Universidad del País Vasco Euskal Herriko Unibertsitatea
INFORMATIKA FAKULTATEA
FACULTAD DE INFORMÁTICA
Konputazio Ingeniaritza eta Sistema Adimentsuak Unibertsitate Masterra
Máster Universitario en Ingeniería Computacional y Sistemas Inteligentes

INGENIERÍA COMPUTACIONAL Y SISTEMAS INTELIGENTES

EXPLORACIÓN Y ANÁLISIS DE DATOS

Entrega 3

Estudiante: Eneko Perez **Estudiante:** Alan García

Curso: 2024-2025 **Fecha:** 20 de octubre de 2024

Índice

1. Métodos

1.1. Clustering jerárquico

La agrupación jerárquica implica la creación de clústeres que tienen un orden predeterminado de arriba a abajo. Existen dos tipos de agrupación jerárquica, **divisivo** y **aglomerativo**. En el método de agrupamiento aglomerativo o de abajo hacia arriba, se asigna cada observación a su propio clúster. Luego, se calcula la similitud (por ejemplo, la distancia) entre cada uno de los clústeres y se unen los dos clústeres más similares en cada iteración. Por último, se repiten estos pasos hasta que solo queda un clúster.

Antes de realizar cualquier agrupamiento, es necesario determinar la matriz de proximidad que contiene la distancia entre cada punto utilizando una función de distancia. Se han utilizado 4 distancias diferentes para determinar la distancia entre clusteres:

- Enlace **simple**: La distancia entre dos clústeres se define como la distancia más corta entre dos puntos de cada clúster.

$$L(r,s) = min(D(x_{ri}, x_{sj}))$$

- Enlace **completo**: La distancia entre dos clústeres se define como la distancia más corta entre dos puntos de cada clúster.

$$L(r,s) = max(D(x_{ri}, x_{sj}))$$

- Enlace **promedio**: La distancia entre dos clústeres se define como la distancia media entre cada punto de un clúster y cada punto del otro clúster.

$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

- Enlace de **Ward**: El enlace de Ward es un método de clustering jerárquico que busca minimizar la varianza dentro de los clusters en cada paso. En lugar de simplemente medir distancias entre puntos o clusters, Ward selecciona los clusters a combinar de manera que el aumento de la suma de las varianzas dentro de los clusters sea el menor posible.

$$d_{\text{Ward}}(r,s) = \frac{|r| \cdot |s|}{|r| + |s|} \cdot \|\mu_r - \mu_s\|^2$$

Este método se utiliza como una aproximación inicial debido a que no se conoce el número óptimo de clusters k para este conjunto de datos.

1.2. K-Means

K-means es un algoritmo de clustering divisivo que agrupa los datos en un número predefinido de grupos o clusters, basándose en la proximidad de los puntos a los centros de los clusters (**centroides**). El funcionamiento del algoritmo es el siguiente:

1. Se seleccionan aleatoriamente k puntos del conjunto de datos como centroides iniciales.
2. Cada punto de los datos se asigna al centroide más cercano, creando k clusters.
3. Se actualizan los centroides como el promedio de los puntos asignados a cada cluster.
4. Se repiten los dos últimos pasos hasta que los centroides no cambien significativamente entre iteraciones o hasta que alcance el número máximo de iteraciones.

1.3. Partitioning Around Medoids

Partitioning around medoids (PAM) o K-medoids es un algoritmo de agrupamiento relacionado con el algoritmo K-means, también divide los datos en k clusters, pero en lugar de utilizar centroides, utiliza **medoides**. Un medoide es el punto real en el conjunto de datos que mejor representa a su cluster (en lugar de un promedio, como en el k-means). El funcionamiento del algoritmo es el siguiente:

1. Se seleccionan aleatoriamente k puntos del conjunto de datos como medoides iniciales.
2. Cada punto del conjunto de datos se asigna al medoide más cercano, creando k clusters.
3. Para mejorar la calidad de los clusters, el algoritmo intenta reemplazar un medoide con cualquier otro punto del conjunto de datos y evalúa si esta sustitución reduce la suma de las distancias dentro de los clusters.
4. El proceso se repite hasta que no se pueda reducir más el coste de los clusters.

PAM es más robusto frente a **outliers** y clusters con formas complejas que K-means, ya que los outliers afectan mucho a los centroides (que son un promedio), mientras que los medoides son puntos reales del conjunto de datos, lo que los hace menos sensibles a valores extremos.

2. Resultados

Para comenzar, dado que se desconocía el número óptimo de clusters para obtener el mayor índice de silhouette, se generaron varios modelos utilizando el método de clustering jerárquico, para la distancia euclídea y la distancia correlación.

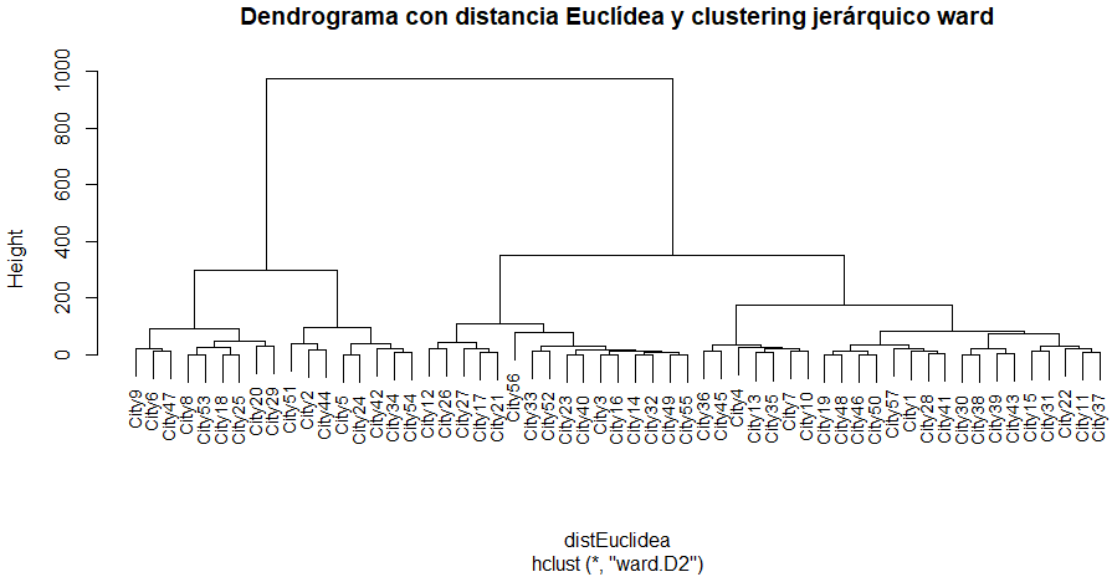

Figura 1: Dendograma distancia euclídea y enlace ward

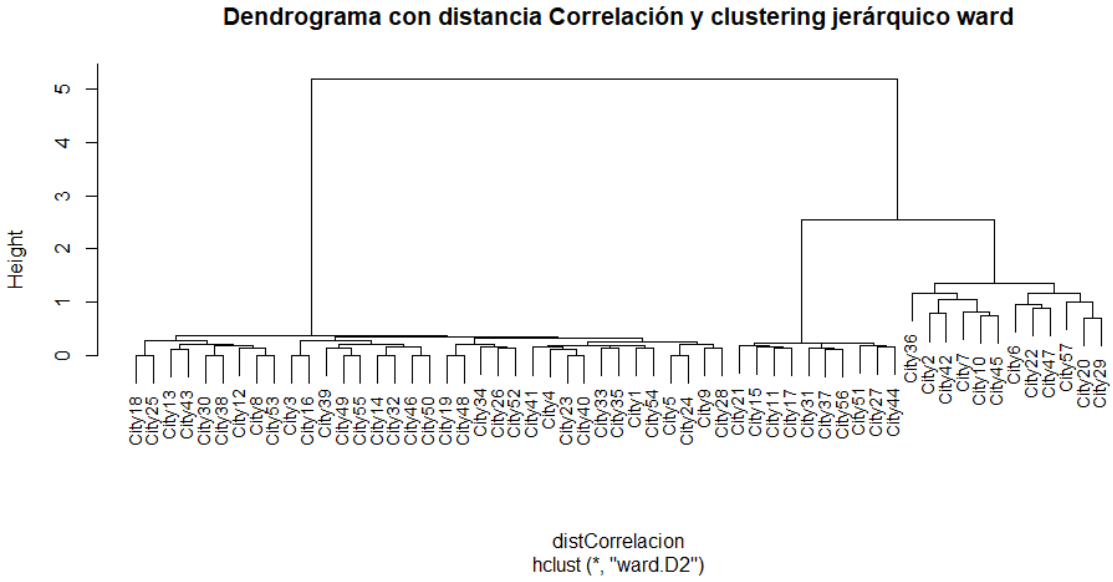

Figura 2: Dendograma distancia correlación y enlace ward

Estos dendograma se han generado con el método de clustering jerárquico y el enlace ward. Se puede observar como en un nivel muy bajo hay muchas agrupaciones, lo que significa que los puntos en esa región están muy concentrados, lo que lleva a una pronta agrupación. En cambio, en un nivel alto se observa como se unen dos agrupaciones, lo que significa que esos clusters o grupos eran muy diferentes entre sí en términos de distancia o similitud.

Para verificar si las suposiciones observadas en los dendrogramas son correctas, se ha calculado el índice de silhouette para varios modelos.

Distancia	**Enlace**	**Número de cluster**	**Valor**
Euclídea	Simple	16	0.473
	Completo	2	0.588
	Promedio	2	0.588
	Ward	2	0.653
Correlación	Simple	12	0.631
	Completo	9	0.650
	Promedio	9	0.650
	Ward	6	0.651

Tabla 1: Índice de silhouette para varios modelos

En la tabla se puede observar que, para la distancia euclídea, el número óptimo de clusters es 2, mientras que para la distancia de correlación es 9. Además, el enlace de Ward ha logrado el mayor índice de silhouette para ambas distancias.

Para confirmar que esos números de clusters son adecuados, se han creado los siguientes modelos utilizando el algoritmo de Partitioning Around Medoids (PAM).

El primer gráfico (Figura 3) muestra dos clusters con un índice de silhouette bastante similares. En ambos clusters, hay varias ciudades con índices de silhouette bajos, lo que sugiere que estas ciudades no están bien representadas por los clústeres, probablemente porque se encuentran demasiado alejadas de los centros de los mismos.

El segundo gráfico (Figura 4) presenta nueve clústeres con índices de silueta variados. Más de la mitad de las ciudades están agrupadas en el clúster 1, que tiene un índice de silueta elevado, lo que indica que están bien clasificadas. Sin embargo, algunos clústeres muestran un índice de 0 y contienen solo una instancia, están lo suficientemente alejado de otros datos como para no ser incluidas en ningún otro cluster. Finalmente, en el clúster 2 hay dos valores negativos, lo que indica que estos objetos probablemente están mal clasificados.

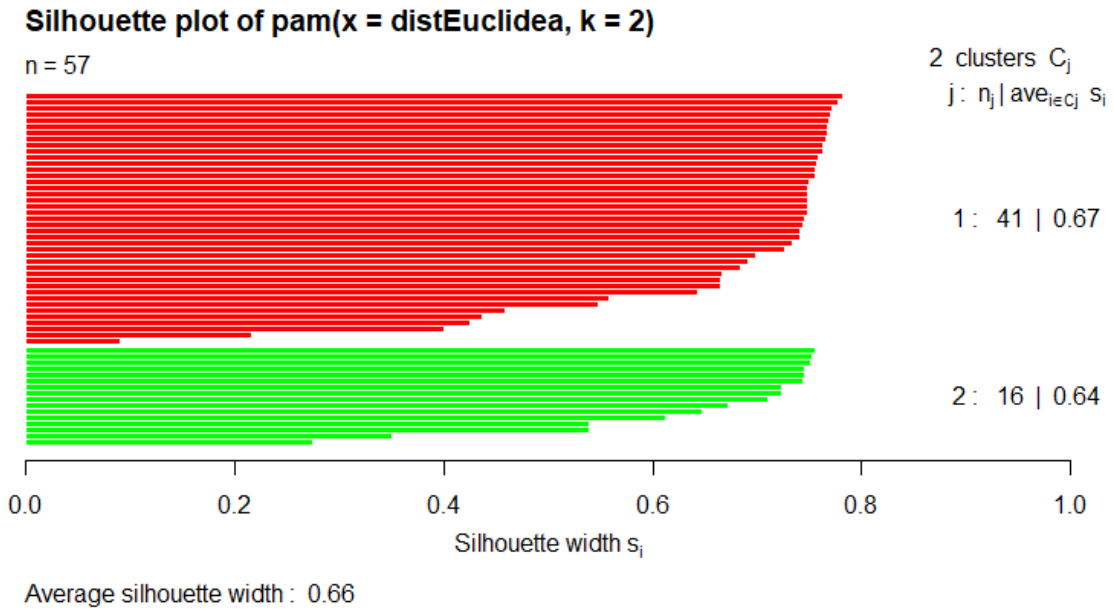

Figura 3: Valores de silhouette con método PAM para distancia euclídea

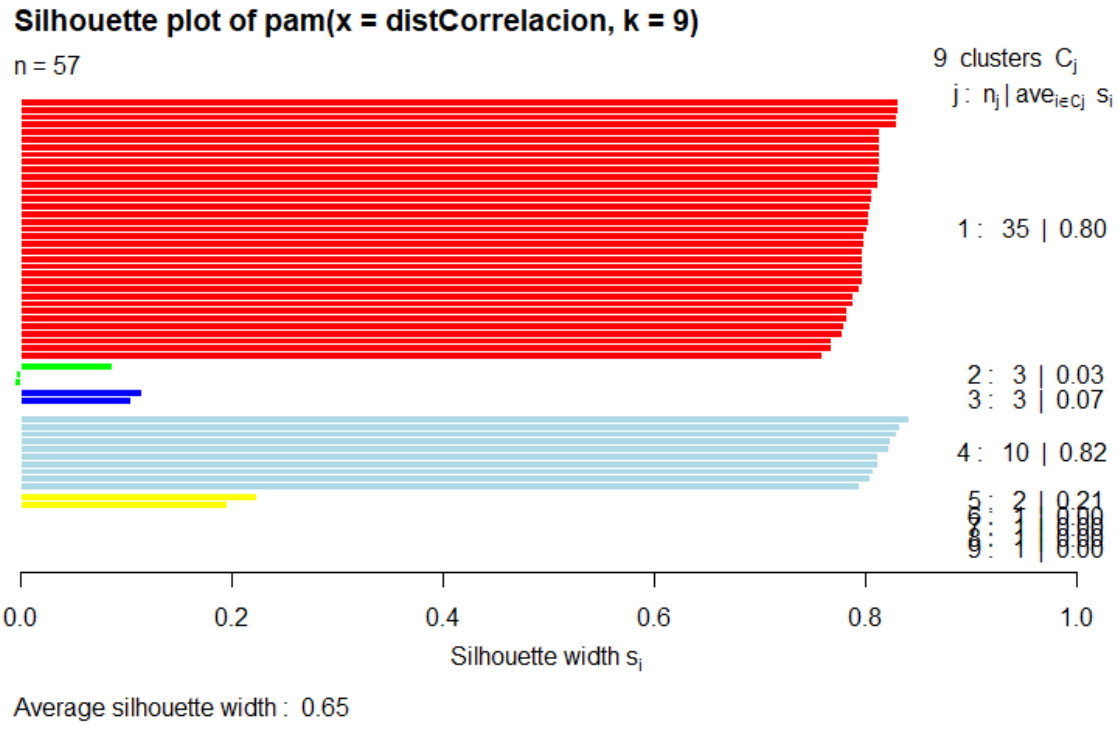

Figura 4: Valores de silhouette con método PAM para distancia correlación

A continuación, se van a comparar los clusters en los que se agrupan las ciudades, utilizando una **matriz de confusión**. La primera comparación será entre la agrupación jerárquica promedio y la agrupación PAM para la distancia euclídea.

		PAM	
		C1	C2
Jerárquico	C1	40	0
	C2	1	16

Tabla 2: Matriz de confusión de 2 clusters

En la tabla 2 se puede observar, en la diagonal, como dos métodos diferentes agrupan todas las ciudades menos una al mismo cluster.

La segunda comparación será entre la agrupación jerárquica completa y la agrupación PAM para la distancia correlación.

		PAM								
		C1	C2	C3	C4	C5	C6	C7	C8	C9
Jerárquico	C1	35	0	0	0	0	0	0	0	0
	C2	0	2	0	0	0	0	0	0	0
	C3	0	0	3	0	0	0	0	0	0
	C4	0	0	0	10	0	0	0	0	0
	C5	0	0	0	0	2	0	0	0	0
	C6	0	0	0	0	0	1	0	1	0
	C7	0	0	0	0	0	0	1	0	0
	C8	0	1	0	0	0	0	0	0	0
	C9	0	0	0	0	0	0	0	0	1

Tabla 3: Matriz de confusión de 9 clusters

En la tabla 3 se puede observar, en la diagonal, como dos métodos diferentes agrupan todas las ciudades menos dos al mismo cluster.

Por último, se han representado los municipios en un espacio de dimensión reducida, en este caso 2 dimensiones.

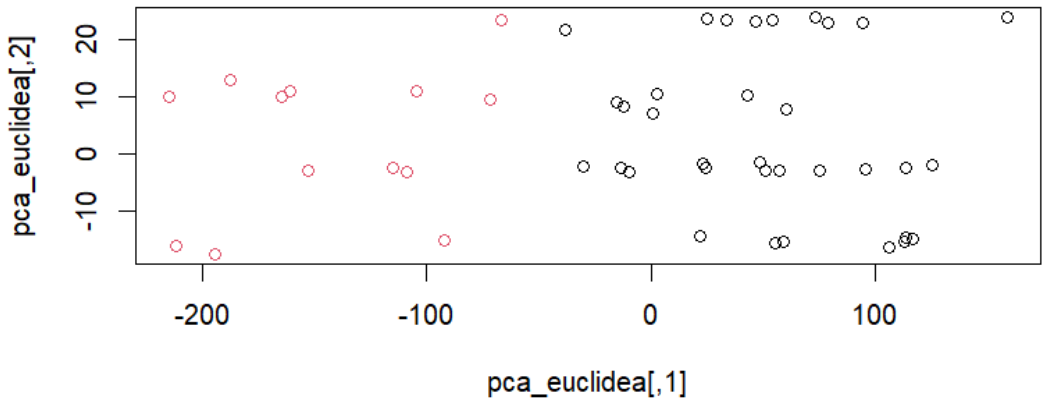

Figura 5: PCA distancia euclídea

En este gráfico, se han proyectado los datos utilizando la distancia euclídea y se han dividido en 2 grupos, que representan el número óptimo de clústeres para esta métrica. Sin embargo, los puntos aparecen bastante dispersos a lo largo del gráfico.

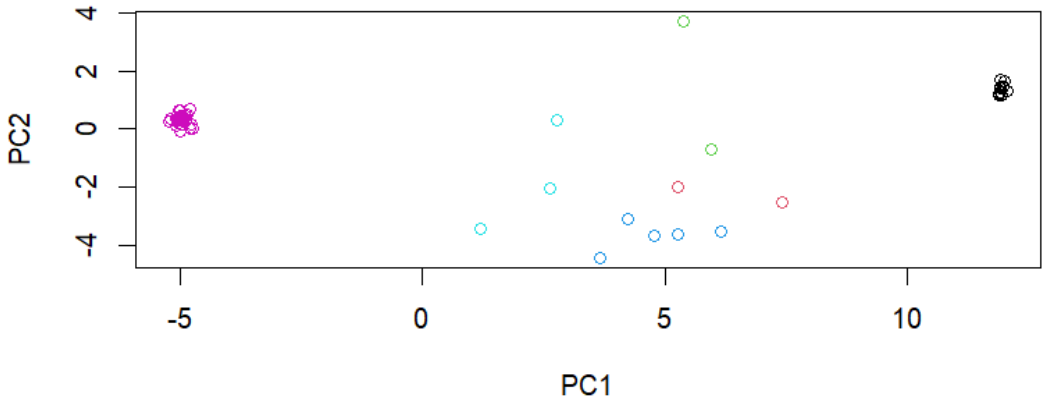

Figura 6: PCA distancia correlación

En este otro gráfico, los datos se han proyectado usando la distancia de correlación. Aquí se observan dos agrupaciones bien definidas en los bordes. Además, hay varios puntos en el centro que se dividen en diferentes clústeres.

3. Conclusiones

Para concluir, se han extraído varias observaciones a partir de las gráficas y tablas presentadas en el apartado anterior.

La selección entre la distancia euclídea y la distancia de correlación para el clustering influye notablemente en los resultados, ya que cada métrica capta diferentes aspectos de las relaciones entre los datos. La distancia de correlación resulta más adecuada para identificar similitudes en los patrones o relaciones entre variables, mientras que la distancia euclídea es más útil cuando se busca agrupar puntos cercanos físicamente.

Al emplear la distancia euclídea, el número óptimo de clústeres es 2, mientras que con la distancia de correlación se obtienen 9 clústeres. Con un número mayor de clústeres, algunos se centran en agrupar datos más alejados, lo que facilita su detección, pero aporta poca información sobre con qué otros datos deberían agruparse.

Las tablas 2 y 3 muestran que, a pesar de utilizar diferentes métodos, la mayoría de los datos se agrupan en los mismos clústeres, lo que indica que están bastante relacionados entre sí.

Finalmente, al reducir la dimensionalidad con PCA y coordenadas principales, el componente principal 1 (PC1) podría estar capturando si los delitos en los municipios han seguido una tendencia general al alza o a la baja. Esto sugiere que el PC1 refleja la dirección de variación entre municipios que han experimentado un aumento constante en los delitos frente a aquellos con una disminución.