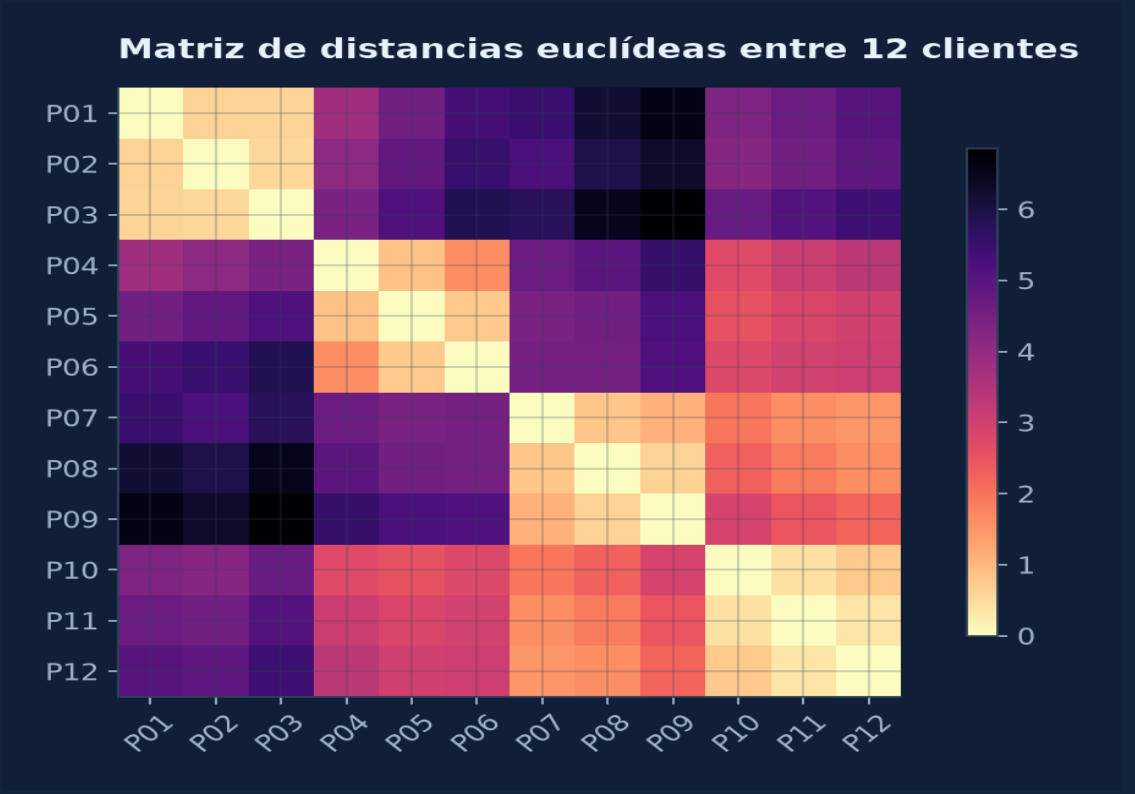
Se comienza con n grupos unitarios y se repite la fusión más cercana hasta obtener uno.



Una matriz de distancias es el verdadero punto de partida

IDEAS CLAVE

- Distancia euclídea $d(i,j) = \sqrt{\Sigma_m (x_{im} - x_{jm})^2}$.
- La diagonal de la matriz vale cero y la matriz es simétrica.
- La escala define cuánto pesa cada variable en la proximidad.
- La tabla de distancias por sí sola no decide cómo combinar clusters.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Métrica y linkage son elecciones distintas, que deben explicitarse.**

Fusión paso a paso: 12 → 11 → … → 1

IDEAS CLAVE

- Buscar el par de clusters con mínima disimilitud según el linkage.
- Unir ese par y registrar altura de fusión.
- Actualizar distancias entre nuevo cluster y restantes.
- Repetir hasta que todos los puntos pertenezcan a un único árbol.

SECUENCIA DE ANÁLISIS

1. 12 clusters individuales
2. Mínimo par de distancias
3. Fusionar y registrar altura
4. Recalcular disimilitudes
5. **Hasta una jerarquía completa**

IDEA QUE DEBE QUEDAR **Una vez fusionados dos grupos, la versión aglomerativa no los separa.**

Ejemplo manual con cuatro observaciones

IDEAS CLAVE

- A=(0,0), B=(0,1), C=(4,0), D=(4,1).
- Distancias AB=CD=1; AC=BD=4; AD=BC=√17.
- Se fusionan AB y CD (hay empate; el orden puede variar).
- Single linkage entre AB y CD usa la menor distancia entre pares: 4.

MODELO MATEMÁTICO

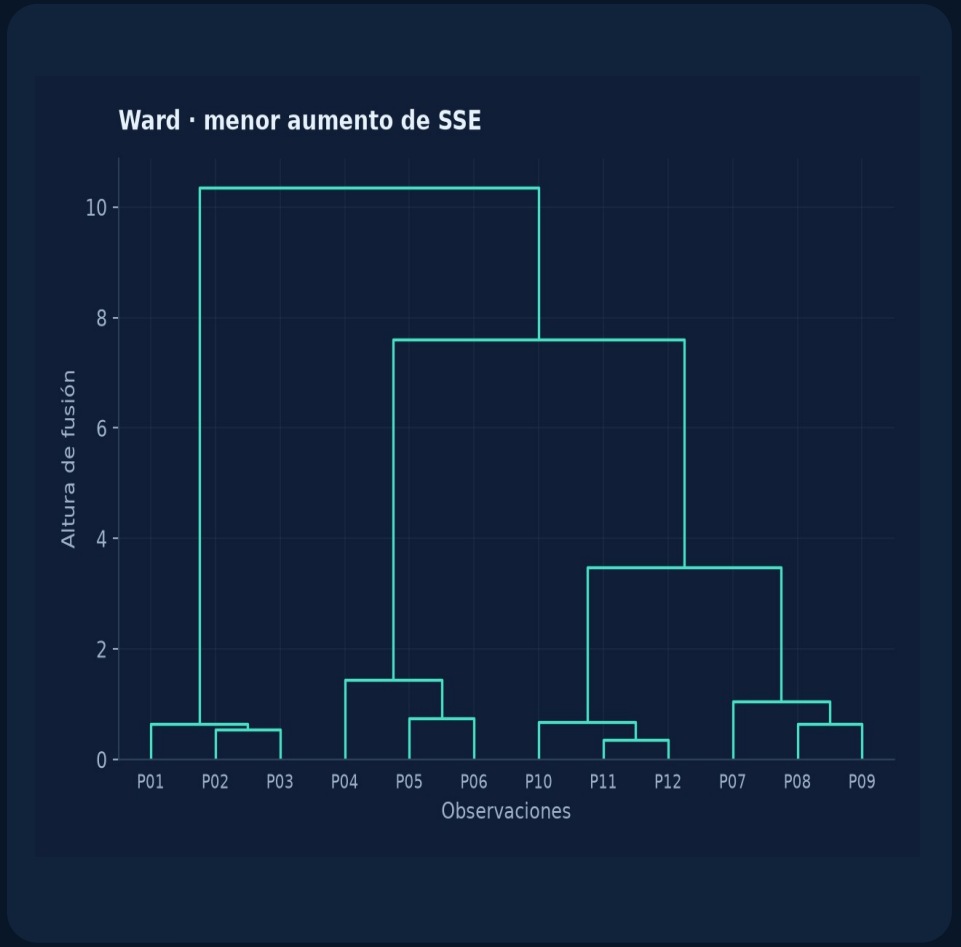
$$d_{single}(AB, CD) = \min\{4, \sqrt{17}, \sqrt{17}, 4\} = 4$$

- Altura 1: dos pares A–B y C–D.
- Altura 4: fusionar ambos clusters.
- Con complete: altura √17 ≈ 4,12.
- La elección de linkage cambia el árbol.

IDEA QUE DEBE QUEDAR **Un dendrograma no es único si hay empates o cambia el criterio de fusión.**

02 · Aprender a leer dendrogramas

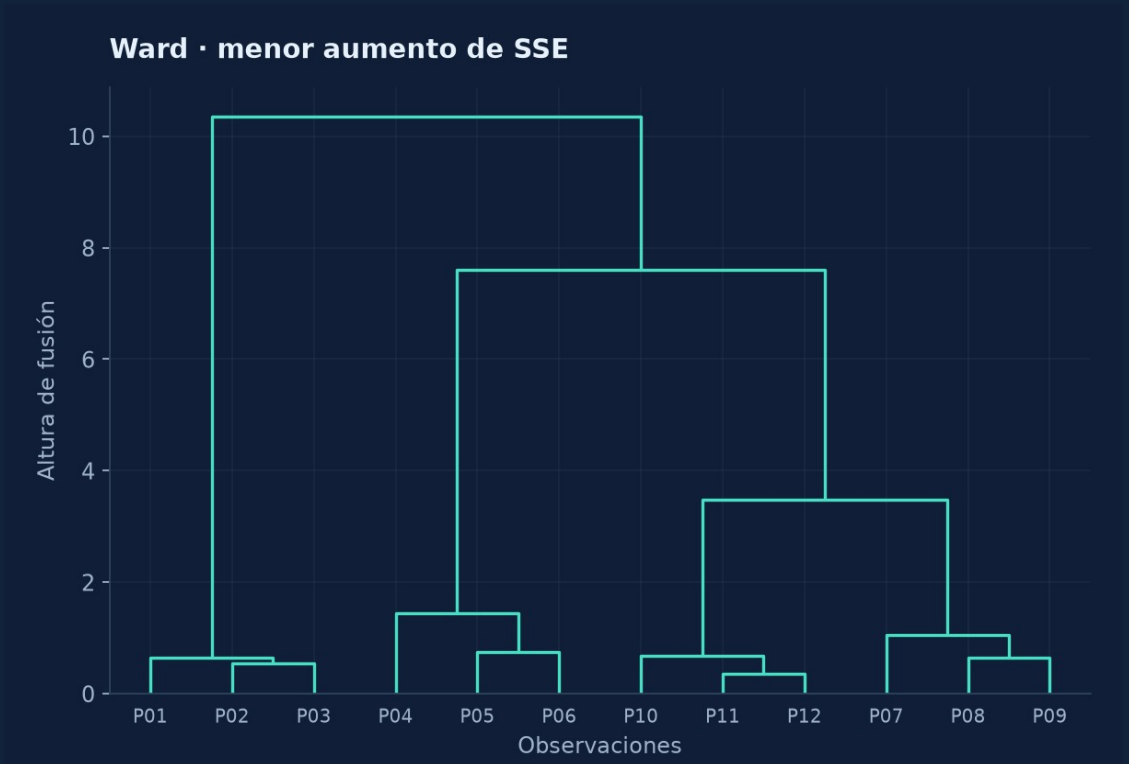
La información crucial está en la altura de fusión y en la estructura, no en el espacio horizontal.



Qué representa cada parte del dendrograma

IDEAS CLAVE

- Las hojas corresponden a observaciones originales.
- Las ramas indican combinaciones sucesivas de grupos.
- La altura codifica el valor del criterio de disimilitud al fusionar.
- Reordenar hojas puede cambiar el aspecto sin cambiar la jerarquía.



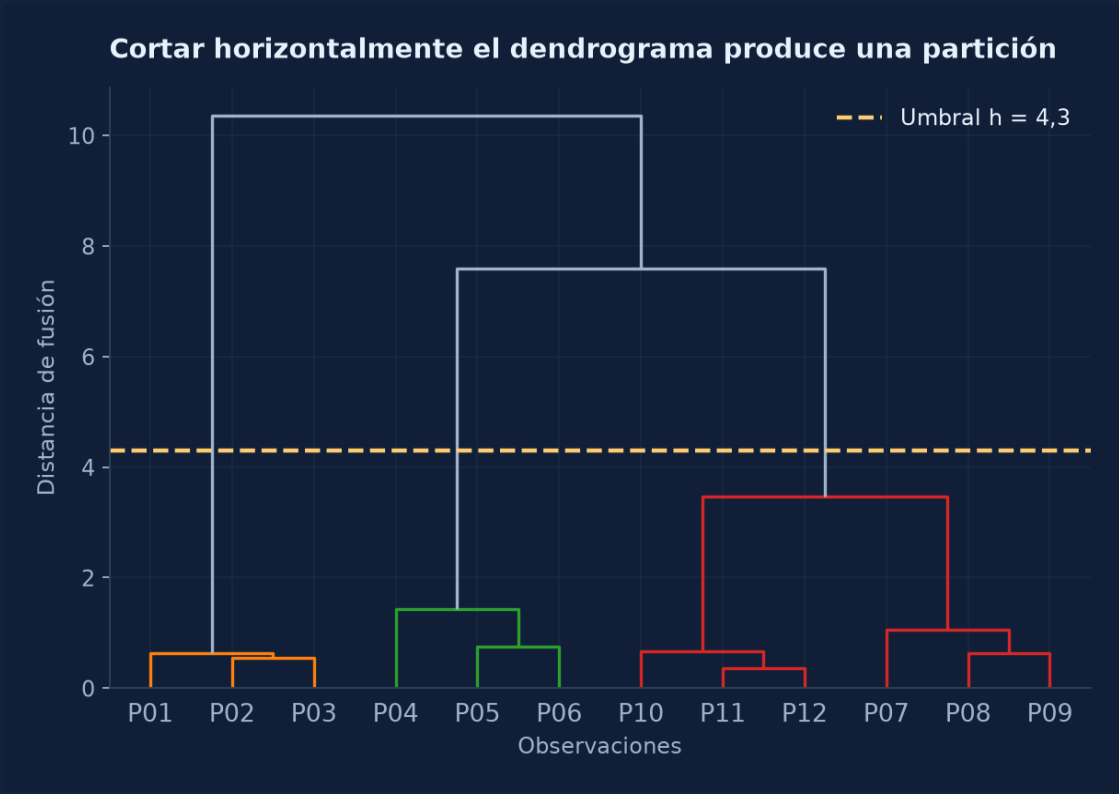
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **La posición horizontal no representa automáticamente una distancia medible.**

Cortar un árbol: del continuo jerárquico al K elegido

IDEAS CLAVE

- Un umbral h divide el árbol en componentes desconectadas.
- Alternativamente, solicitar directamente un número de clusters.
- Comparar niveles según estabilidad, tamaños y acción operativa.
- El mismo árbol admite segmentaciones gruesas y finas.

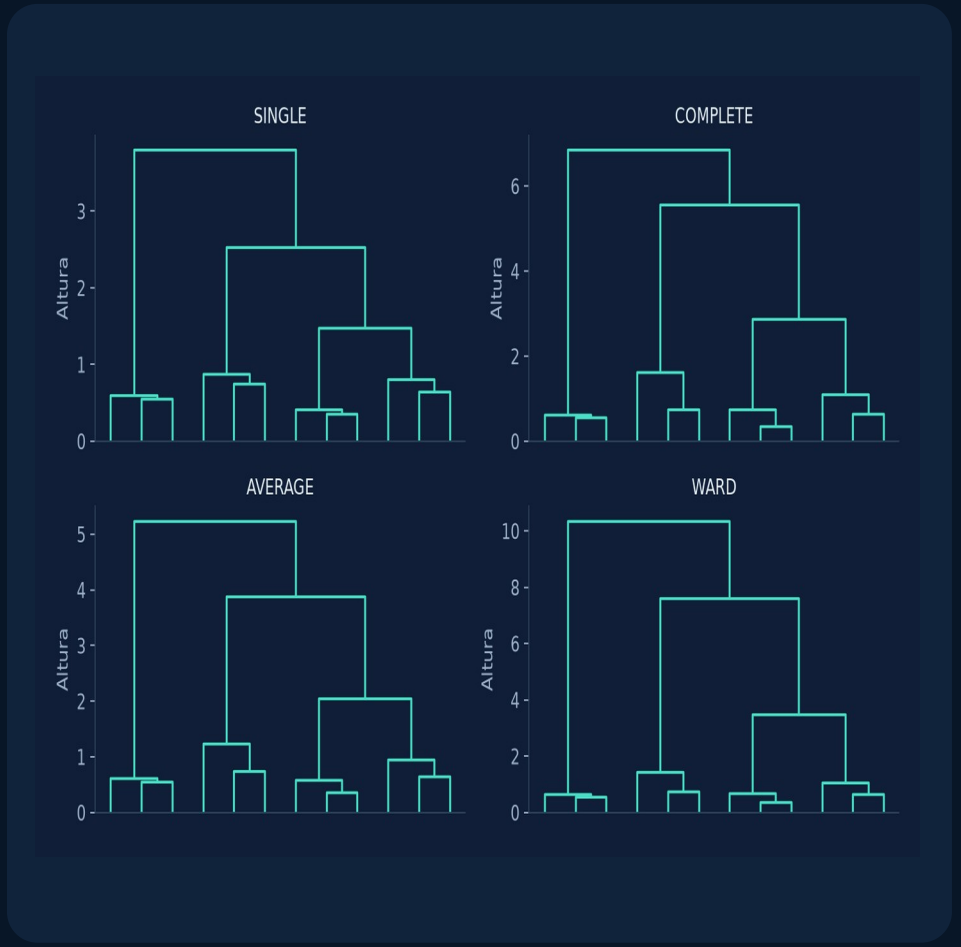


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El corte es una decisión analítica separada de la construcción del árbol.**

03 · Cuatro criterios de linkage

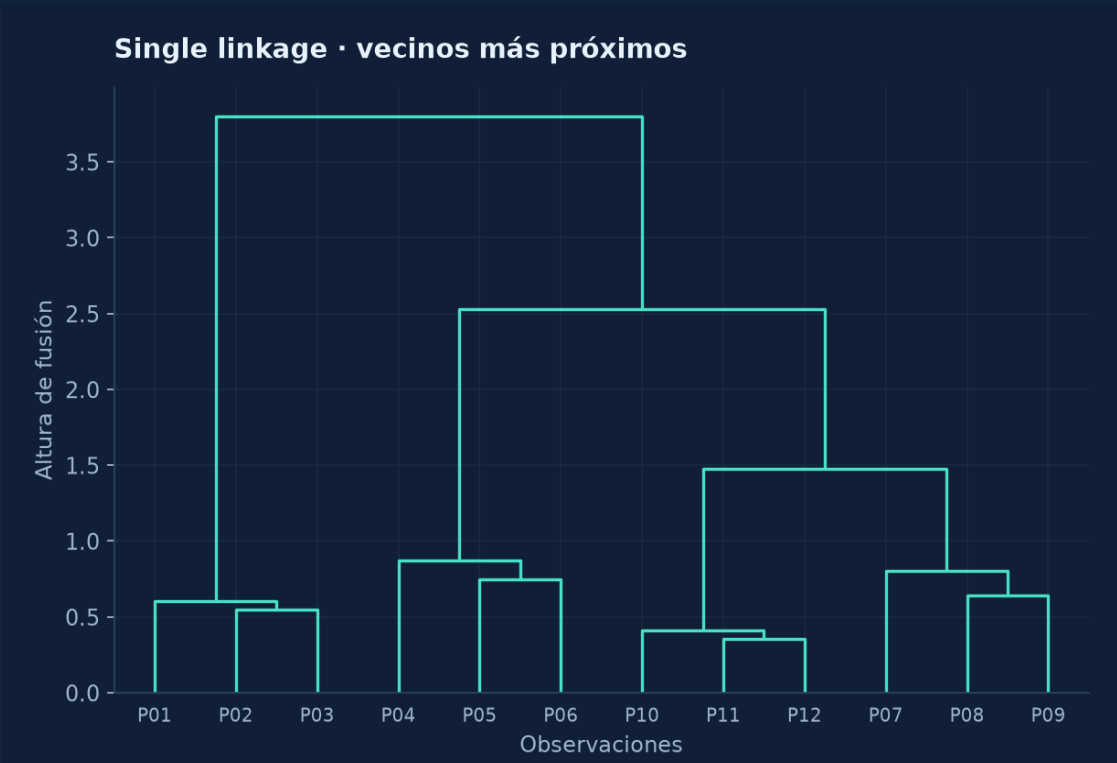
Single, complete, average y Ward producen estructuras diferentes aun con los mismos puntos.



Single linkage: distancia mínima entre clusters

IDEAS CLAVE

- D(A,B) = min d(a,b) para a∈A y b∈B.
- Puede capturar formas alargadas y conectividad local.
- Sufre el efecto chaining: un puente de puntos conecta grupos.
- Es sensible a contactos accidentales entre regiones.



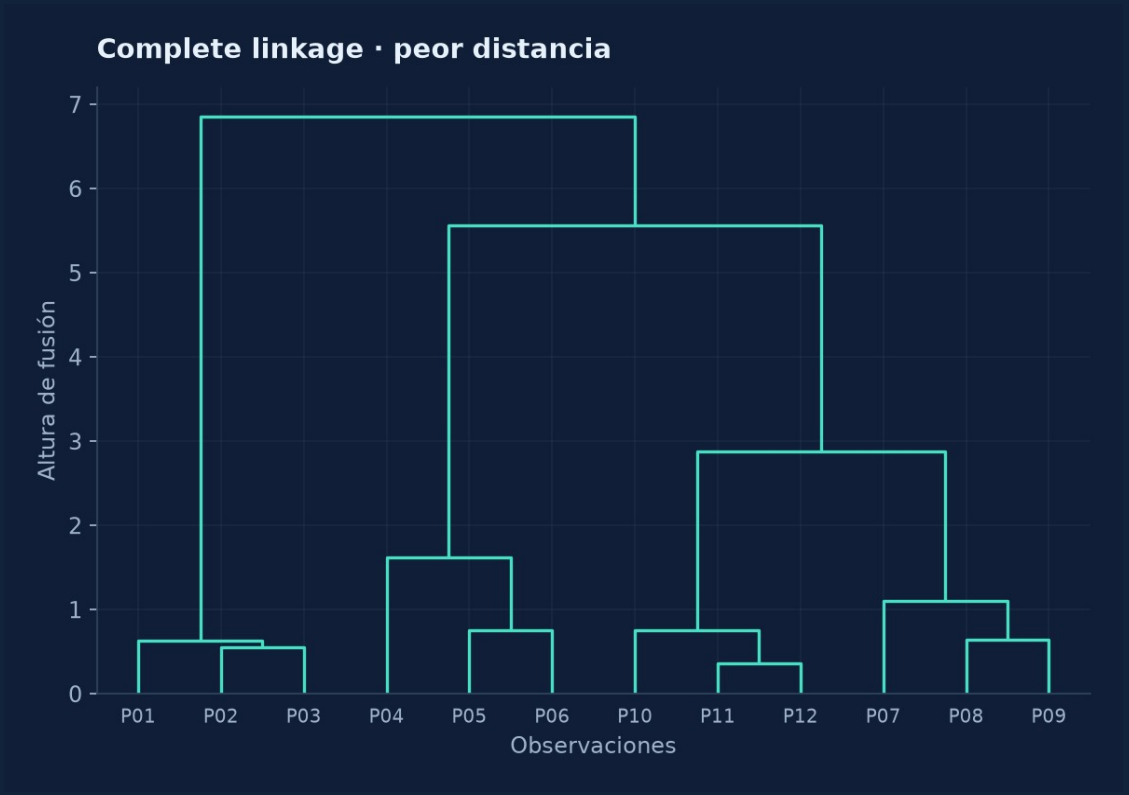
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Single prioriza conectividad; no necesariamente compacidad.**

Complete linkage: la distancia máxima

IDEAS CLAVE

- D(A,B) = max d(a,b) para a∈A y b∈B.
- Favorece grupos de diámetro contenido y relativamente compactos.
- Puede dividir estructuras largas que single sí conectaría.
- La elección depende de la topología esperada y el uso.



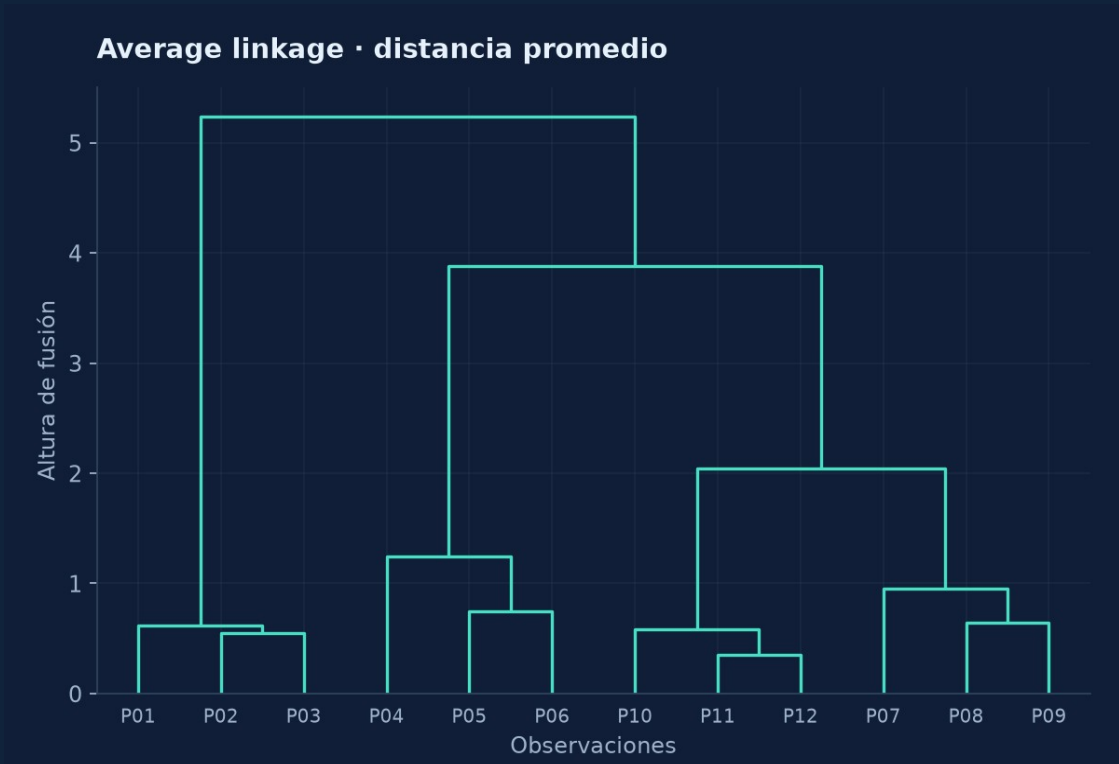
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Complete minimiza la peor separación entre puntos de grupos unidos.**

Average linkage: promedio de distancias por pares

IDEAS CLAVE

- D(A,B) = promedio de d(a,b) sobre a∈A, b∈B.
- Se ubica entre el énfasis conectivo de single y el compacto de complete.
- No equivale a distancia entre centroides en general.
- Es sensible a la métrica y al peso de grupos grandes.



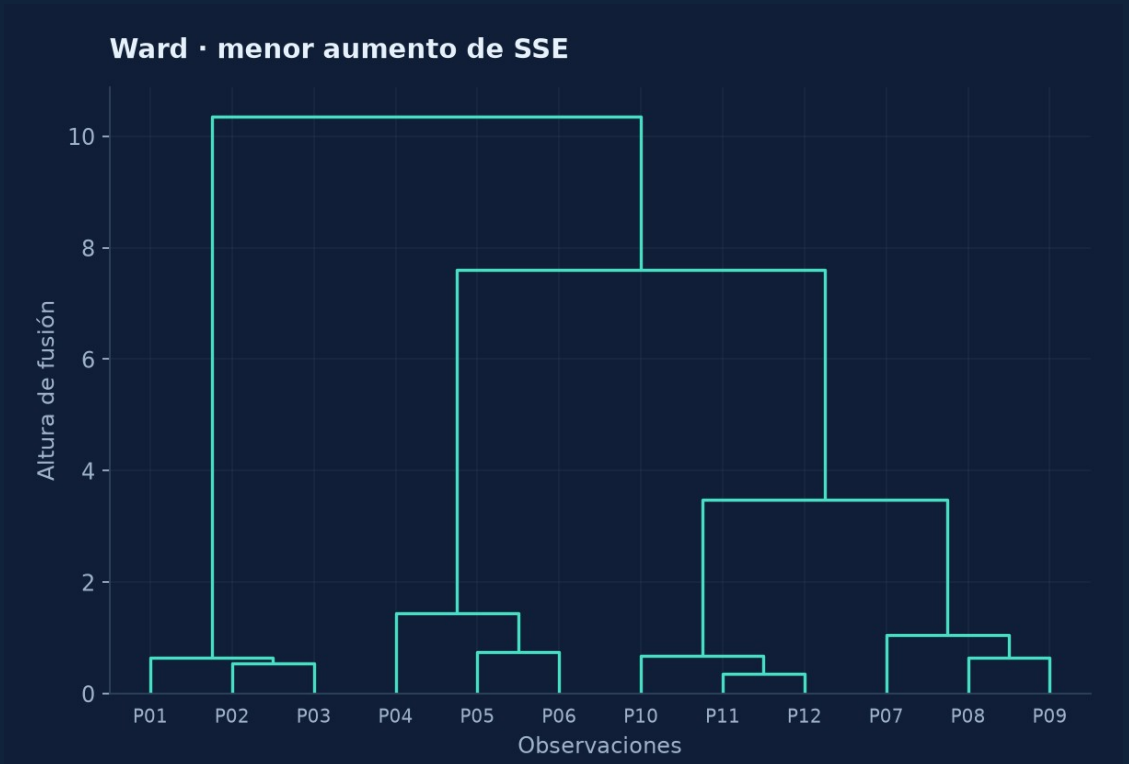
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Average balancea múltiples relaciones en vez de mirar solo extremos.**

Ward: fusionar donde menos aumenta la inercia

IDEAS CLAVE

- Selecciona la fusión que menos eleva la SSE total dentro de clusters.
- Para grupos A y B: ΔSSE = (nA nB)/(nA+nB) ‖μA−μB‖².
- Tiende a generar conjuntos compactos y de tamaño relativamente equilibrado.
- Requiere trabajar coherentemente con distancia euclídea en su formulación.



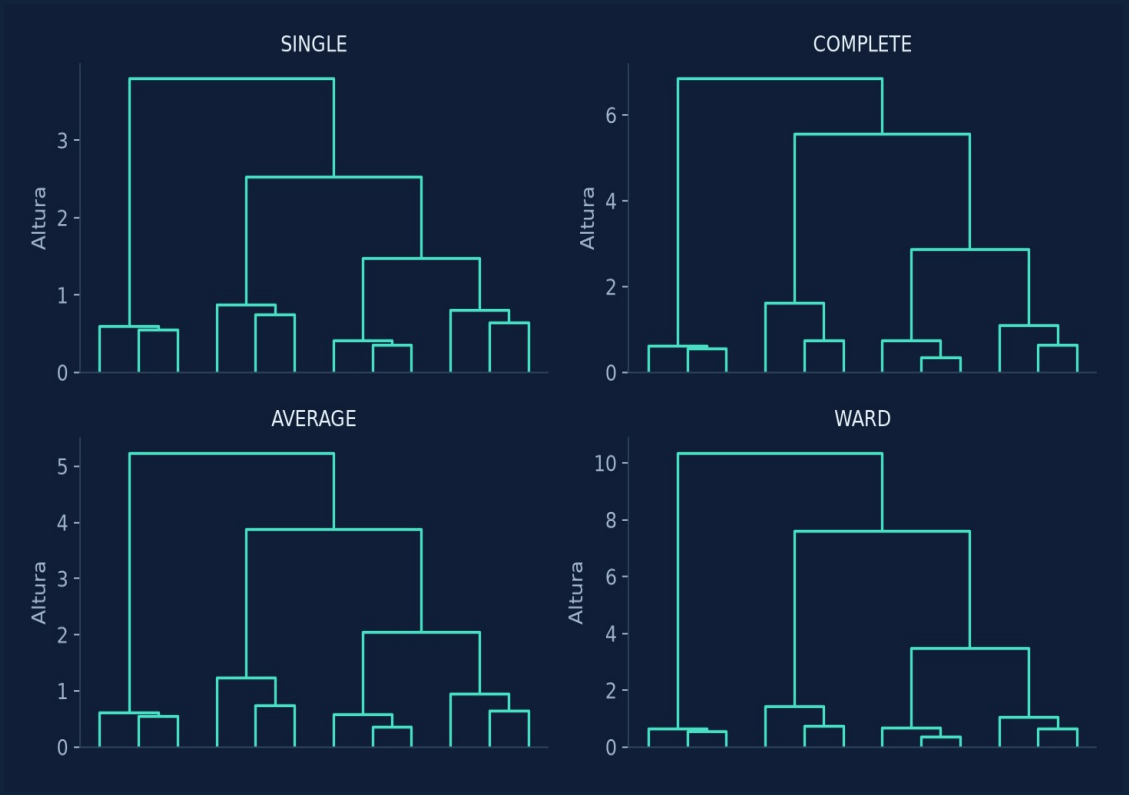
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **La altura graficada por software Ward depende de su convención; no siempre es ΔSSE literal.**

Comparar los cuatro árboles sobre la misma evidencia

IDEAS CLAVE

- Mantener puntos y escala exactamente iguales.
- Cambiar exclusivamente el criterio de enlace.
- Observar altas fusiones, grupos aislados y puentes.
- Evaluar más de una partición: la comparación visual no basta.



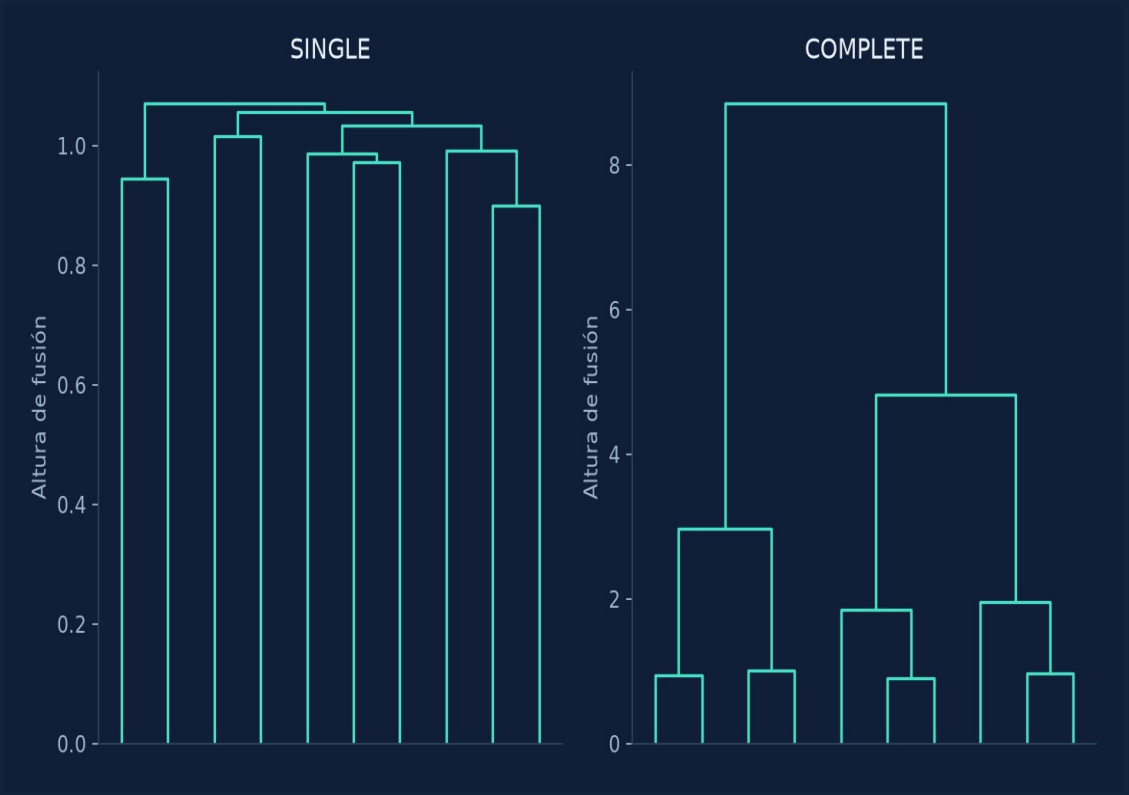
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El cambio de árbol no surge de nuevos datos sino de la definición de proximidad.**

Chaining: cuando un puente “pega” estructuras diferentes

IDEAS CLAVE

- Single fusiona observaciones a través de vecinos cercanos sucesivos.
- Una cadena de puntos puede conectar extremos muy separados.
- Complete penaliza el diámetro y ofrece un comportamiento diferente.
- La “mejor” solución depende de la noción de grupo que se quiera detectar.

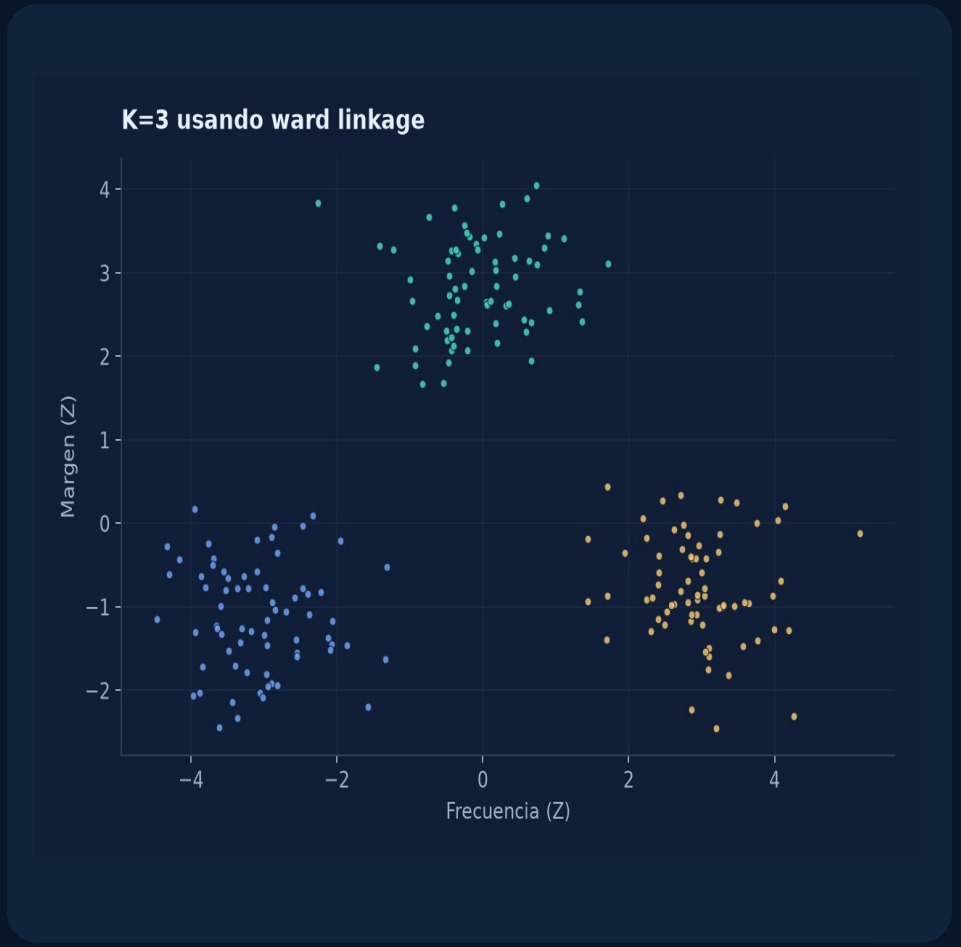


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El chaining no es un bug: es una consecuencia de la definición de single.**

04 · Validez y límites

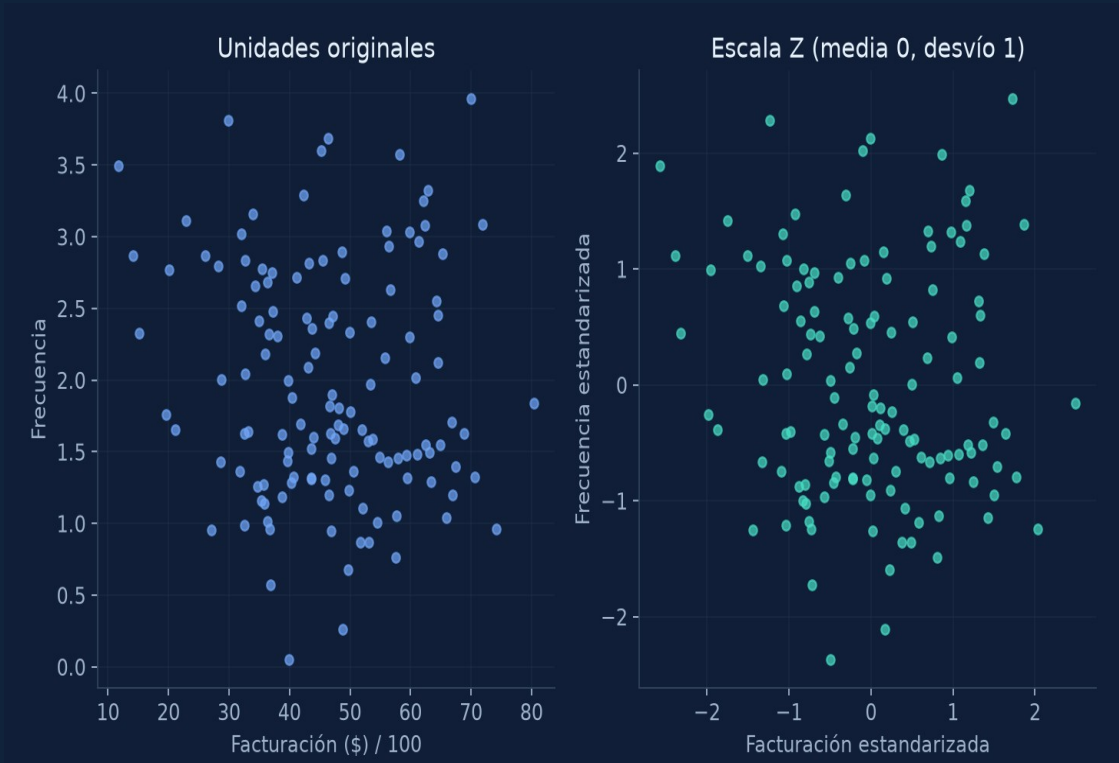
Los algoritmos construyen árboles completos aun cuando los datos no contienen agrupaciones significativas.



Las variables y la escala siguen gobernando el resultado

IDEAS CLAVE

- Estandarizar si las unidades y dispersiones no son comparables.
- Las transformaciones pueden alterar distancias y orden de fusión.
- En Ward, la geometría euclídea es parte del criterio cuadrático.
- Justificar la inclusión de variables relevantes para la acción.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **La jerarquía no corrige datos mal definidos o mal escalados.**

Distancia y linkage: combinaciones compatibles

IDEAS CLAVE

- Single/complete/average pueden usar diferentes métricas de distancia.
- Ward está atado a un objetivo de suma de cuadrados euclídeos.
- La distancia de Manhattan puede ser útil, pero cambia interpretación.
- Nunca reportar “Ward + cualquier métrica” sin verificar implementación.

COMPARACIÓN

SINGLE	Mínimo par / diversas métricas.
COMPLETE	Máximo par / diversas métricas.
AVERAGE	Promedio / diversas métricas.
WARD	Incremento SSE / marco euclídeo.

IDEA QUE DEBE QUEDAR **La consistencia matemática entre métrica y criterio es una exigencia, no un detalle.**

Coeficiente cofenético: cómo deforma las distancias

IDEAS CLAVE

- La distancia cofenética de dos puntos es la altura a la que se unen.
- Se correlacionan distancias originales con cofenéticas.
- Correlación alta significa fidelidad global del árbol a esas distancias.
- No garantiza separación real ni utilidad de clusters para decisiones.

MODELO MATEMÁTICO

$$r_coph = corr(d_{ij}, c_{ij})$$

- d_{ij}: distancia original entre i y j.
- c_{ij}: altura de fusión en el árbol.
- Comparar linkages manteniendo misma matriz.
- No confundir fidelidad geométrica con utilidad.

IDEA QUE DEBE QUEDAR **Es una medida de representación de distancias, no un “certificado de clusters”.**

Costo computacional: por qué no escalar ciegamente

IDEAS CLAVE

- La matriz de distancias completa contiene $n(n-1)/2$ pares.
- Requiere $O(n^2)$ memoria para distancias densas.
- El tiempo depende de algoritmo, linkage e implementación.
- Para cientos de miles de clientes conviene submuestrear o usar alternativas.

REGLAS DE DECISIÓN

n^2

Número de distancias crece cuadráticamente

10k

Casi 50 M de pares

RAM

Puede ser el cuello de botella

MUESTRA

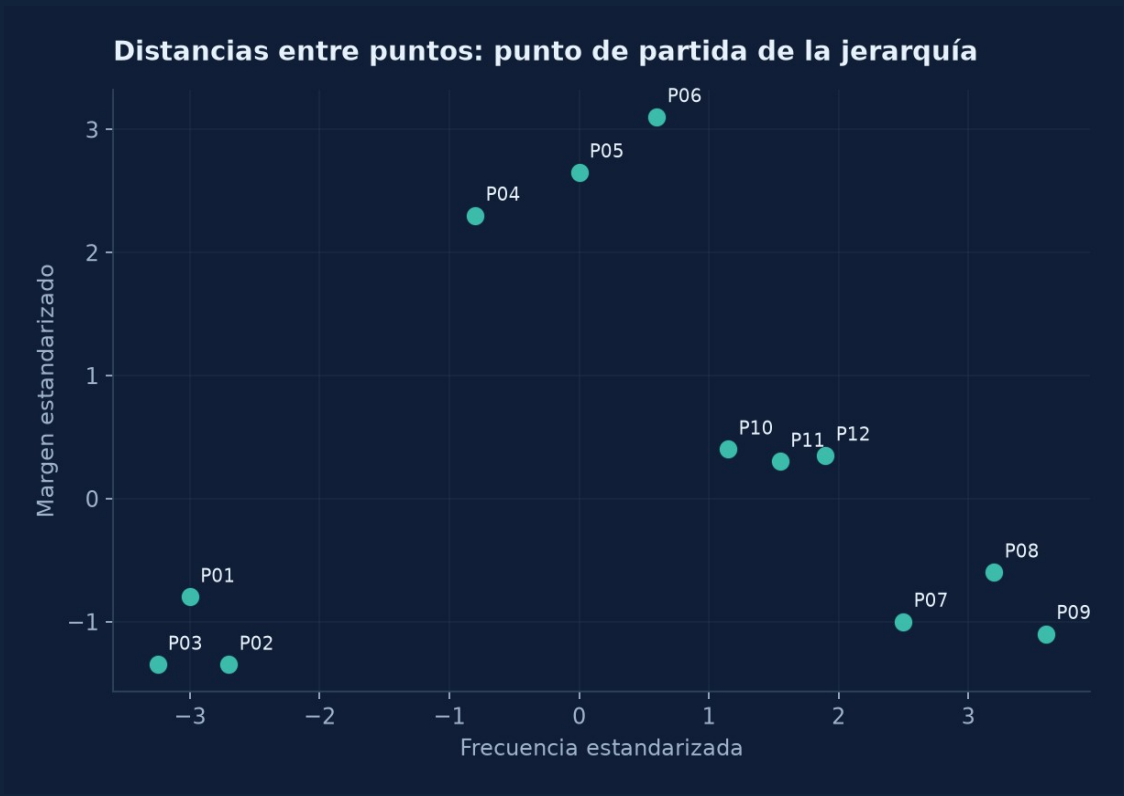
Evaluar escalabilidad antes de desplegar

IDEA QUE DEBE QUEDAR **Un dendrograma completo no siempre es la herramienta operacional escalable.**

Ruido, outliers y estabilidad del árbol

IDEAS CLAVE

- Una observación atípica puede unirse al final y alterar alturas.
- Pequeñas perturbaciones de datos pueden modificar ramas cercanas.
- Comparar cortes en bootstrap, ventanas temporales o submuestras.
- Medir concordancia de particiones sin comparar IDs arbitrarios.



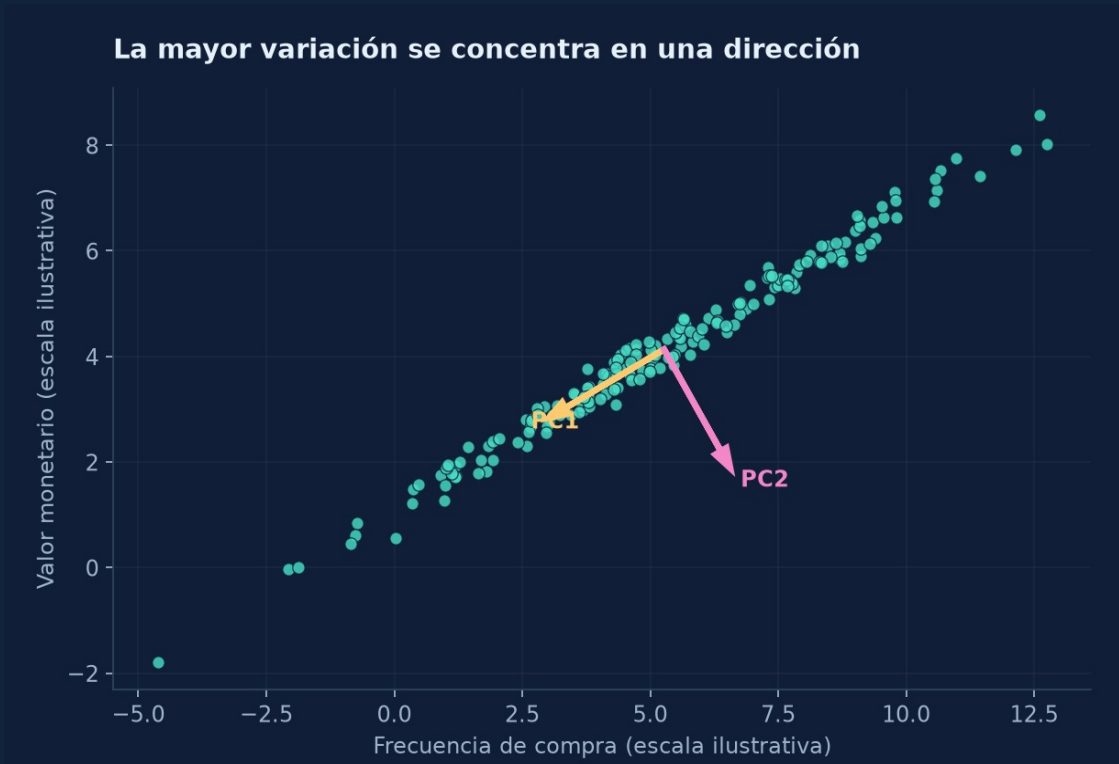
ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **Priorizar las divisiones reproducibles por sobre historias atractivas del árbol.**

¿PCA antes del clustering jerárquico?

IDEAS CLAVE

- PCA puede atenuar redundancia y facilitar visualización de clusters.
- Descartar componentes cambia distancias y, por lo tanto, fusiones.
- Si se usan todas las componentes de PCA centrado, las distancias euclídeas se preservan.
- Si se estandarizan scores o se truncan, se modifica la geometría.

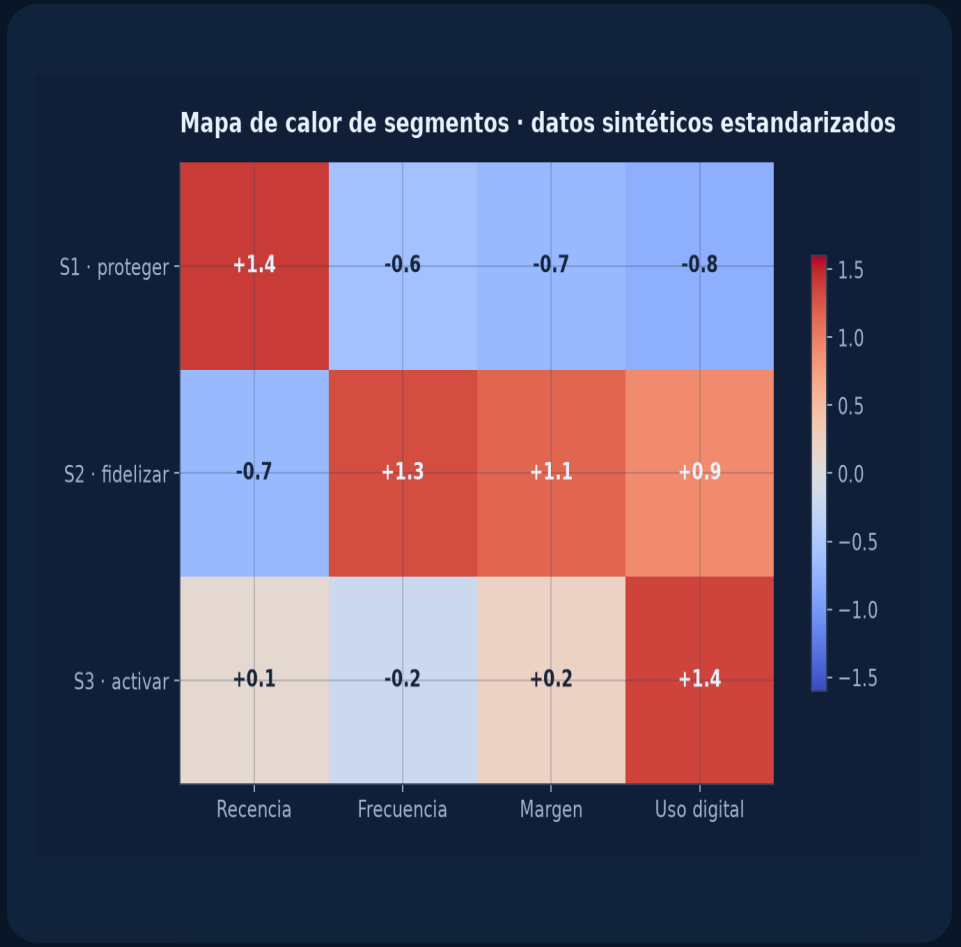


ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **La rotación ortogonal completa preserva distancias; la reducción no.**

05 · Interpretación y aplicación

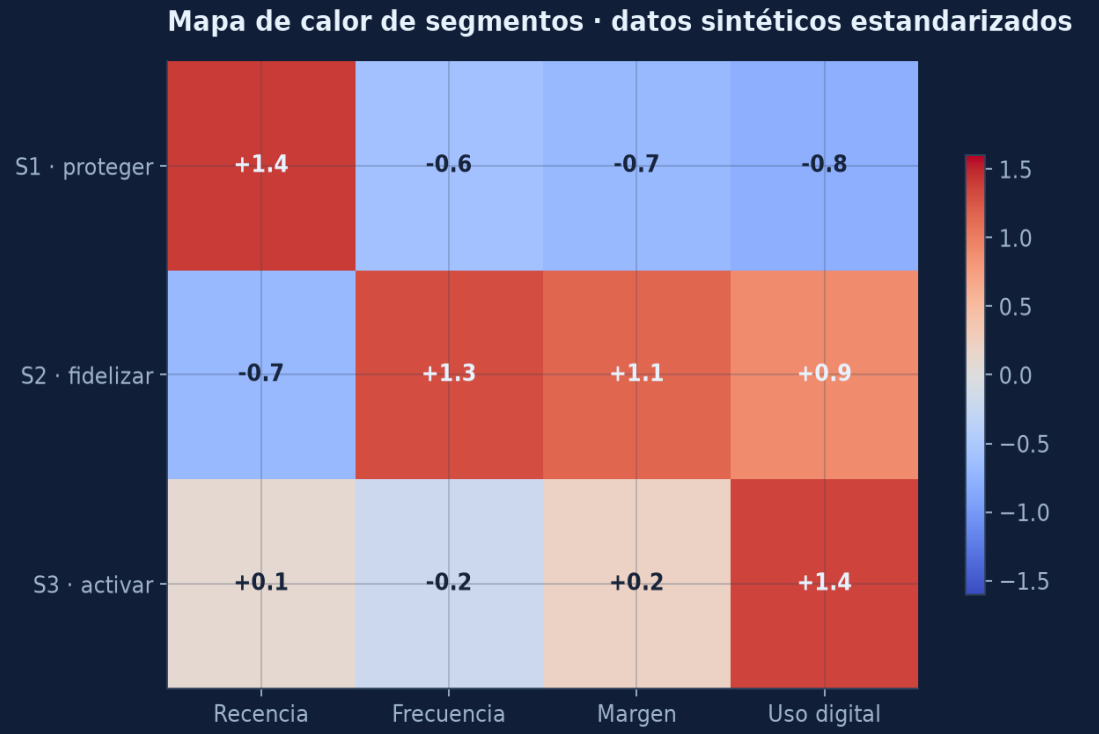
Los niveles del árbol pueden corresponder a decisiones de distinta granularidad si se validan.



Heatmap de perfiles: ver si los segmentos son accionables

IDEAS CLAVE

- Después del corte, calcular medias o medianas en variables originales.
- En el ejemplo sintético, los grupos difieren en recencia, margen y canal.
- Contrastar cada perfil con dispersión, tamaño y estabilidad.
- No tratar la visualización del mapa de calor como evidencia causal.



ESQUEMA / DATOS SINTÉTICOS

IDEA QUE DEBE QUEDAR **El mapa de calor transforma una partición en hipótesis comerciales contrastables.**

Jerárquico versus K-Means: dos preguntas diferentes

IDEAS CLAVE

- K-Means: partición de K grupos que minimiza inercia localmente.
- Jerárquico: secuencia de fusiones sujeta a distancia y linkage.
- Ward comparte con K-Means énfasis en SSE, pero no optimizan idénticamente.
- La elección depende de escalabilidad, forma, estabilidad y posibilidad de acción.

COMPARACIÓN

K-MEANS	K inicial, centroides, reasignación.
JERÁRQUICO	Árbol y cortes con niveles múltiples.
SIMILITUD	Ambos dependen de escala y variables.
PRUEBA	Perfiles + estabilidad + utilidad externa.

IDEA QUE DEBE QUEDAR **El dendrograma aporta lectura multiescala, no superioridad universal.**

Cuatro criterios para adoptar una segmentación

IDEAS CLAVE

- Estructura: separación, cohesión y alturas de fusión.
- Reproducibilidad: estabilidad frente a remuestreo y perturbaciones.
- Interpretabilidad: variables originales distinguen grupos.
- Viabilidad: costo de asignación, tamaño mínimo y tratamiento de clientes nuevos.

REGLAS DE DECISIÓN

01

Estructura geométrica

02

Estabilidad de grupos

03

Perfiles interpretables

04

Valor y operación

IDEA QUE DEBE QUEDAR **Sin valor diferencial verificable, una jerarquía compleja puede sobrar.**

Simulaciones para reproducir cada transformación

IDEAS CLAVE

- S10: construir matriz de distancias a partir de puntos movibles.
- S11: animar fusiones y actualizar dendrograma paso a paso.
- S12: comparar single, complete, average y Ward.
- S13: mover el umbral de corte y observar clusters resultantes.
- S14: estudiar chaining, sensibilidad y perfiles de negocio.

SECUENCIA DE ANÁLISIS

S10 · Distancias interactivas

↓

S11 · Fusiones una a una

↓

S12 · Linkages lado a lado

↓

S13 · Corte y partición

↓

S14 · Puentes / estabilidad / negocio

IDEA QUE DEBE QUEDAR **Cada control tiene que conectar cálculo, gráfico e interpretación.**

Evaluación integradora: el árbol no decide por nosotros

IDEAS CLAVE

- Single muestra dos grupos unidos por un puente; Ward muestra tres compactos.
- Cortar Ward en K=4 crea un grupo de 3 clientes con alto margen.
- En una submuestra desaparece ese microgrupo.
- Diseñar validaciones antes de asignar presupuesto comercial.

PREGUNTAS DE DECISIÓN

1. ¿Cuál es el supuesto de cada linkage?
2. ¿Qué evidencia favorece cierto corte?
3. ¿Cuánto pesan tres clientes en el negocio?
4. ¿Qué experimento validarías primero?

IDEA QUE DEBE QUEDAR **El método ofrece alternativas; el analista construye la decisión defendible.**

Errores de lectura y fuentes esenciales

IDEAS CLAVE

- Una hoja próxima en el gráfico no implica una distancia baja.
- La altura de Ward no debe leerse siempre como distancia euclídea literal.
- Murtagh & Contreras (2012). Algorithms for hierarchical clustering.
- Hastie, Tibshirani & Friedman (2009); SciPy hierarchy y scikit-learn Clustering.

SECUENCIA DE ANÁLISIS

1. Definir métrica + linkage
2. Construir y leer el árbol
3. Elegir y justificar el corte
4. Validar estabilidad + perfiles
5. **Comprobar impacto del negocio**

IDEA QUE DEBE QUEDAR **Las ramas sirven para investigar hipótesis, no para sustituir validación.**