Clustering

Diego Garrido

Departamento de Ingeniería Industrial
Universidad de Chile

13 de Septiembre de 2019

Es una tarea de aprendizaje no supervisado, aquí solo existen *inputs* $D = \{x_i\}_{i=1}^{N}$, $x_i \in \mathbb{R}^M$.

Definición:Consiste en agrupar los datos *clusters*, es decir, en una colección de objetos con una alta similitud entre sí y una alta disimilitud con objetos de otros grupos.
Características relevantes:

- **Similitud**: Métrica con la que se mide el grado de semejanza ue possen los objetos entre sí.**¿Cómo medimos la similitud?**
- **Centroide**: Punto promedio de un cluster.
- Calidad del cluster:
  - Enfoque experto: se define a priori que es un buen cluster dada las restricciones del negocio.
  - Métricas de similitud internas del cluster (no supervisado) o externas (supervisado).
- Usos: Segmentación de mercado, perfilamiento de delitos, organización de contenido web.

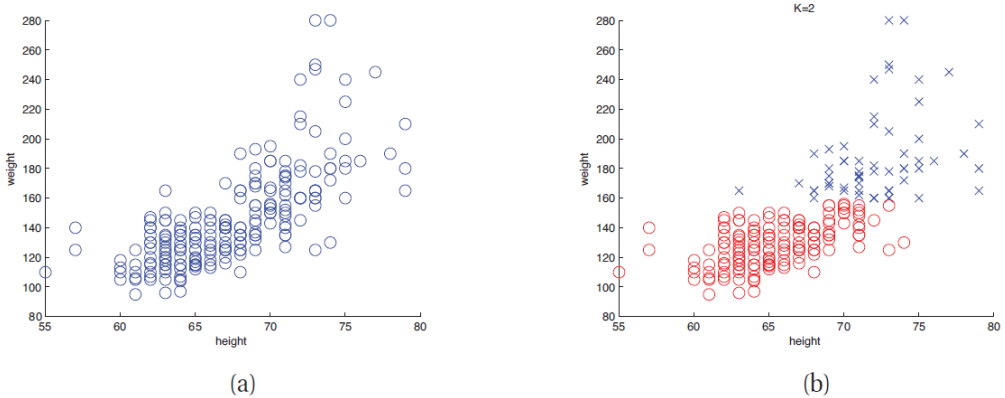


Fig. 1: (a) La altura y peso de algunas personas. (b) Una posible segmentación usando Kmeans con K=2 *clusters*.

Diremos que dos objetos tienen una alta similitud si hay poca distancia entre ellos. El objetivo es poder comparar objetos del estilo $x = (x_1, ..., xM)$ e $y = (y_1, ..., yM)$.

Algunas medidas de distancia (disimilitud) y similitud entre vectores son las siguientes:

$$d_{Minkowski(x,y)} = (\sum_{i=1}^{M} |x_i - y_i|^p)^{\frac{1}{p}}, \; d_{Euclidean(x,y)} = \sqrt{\sum_{i=1}^{M} |x_i - y_i|^2}$$

$$d_{Manhattan(x,y)} = \sum_{i=1}^{M} |x_i - y_i|, \; s_{Cosine(x,y)} = \frac{x^T y}{\sqrt{\sum_{i=1}^{M} x_i^2}\sqrt{\sum_{i=1}^{M} y_i^2}}$$

Métricas de calidad

Algunas métricas de calidad de un análisis de clustering:

- **Within cluster Sum of Squares (WSS)**: mide que tan cohesionados son los clusters, mientras menor mayor cohesión o más compactos son los clusters, también es conocida como *Inertia*.

$$WSS = \sum_{k=1}^{K} \sum_{x \in C_k} (x - \bar{x_k})^2 \tag{1}$$

- **Between cluster Sum of Squares (BSS)**: mide que tan separados está un cluster de otro cluster, mientras mayor más separados o disimiles son los clusters:

$$BSS = \sum_{k=1}^{K} |C_k| (\bar{x_k} - \bar{X})^2 \tag{2}$$

Mezcla de cosas
Nota: la distancia euclideana puede ser reemplazada por otra distancia en la formula.

Input K: número de clusters a encontrar.

1. Inicialización: escoger K puntos al azar en el plano, estos serán los centroides iniciales.
2. Asignar cada observación al centroide más cercano.
3. Recalcular los centroides como el promedio de las observaciones pertenecientes a el.
4. Repetir puntos 2 y 3 hasta que:
   - Hasta que el cambio de la *Inertia* de una iteración a otra este por debajo de un *threshold*.
   - Se supera el limite máximo de iteraciones ingresado por el usuario.

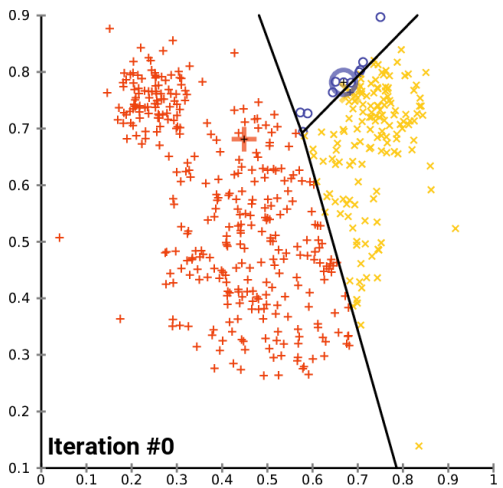


Fig. 2: Ejemplo[1] de convergencia del algoritmo K-means para K=3.

[1] https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

1. Preprocesamiento
   1. Escalar los datos, así todos los atributos poseen la misma posibilidad de influir en los clusters.
   2. Eliminar *outliers* si no son relevantes para el problema.
   3. Realizar selección de atributos (de lo contrario se puede afectar seriamente los resultados).
2. Postprocesamiento
   1. Filtrar cluster pequeños que pueden representar outliers que no son interés del problema.
   2. Dividir clusters con poca cohesión interna.
   3. Unir clusters cercanos con alta cohesión interna.
3. Cómo escoger el número óptimo de clusters:
   1. Regla del codo sobre una métrica monótona.
   2. Métricas que penalizan la cantidad de parámetros.
   3. Conocimiento del negocio.

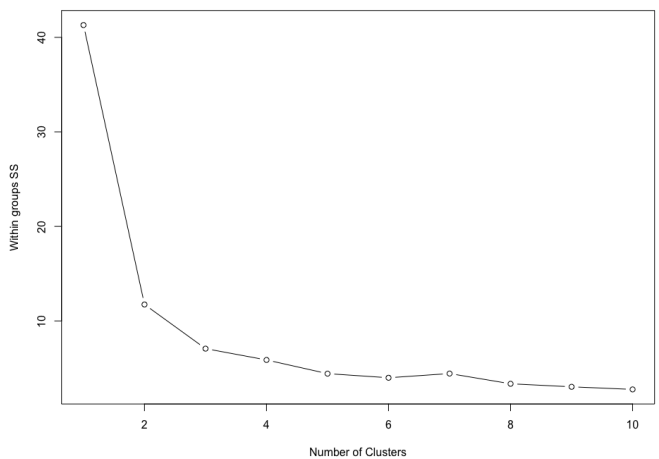


Fig. 3: Criterio del codo sobre la métrica *Inertia*, el punto de máxima inflexión es con dos clusters.

Algunas limitaciones de K-means:

- Clusters con diferentes densidades.
- Clusters con formas no esféricas.
- Clusters con diferentes tamaños.
- No es robusto a outliers.
- Mínimos locales (sensible a la inicialización).

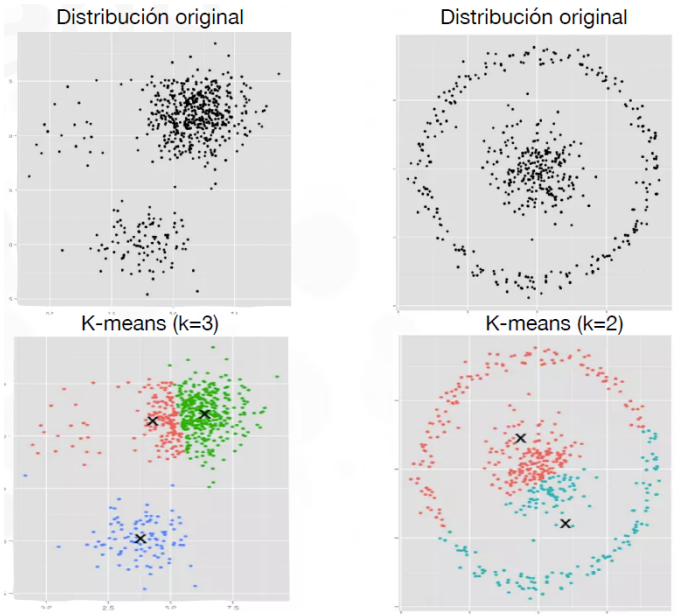


Fig. 4: Ejemplo de limitaciones del algoritmo K-means

Gaussian Mixture Model

Gaussian Mixture Model (GMM) es uno de los modelos de clustering más utilizados que tienen enfoque probabilístico. Asume que cada observación proviene de una mezcla de K gaussianas multivariadas. En el enfoque probabilístico las observaciones pertenecen a todos los clusters pero en diferentes proporciones, además se puede volver al enfoque clásico tomando el argumento máximo.

$$p(x_i|\theta) = \sum_{k=1}^{K} \pi_k \mathbb{N}(x_i|\mu_k, \Sigma_k)$$

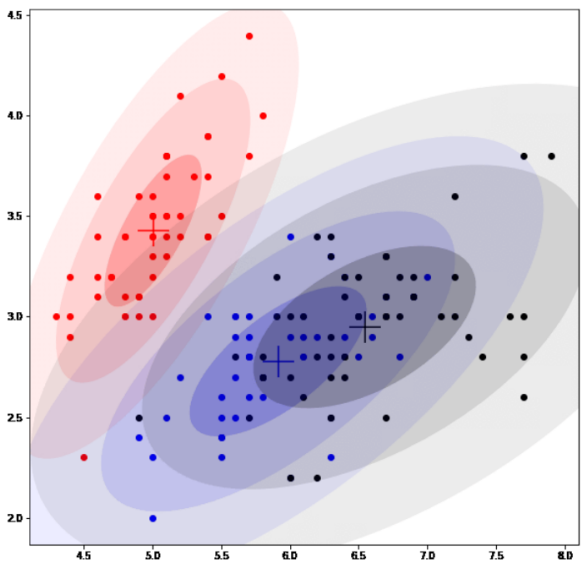


Fig. 5: Ejemplo[2] de GMM con K=3.

[2] https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

Inputs:

- *eps* =distancia máxima que puede estar una observación a otra para ser considerada como vecino de este.
- *min_samples* = tamaño mínimo que debe tener un cluster para ser considerado como tal, sino es considerado como ruido.

Algoritmo:

1. Selecciona una observación al azar.
2. Barrido: agrupar todos los vecinos que están en un radio menor o igual a *eps* de la observación seleccionada si la cantidad de observaciones agrupadas es mayor a *min_samples* se inicia un cluster sobre el mismo, sino el punto es etiquetado como ruido.
3. Si un punto pertenece a un cluster su vencidad entorno a *eps* también se añade si es lo suficientemente densa.
4. Si no hay más puntos vecinos se crea el cluster con todas las observaciones agrupadas.
5. Repetir puntos 1, 2, 3 y 4 hasta que no queden observaciones sin etiquetar.

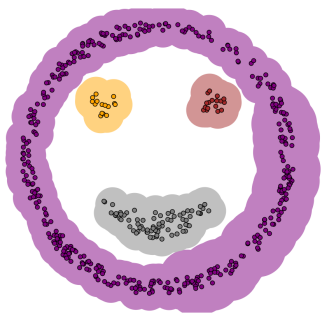


Fig. 6: Ejemplo[3] DBSCAN donde se hallaron 4 clusters.

Clustering Jerárquico

Produce un conjunto de clusters anidados y organizados en un árbol jerárquico.

- Típicamente se visualizan como un dendograma.
- No se tiene que suponer un número a priori de clusters (se puede obtener el número deseado de clusters cortando el dendograma al nivel deseado).

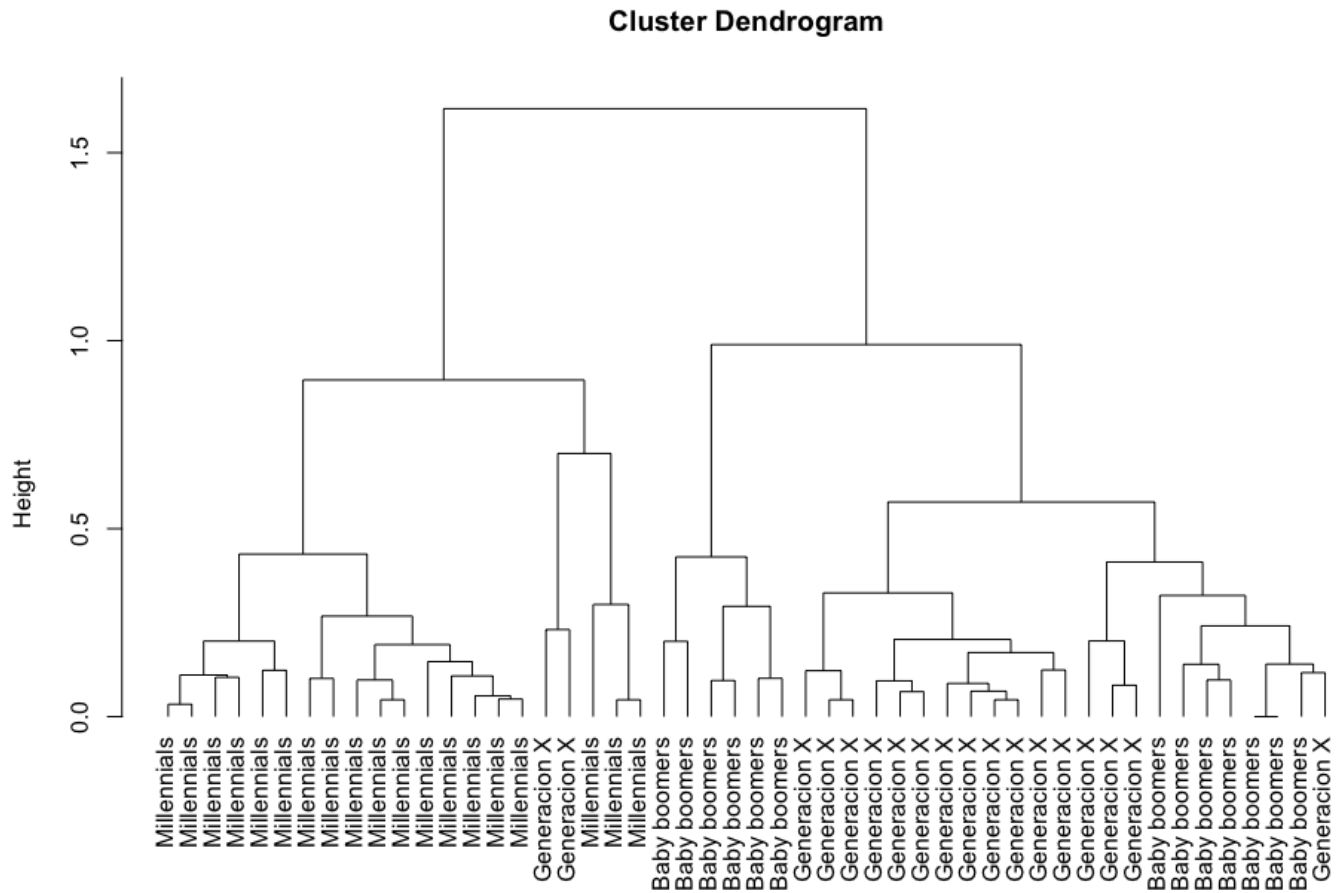


Fig. 7: Eje x observaciones eje y distancia, si el corte es más arriba menos clusters.

1. Clustering Jerárquico Aglomerativo (Bottom Up)

   (a) Comenzar con cada elemento como un cluster inicial.
   (b) Calcular matriz de distancias entre clusters.
   (c) En cada paso, mezclar el par de clusters más cercano hasta que quede sólo un cluster. (o k clusters)

2. Clustering Jerárquico Divisivo (Top Down)

   (a) Empezar con un cluster que contenga todos los puntos.
   (b) Calcular matriz de distancias entre clusters.
   (c) En cada paso, dividir un cluster en dos hasta que todo cluster contenga un solo punto (o haya k clusters).

1. Single: se calcula la distancia entre todos los miembros del cluster y las observaciones en evaluación para ser incluidas en el cluster, seleccionando la menor distancia.
2. Complete: se calcula la distancia entre todos los miembros del cluster y las observaciones en evaluación para ser incluidas en el cluster, seleccionando la mayor distancia.
3. Ward.D: la similitud entre clusters se basa en el incremento del WSS cuando se mezclan dos clusters. Análogo jerárquico de K-means.

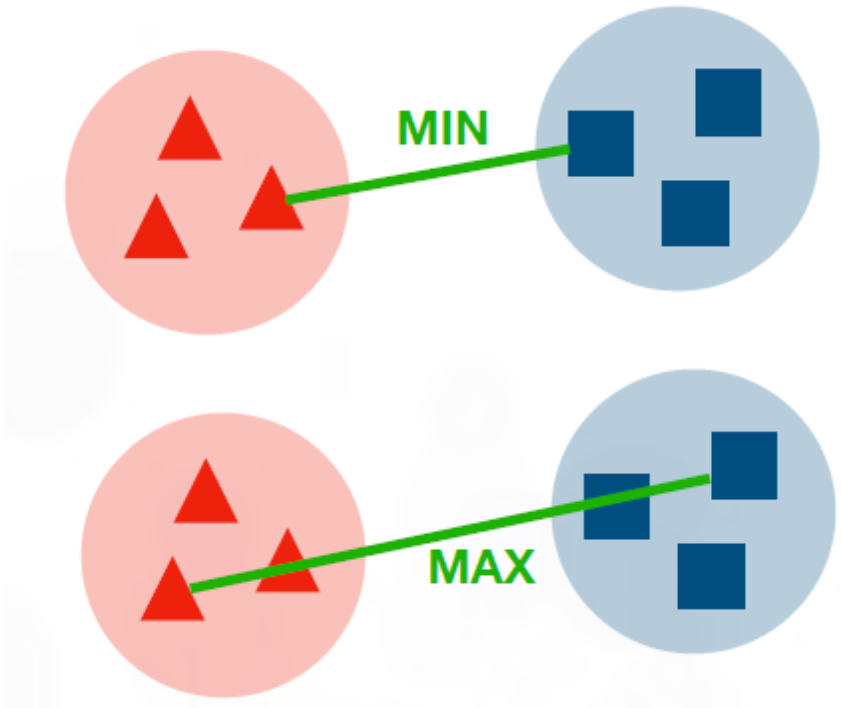
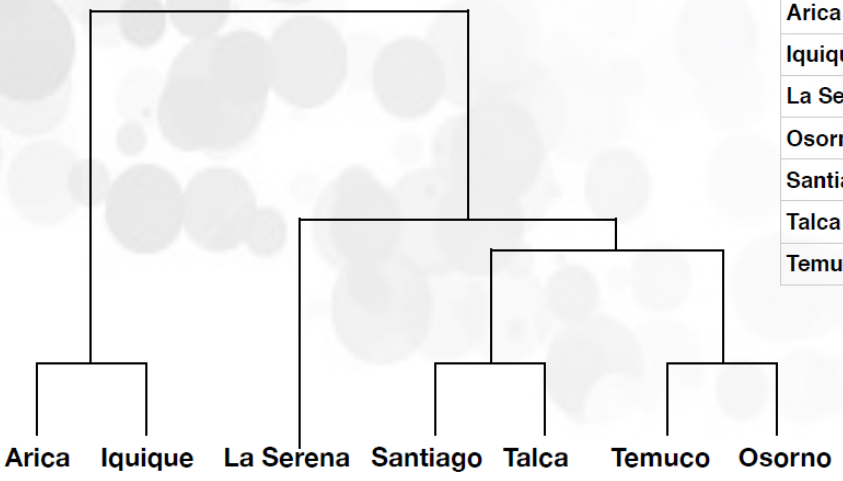
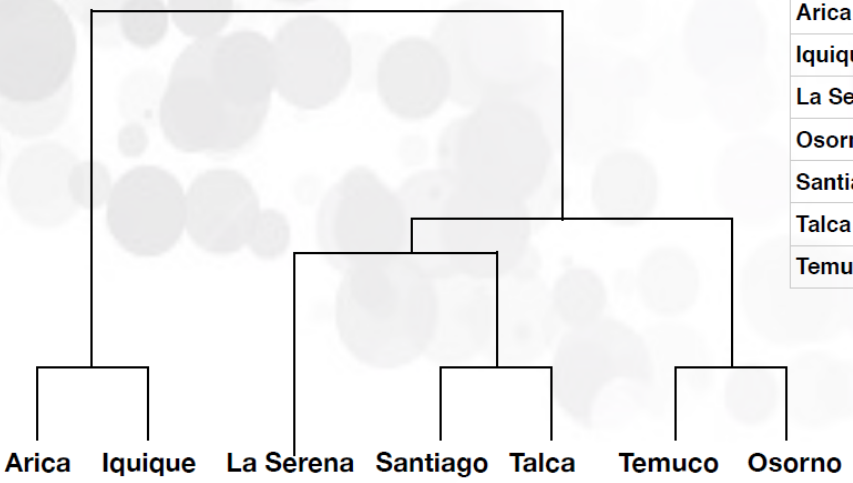


Fig. 8: Single (parte superior), complete (parte inferior).



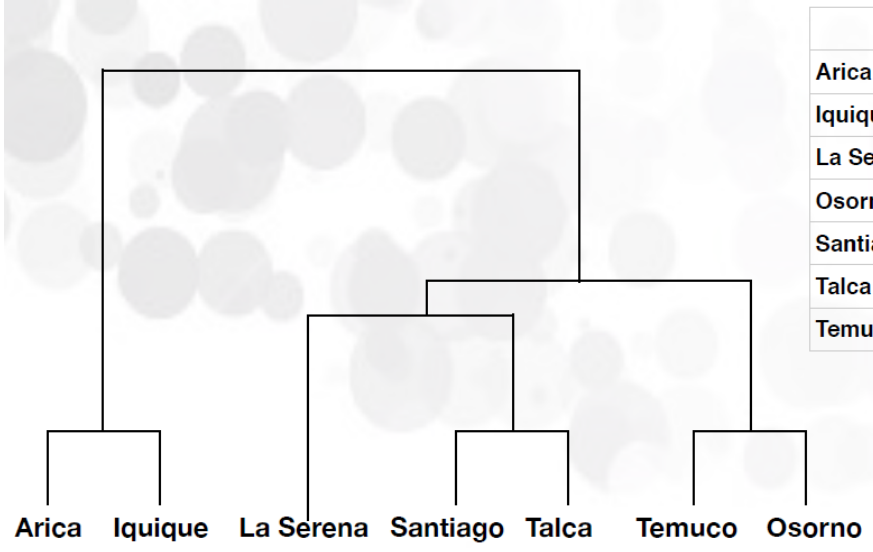
	Arica	Iquique	La Serena	Osorno	Santiago	Talca	Temuco
Arica	0	300	1580	2995	2050	2310	2725
Iquique	300	0	1370	2790	1840	2100	2515
La Serena	1580	1370	0	1415	470	725	1145
Osorno	2995	2790	1415	0	945	695	270
Santiago	2050	1840	470	945	0	255	675
Talca	2310	2100	725	695	255	0	420
Temuco	2725	2515	1145	270	675	420	0

Fig. 9: Dendograme usando Single(izquierda), matriz de distancia entre observaciones (derecha).



	Arica	Iquique	La Serena	Osorno	Santiago	Talca	Temuco
Arica	0	300	1580	2995	2050	2310	2725
Iquique	300	0	1370	2790	1840	2100	2515
La Serena	1580	1370	0	1415	470	725	1145
Osorno	2995	2790	1415	0	945	695	270
Santiago	2050	1840	470	945	0	255	675
Talca	2310	2100	725	695	255	0	420
Temuco	2725	2515	1145	270	675	420	0

Fig. 10: Dendograme usando Complete(izquierda), matriz de distancia entre observaciones (derecha).



	Arica	Iquique	La Serena	Osorno	Santiago	Talca	Temuco
Arica	0	300	1580	2995	2050	2310	2725
Iquique	300	0	1370	2790	1840	2100	2515
La Serena	1580	1370	0	1415	470	725	1145
Osorno	2995	2790	1415	0	945	695	270
Santiago	2050	1840	470	945	0	255	675
Talca	2310	2100	725	695	255	0	420
Temuco	2725	2515	1145	270	675	420	0

Fig. 11: Dendograme usando Ward.D(izquierda), matriz de distancia entre observaciones (derecha).