# Análisis de Series de Vazoes

*Cristian David Pachacama*

*7 de noviembre de 2017*

## Introducción

En el presente documento presentamos un análisis Macro de las series de tiempo asociadas a las estaciones de `VAZOES`, para ello utilizaremos un conjunto de técnicas multivariantes como:

- Escalonamiento Multidimensional
- Análisis Clúster

Mismas que detallamos más a continuación.

### Escalonamiento Multidimensional (MDS)

Es un conjunto de técnicas que permiten visualizar un conjunto de objetos en un espacio de dimensión $N$ ($N = 2$ usualmente, y definida a priori), esto a partir de una matriz de disimilitud (similitud o distancia) entre dichos objetos.

### Análisis Clúster

Es un conjunto de técnicas que buscan agrupar un conjunto de objetos de tal manera que los miembros del mismo grupo (llamado clúster) sean lo más similares posibles, en algún sentido.

## Metodología

Primero construimos la matriz de distacias entre las series de tiempo asociadas a los `VAZOES`, para ello escogemos una de las métricas definidas para series de tiempo. Luego, a partir de la matriz de distancias utilizamos la técnica de MDS clásica, a partir de ella se obtiene una nube de puntos (en dimensión $N = 2$) donde cada punto representa a la serie de `VAZOES` de una estación.

A partir de esta nube de puntos en dimensión $N = 2$, usamos el Análisis Clúster para crear grupos de puntos basandose en su cercanía (esta cercanía depende fundamentalmente de la métrica elegida).

A continuación presentamos los resultados al haber seguido los puntos anteriores escogiendo una métrica específica en cada caso.

## Resultados

Con ayuda del software estadístico R y de sus paquetes `TSdist`,`TSclust` (usado para construir las matrices de distacias) y `smacof` (usado para ejecutar la técnica MDS), mostramos a continuación resultados correspondientes a considerar 9 métricas definidas para series de tiempo, a partir de cada una de ellas y usando MDS se obtiene una nube de puntos (etiquetados por un código que corresponde al código ONS de la estación correspondiente). Luego, formamos los grupos mediante el Análisis Cluster (jerárquico), y en este caso elegiremos el número de grupos con la ayuda de la función `fviz_nbclust()` del paquete `factoextra`, que halla el número óptimo de grupos. Cabe mencionar que los grupos se grafican de distintos colores, además se

delimitan los grupos coloreando la envolvente convexa de cada grupo (delimitando así los grupos en regiones). Adicionalemte se muestra el Dendograma asociado a la creación de los grupos, y finalmente se grafica las series de tiempo de cada agrupación.

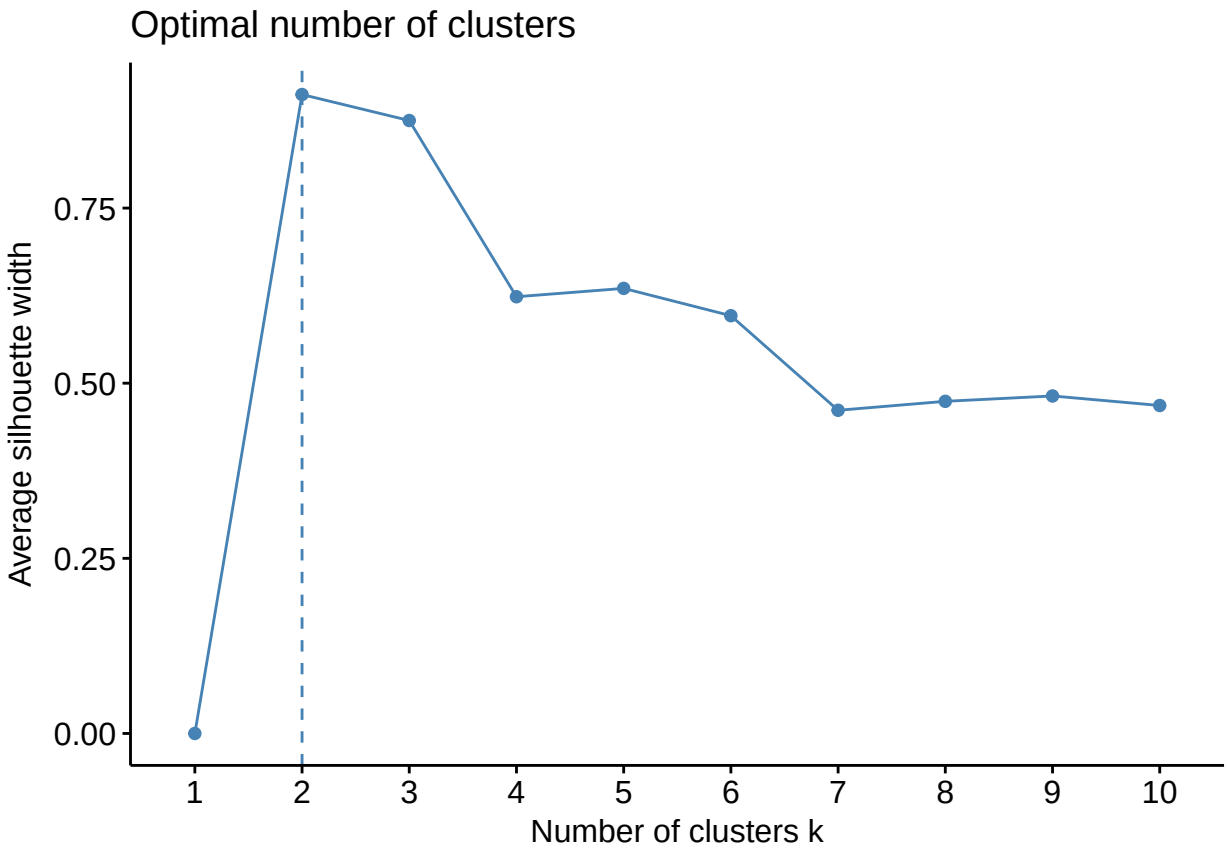
## Sección 1

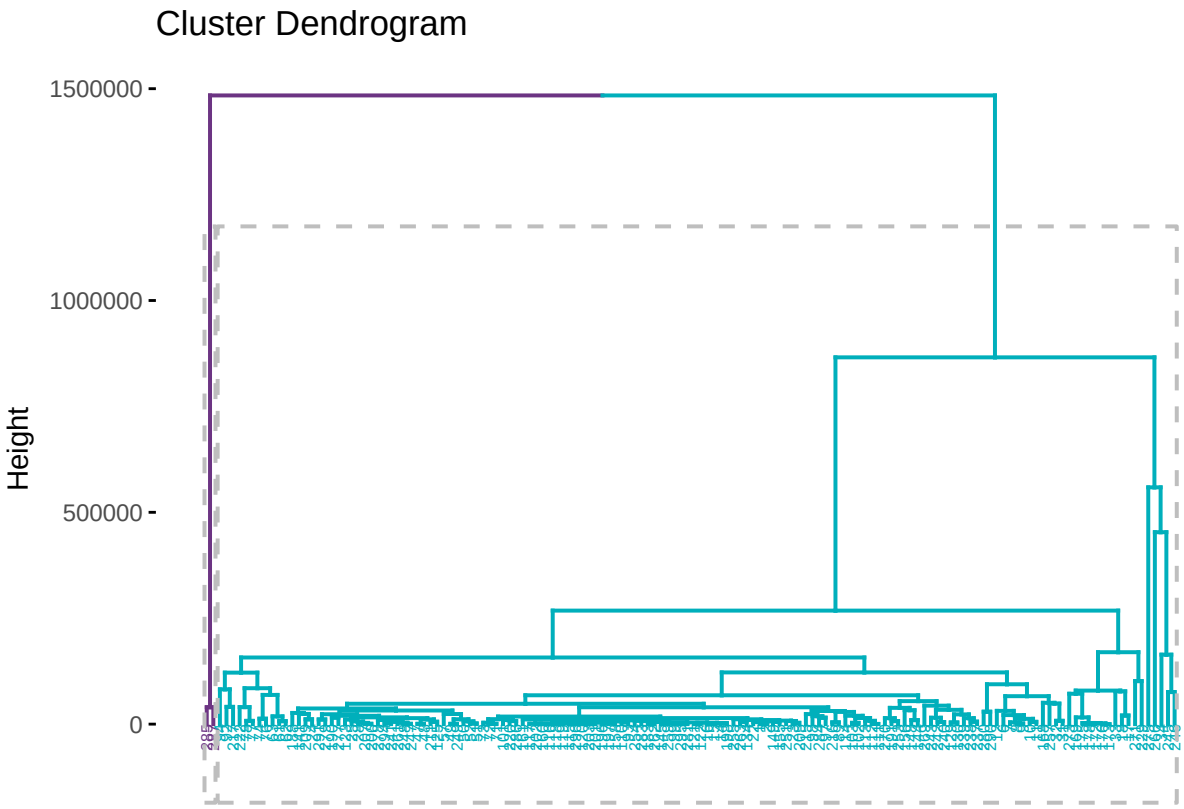
En esta sección mostramos resultados obtenidos al considerar métricas que consideran distancias "geométricas" entre las series.

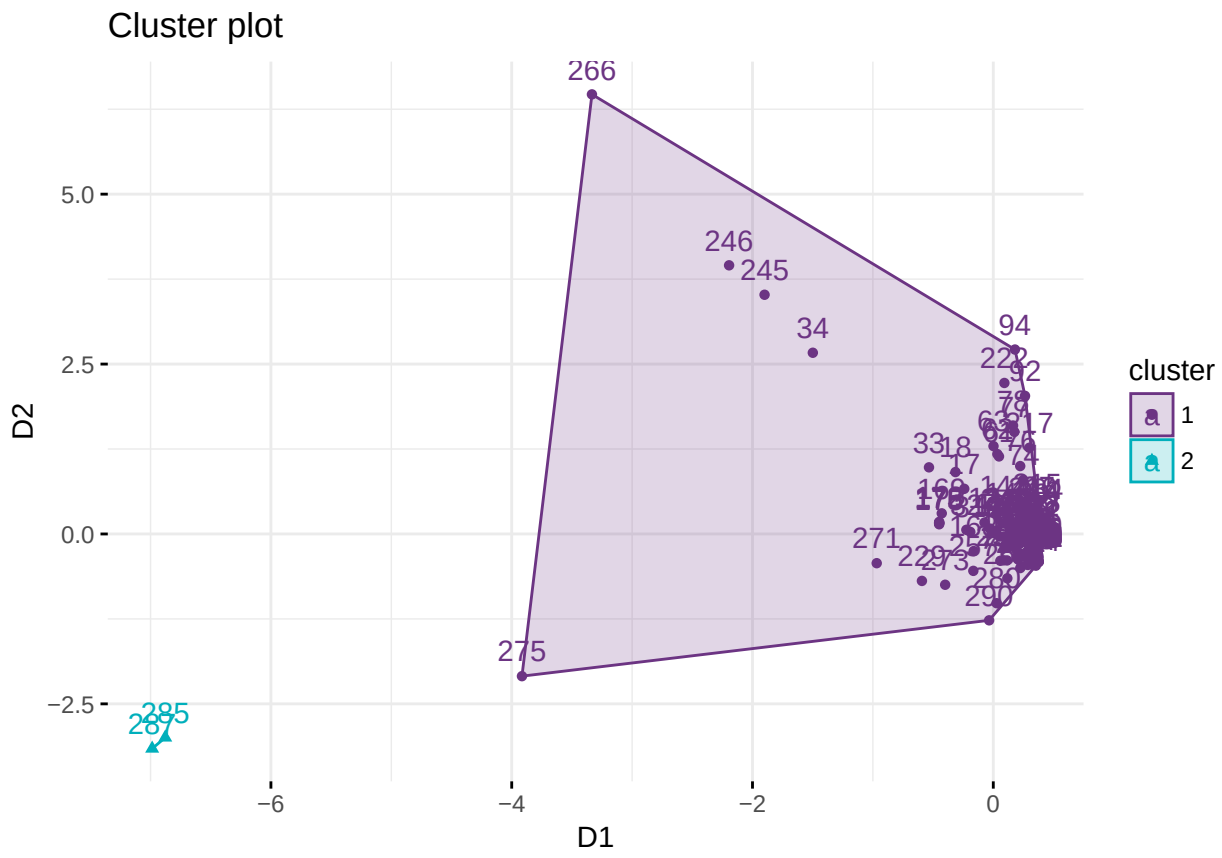
### Métrica Euclidea

Sean $(X_t)$, $(Y_t)$ dos series de tiempo, a valores en $\mathbb{R}$, con $t \in T$. Se definen entonces las siguientes métricas (distancias).

$$d_{euc}(X_t, Y_t) = \sqrt{\sum_{t \in T} (x_t - y_t)^2}$$

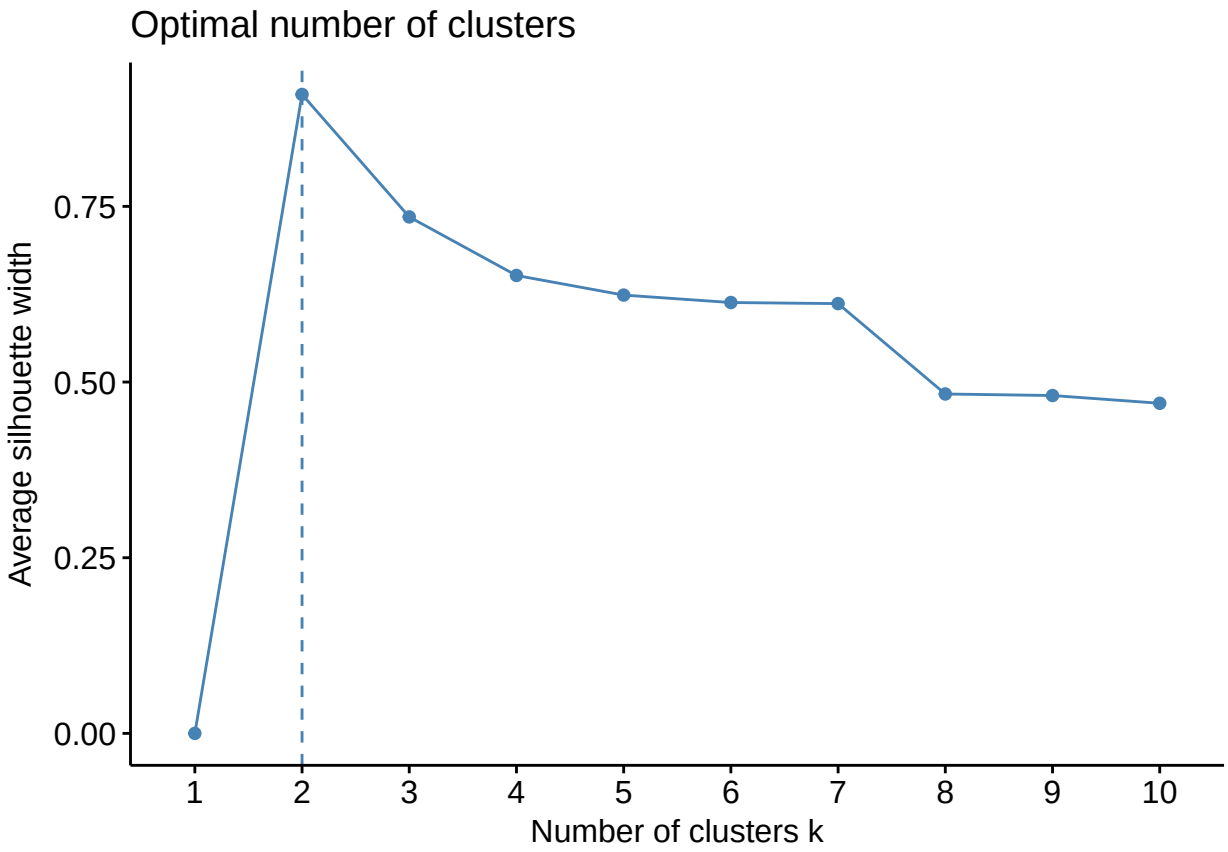


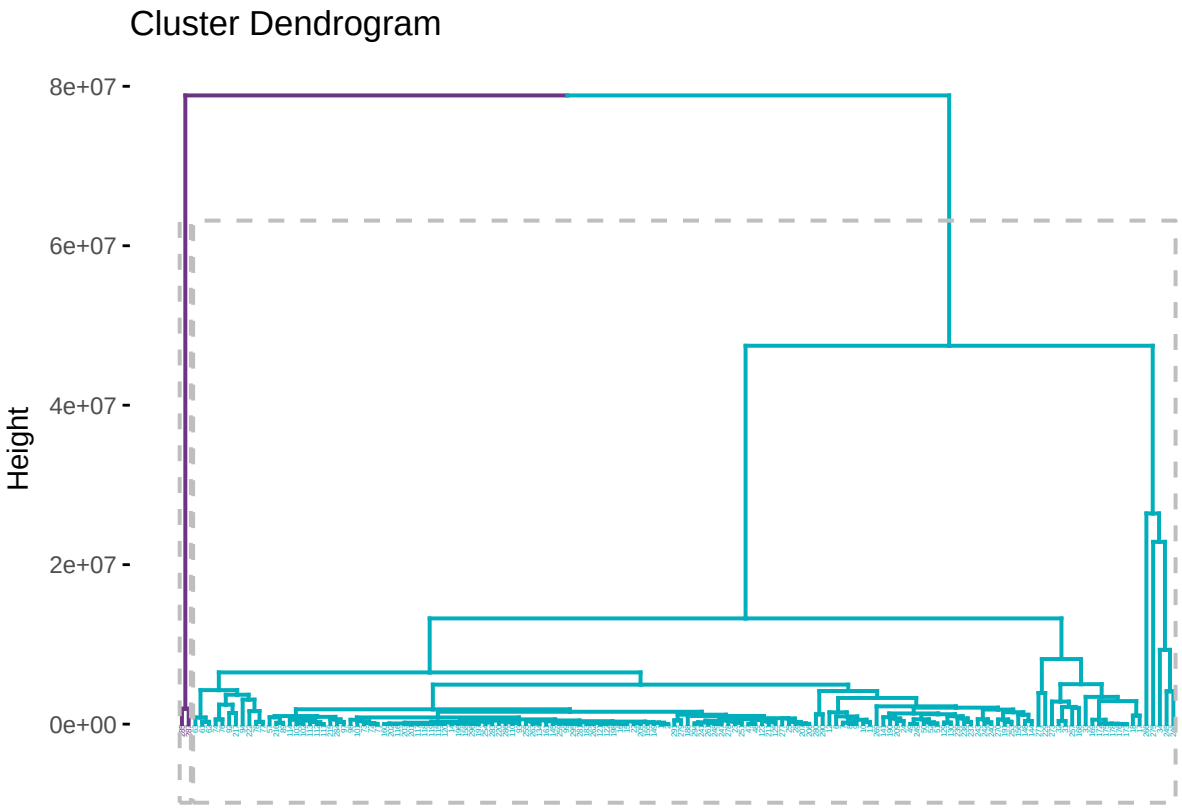

Cluster Dendrogram
Height
1500000
1000000
500000
0

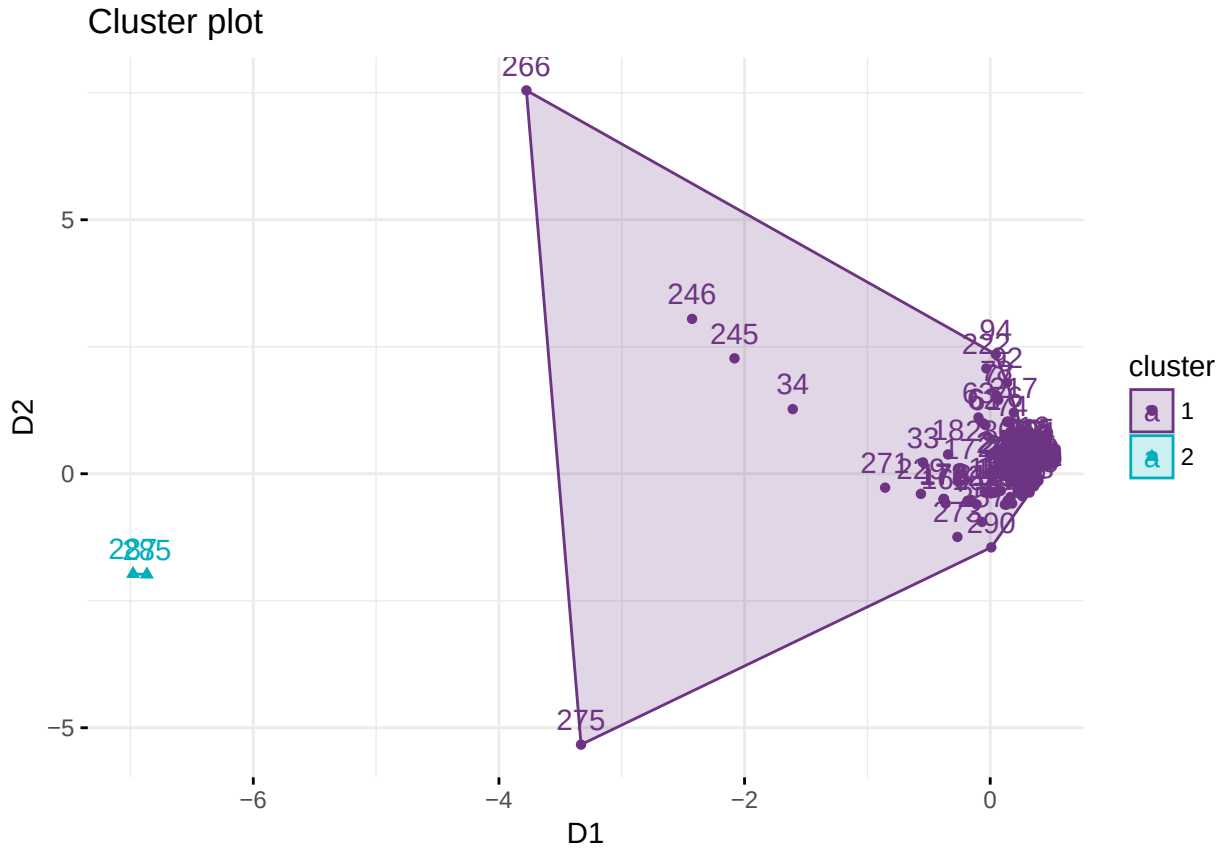



**Distancia Manhattan**

$$d_{manh}(X_t, Y_t) = \sum_{t \in T} |x_t - y_t|$$

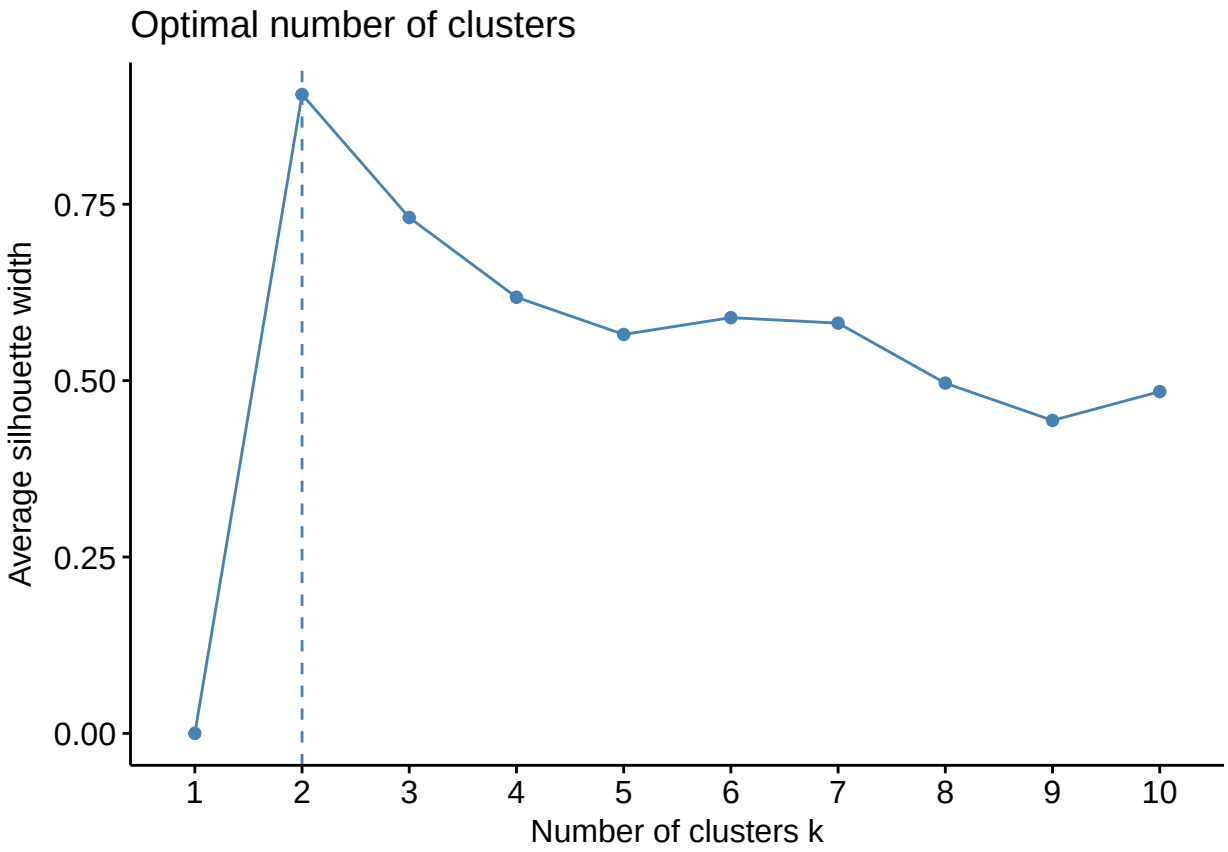

Optimal number of clusters
Average silhouette width
0.00
0.25
0.50
0.75
1
2
3
4
5
6
7
8
9
10
Number of clusters k

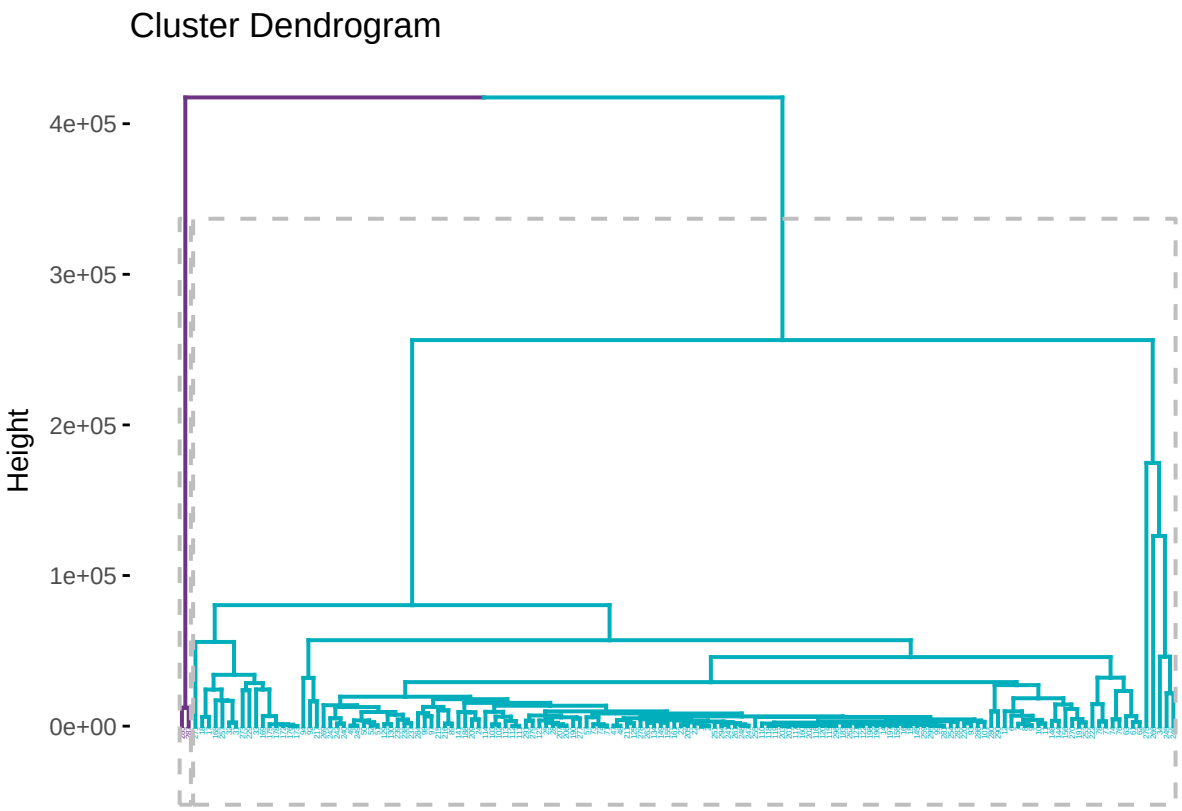

Cluster Dendrogram
Height
8e+07
6e+07
4e+07
2e+07
0e+00

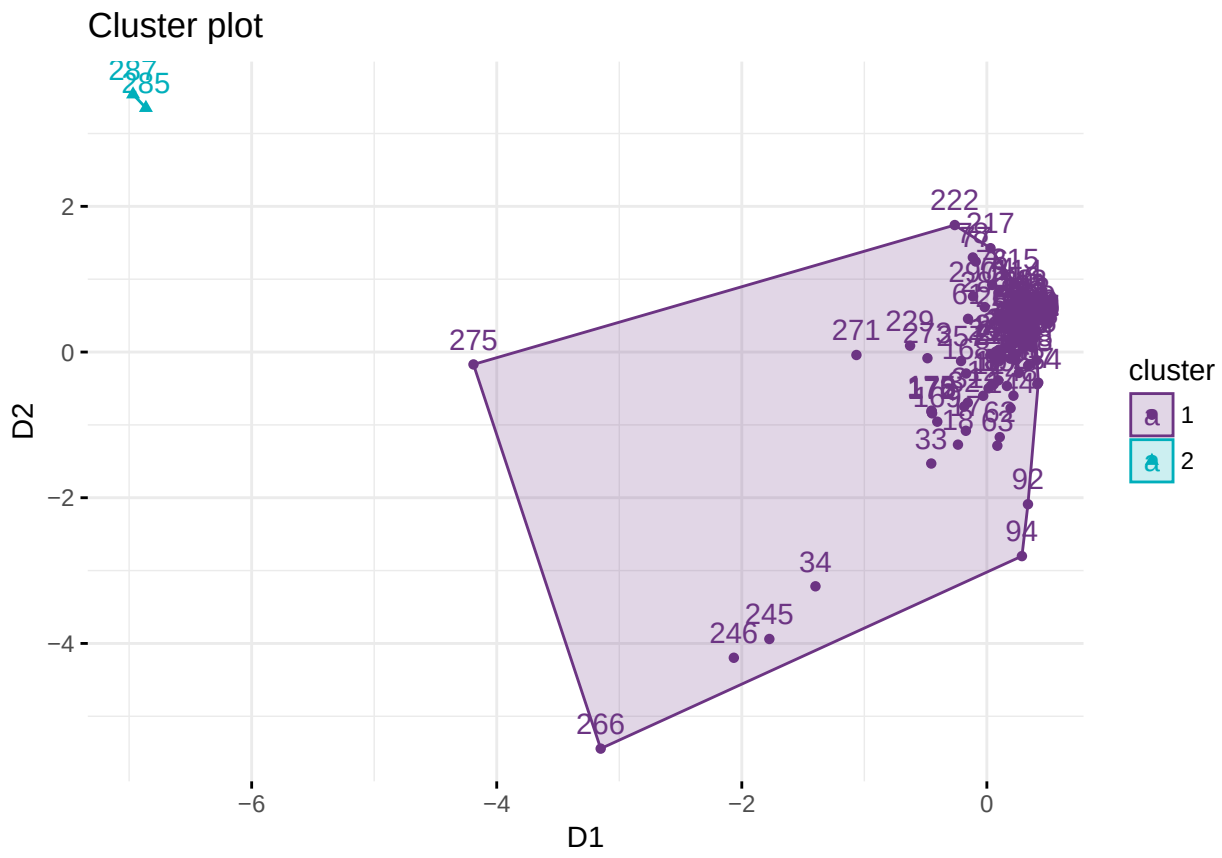



**Métrica de Minkowski**

$$d_{mink}(X_t, Y_t) = \sqrt[p]{\sum_{t \in T} (x_t - y_t)^p}$$

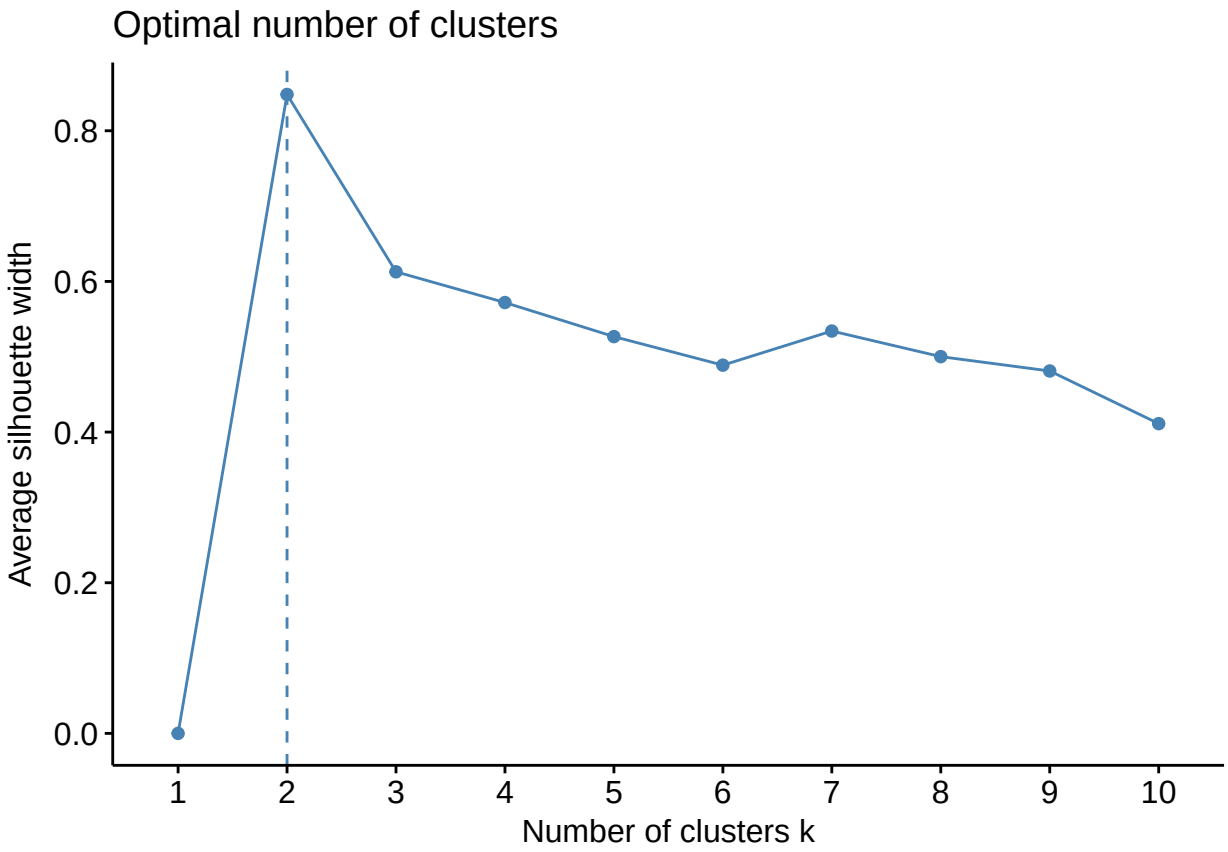

Optimal number of clusters
Average silhouette width
0.00
0.25
0.50
0.75
1
2
3
4
5
6
7
8
9
10
Number of clusters k

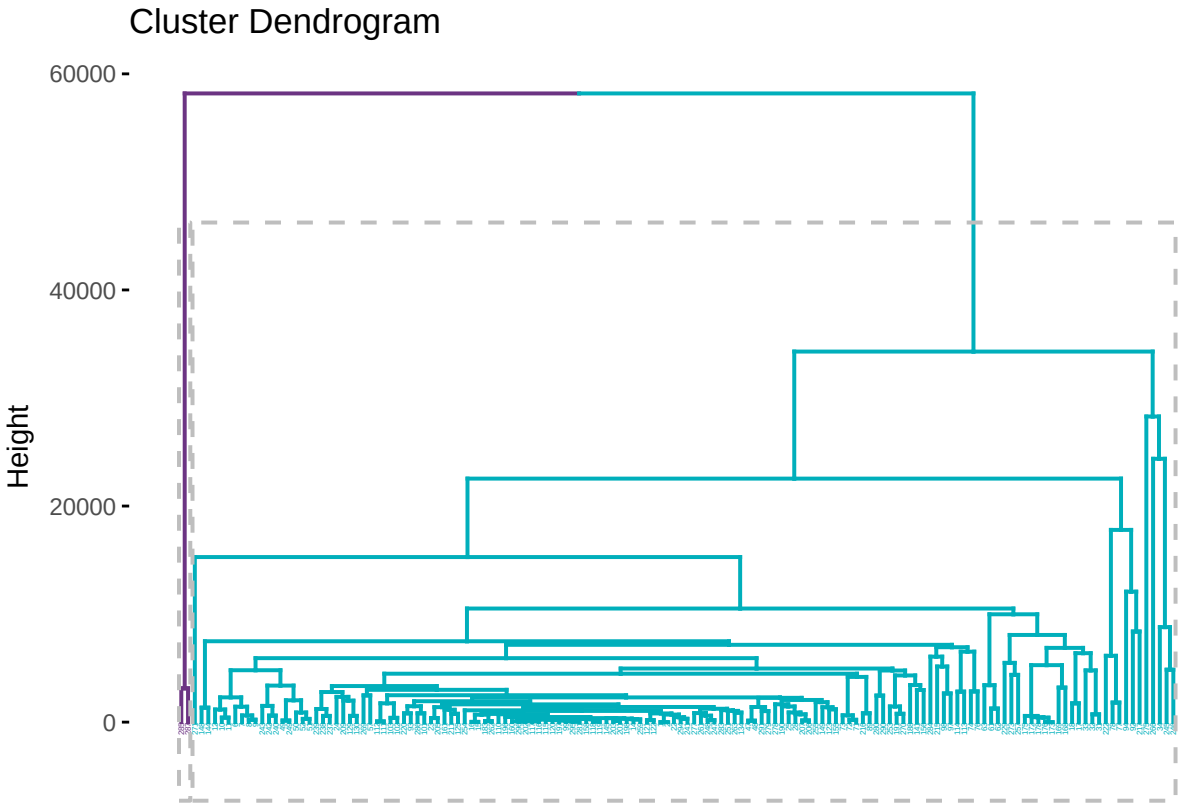

Cluster Dendrogram
Height
4e+05
3e+05
2e+05
1e+05
0e+00

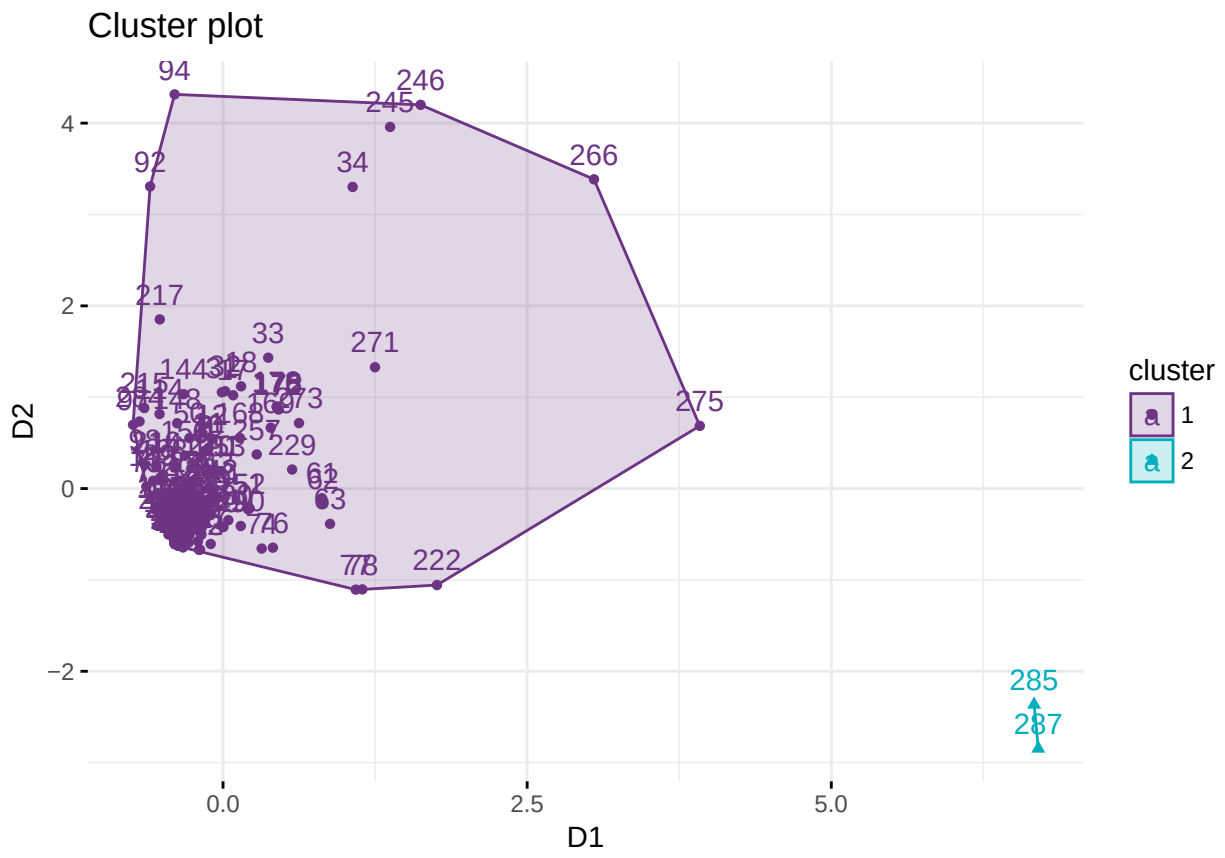



**Norma Infinito**

$$d_{inf}(X_t, Y_t) = \max_{t \in T} |x_t - y_t|$$

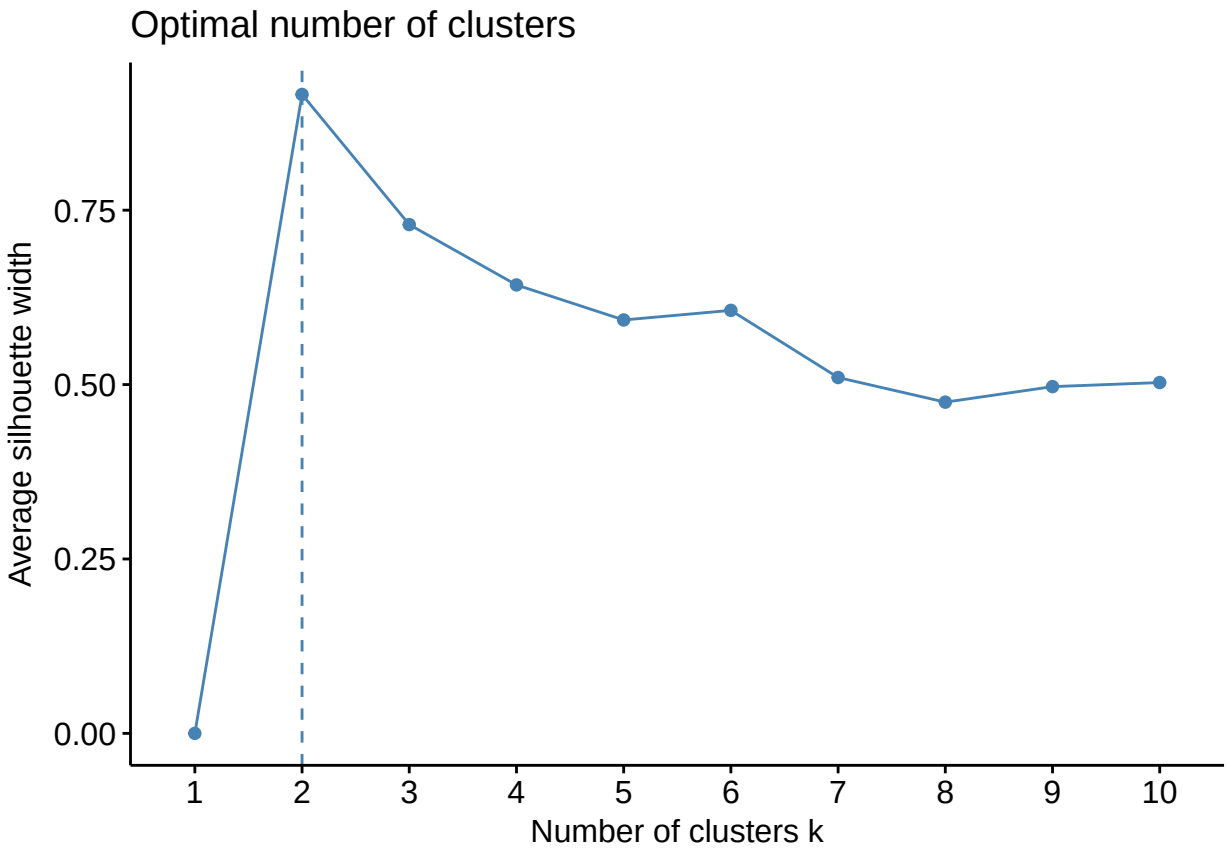

Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
Number of clusters k

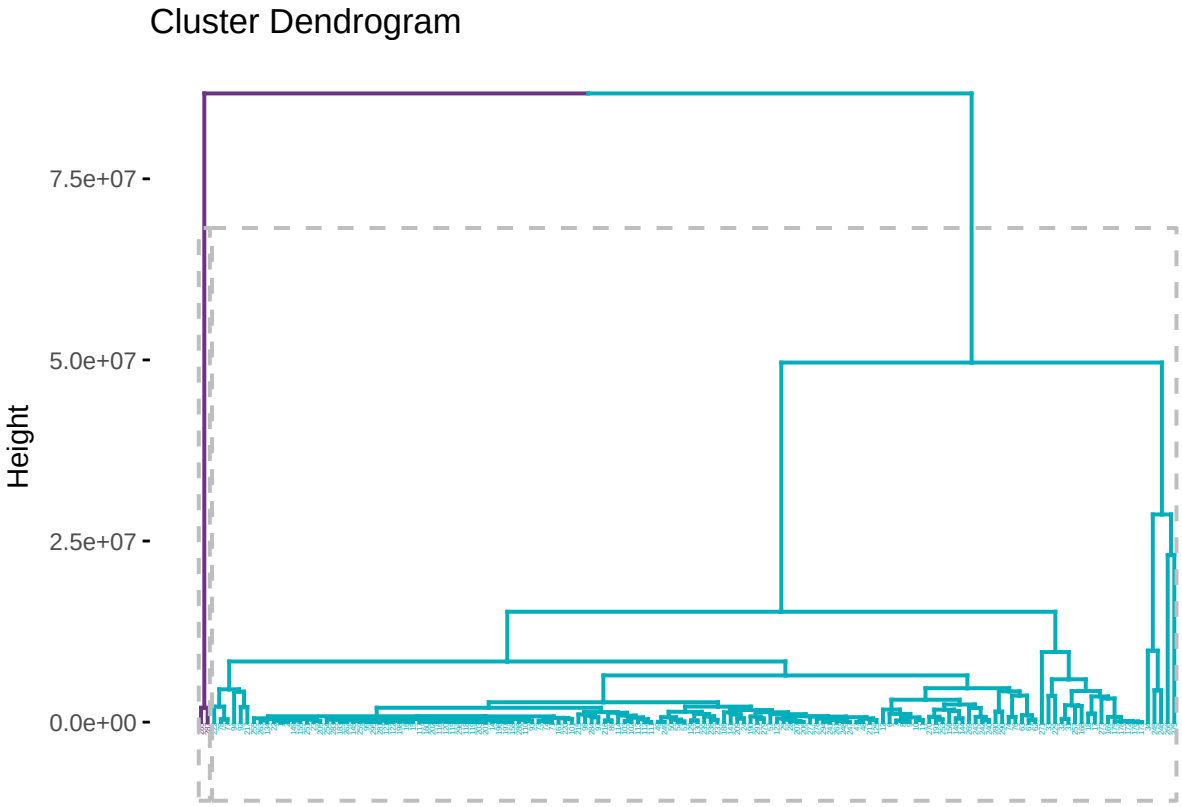

Cluster Dendrogram
Height
60000
40000
20000
0

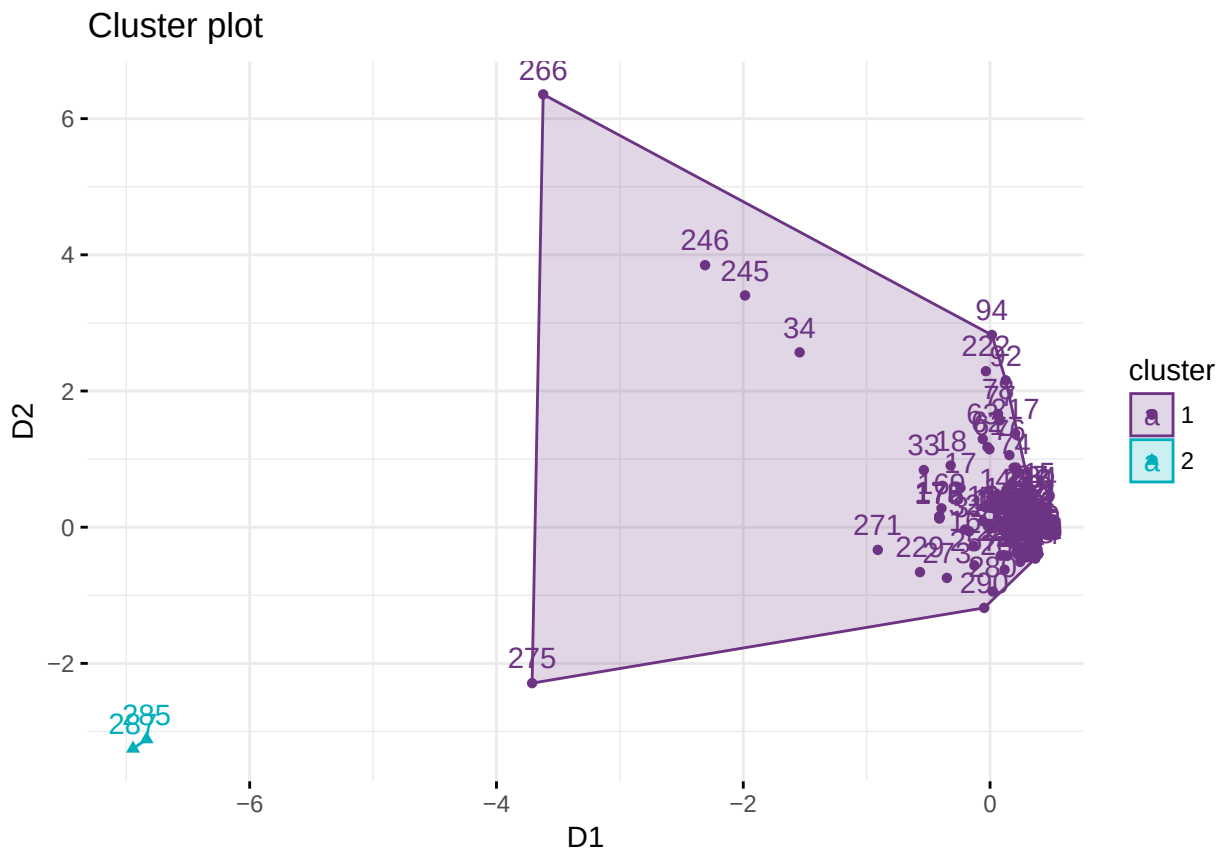



### Distancia de Fourier

Se calcula como la distancia euclidiana entre los primeros $n$ coeficientes de Fourier de las series $x$ e $y$. Las series deben tener la misma longitud. Además, $n$ no debería ser mayor que la longitud de la serie.


Optimal number of clusters
Average silhouette width
0.00
0.25
0.50
0.75
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Cluster Dendrogram
Height
7.5e+07
5.0e+07
2.5e+07
0.0e+00




### Disimilitud CORT

Índice de disimilitud que combina la correlación temporal y los comportamientos de valores sin procesar.

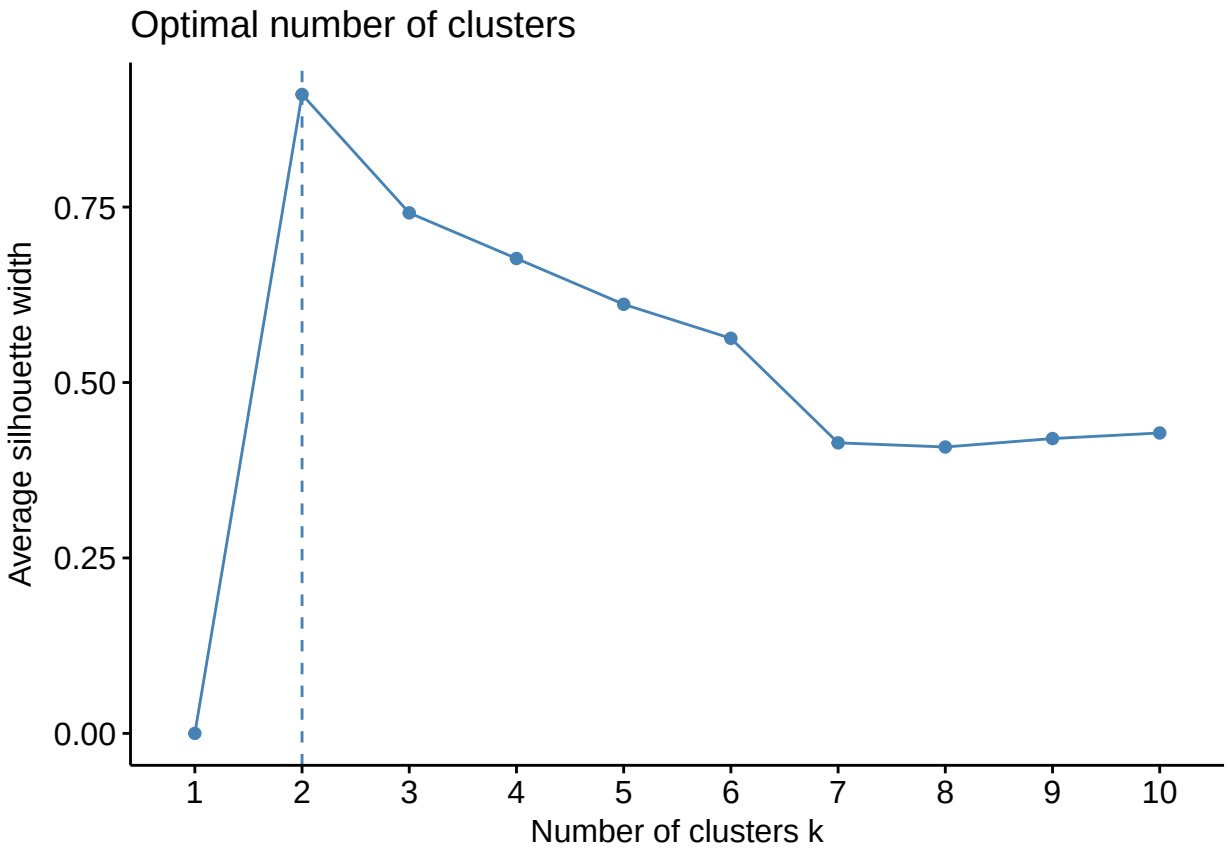
$$d_{cort}(X_t, Y_t) = \Phi[CORT(x, y)]\delta(x, y)$$

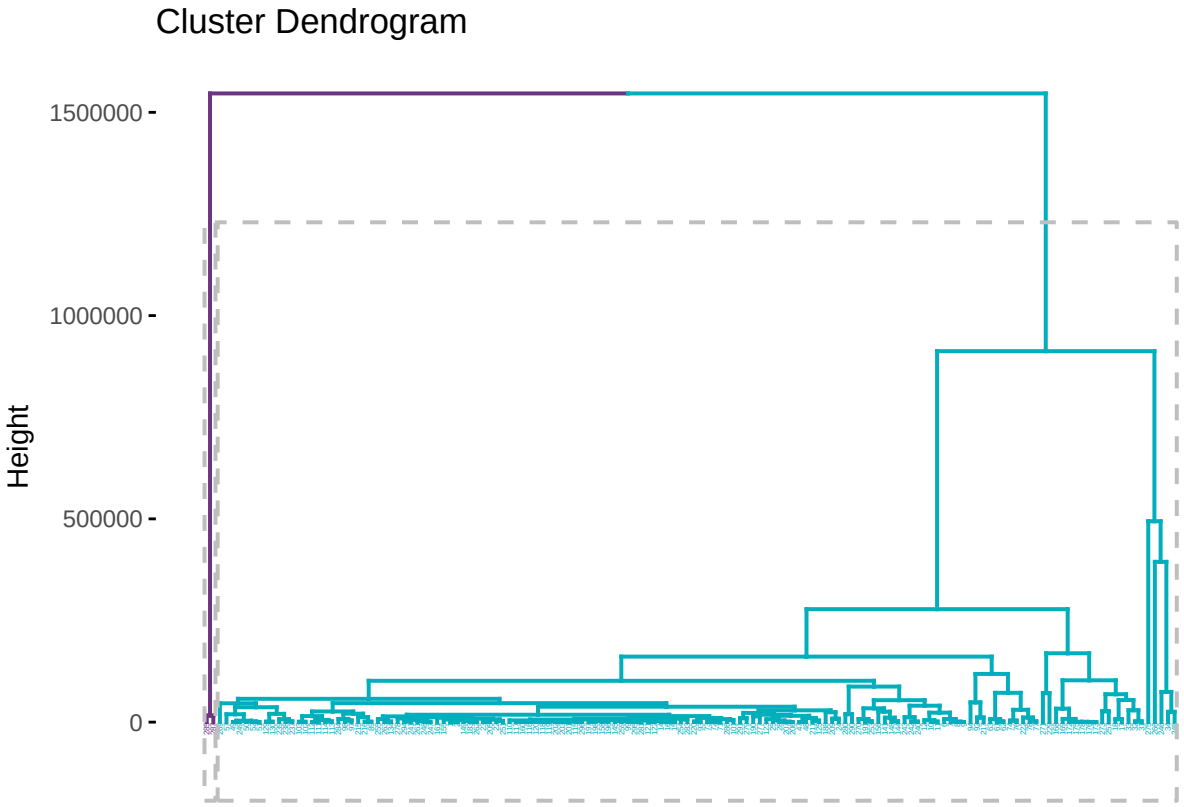
Donde:

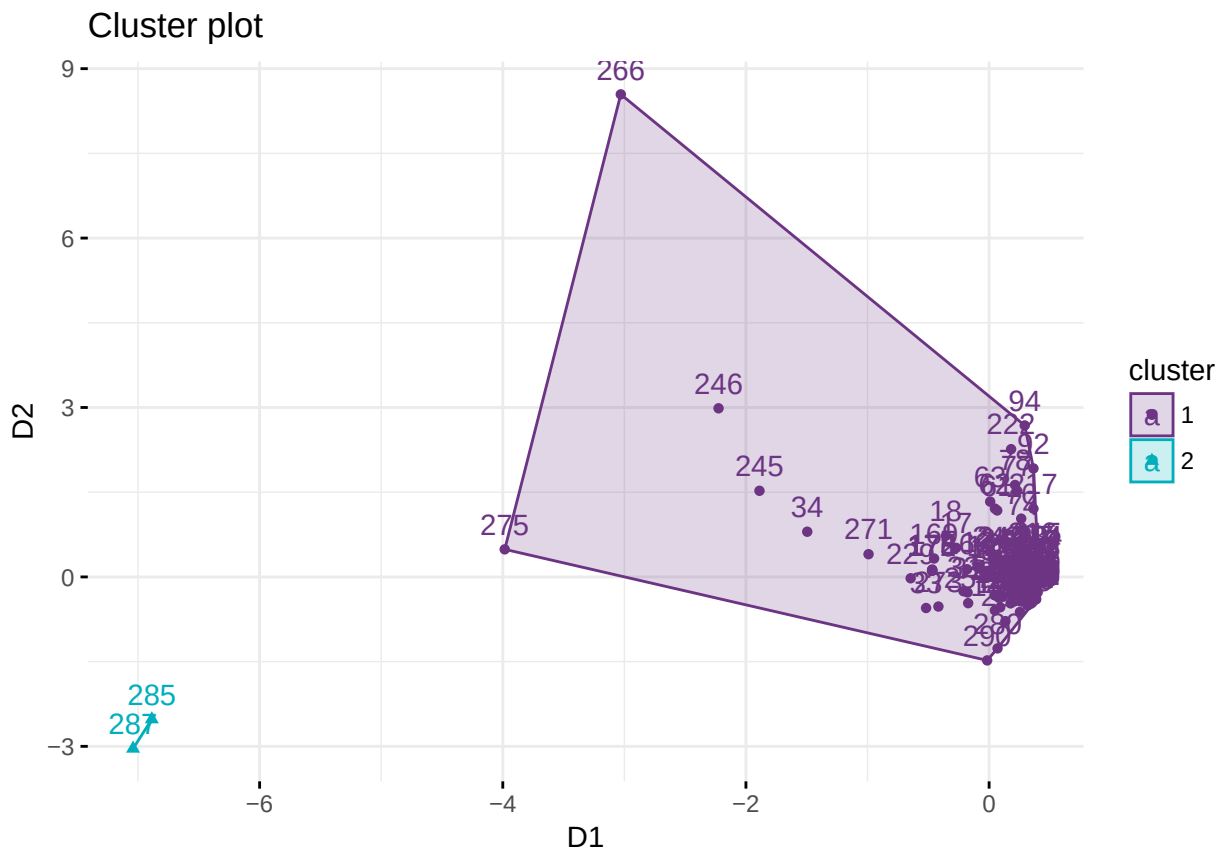
$$CORT(x, y) = \frac{\sum_{t\in T}(x_{t+1} - x_t)(y_{t+1} - y_t)}{\sqrt{\sum_{t\in T}(x_{t+1} - x_t)^2}\sqrt{\sum_{t\in T}(y_{t+1} - y_t)^2}}$$

$$\Phi(u) = \frac{2}{1 + e^{ku}}$$

Y $\delta(x, y)$ es una medida de disimilitud entre los elementos de una misma serie.


Optimal number of clusters
Average silhouette width
0.00
0.25
0.50
0.75
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Cluster Dendrogram
Height
1500000
1000000
500000
0




### Observacion 1

Como se puede observar todas estas métricas desenvocan en resultados parecidos en cuanto al número de grupos que se forman con las series (2 en todos los casos), y además las nubes de puntos (asociados a las estaciones de Vazoes) son bastante similares en todos los casos.

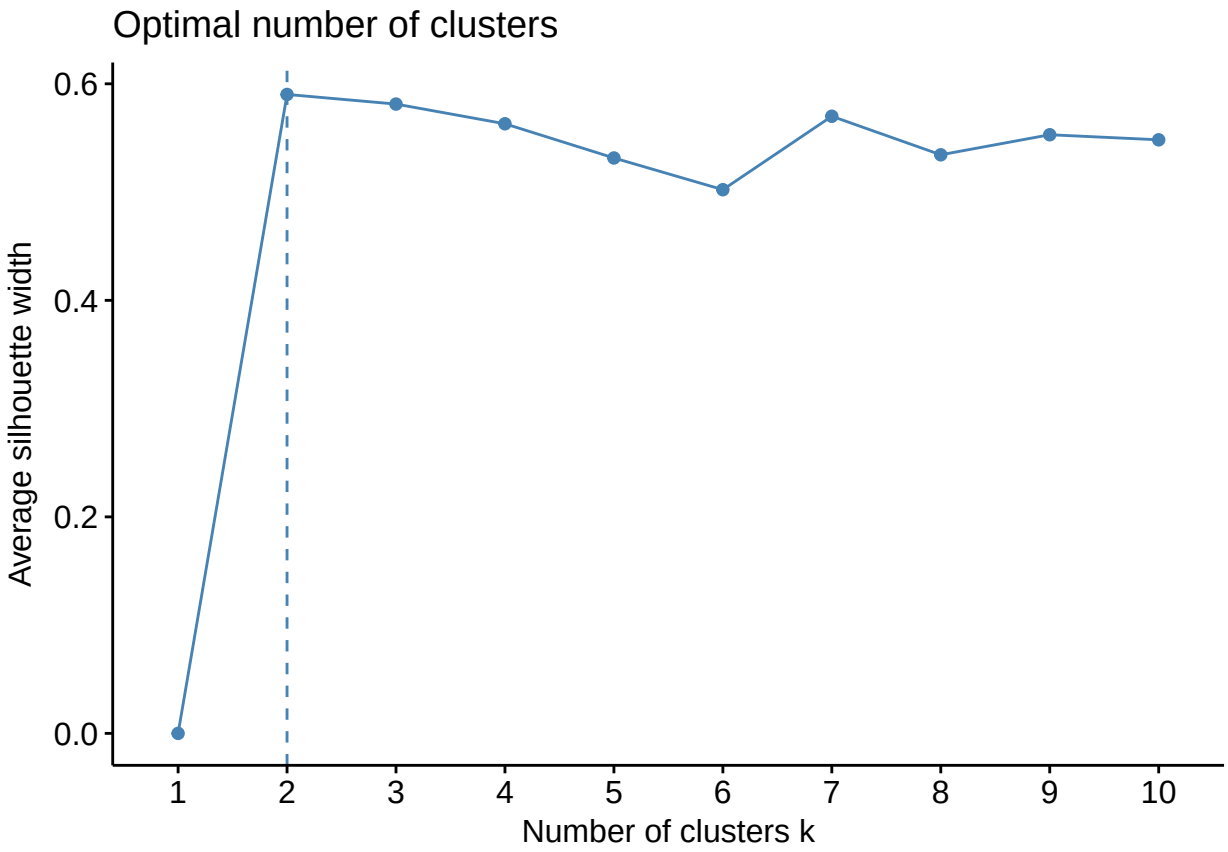
## Sección 2

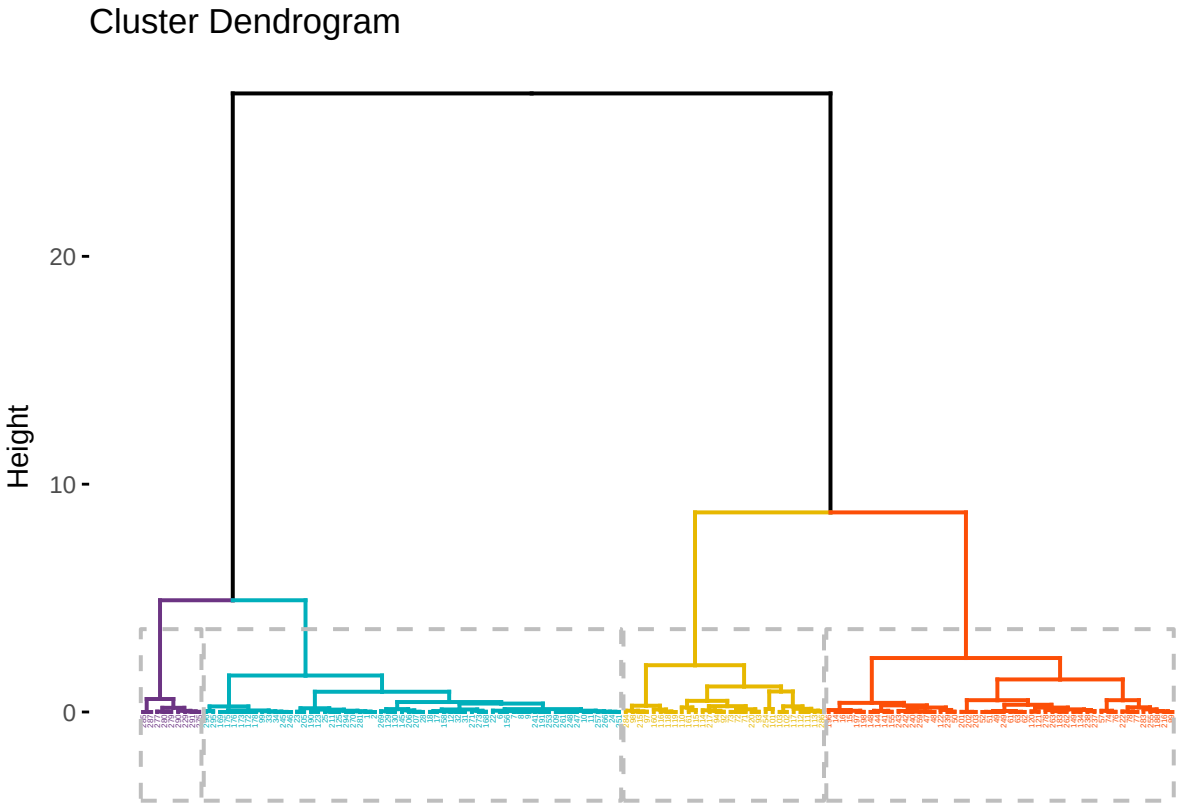
En esta sección mostramos resultados obtenidos al considerar métricas que consideran la correlación en y entre las series.

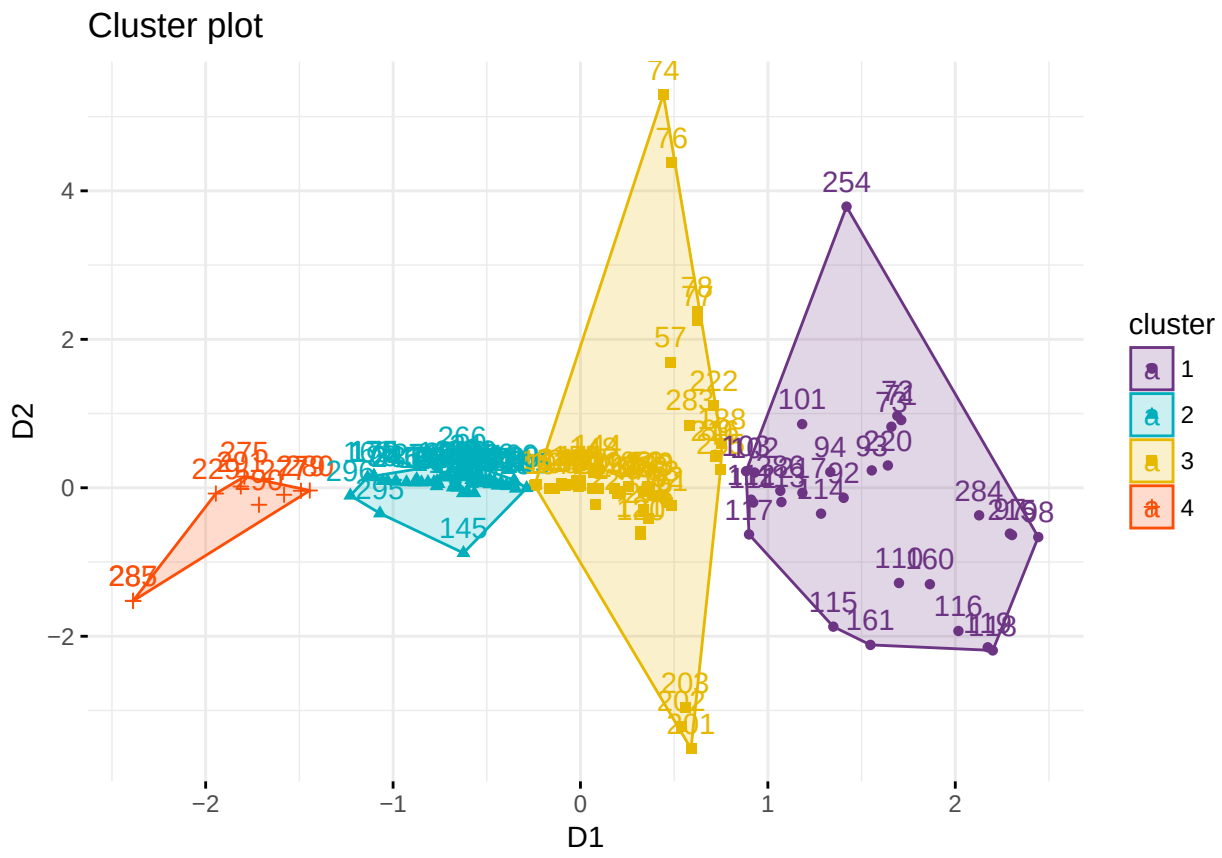
### Disimilitud basada en la Autocorrelación

$$d_{acf}(X_t, Y_t) = \sqrt{(\hat{\rho}_x - \hat{\rho}_y)^t \Omega (\hat{\rho}_x - \hat{\rho}_y)}$$

donde $\hat{\rho}_x$ es el vector con los coeficientes de autocorrelación.


Optimal number of clusters
Average silhouette width
0.0
0.2
0.4
0.6
1
2
3
4
5
6
7
8
9
10
Number of clusters k

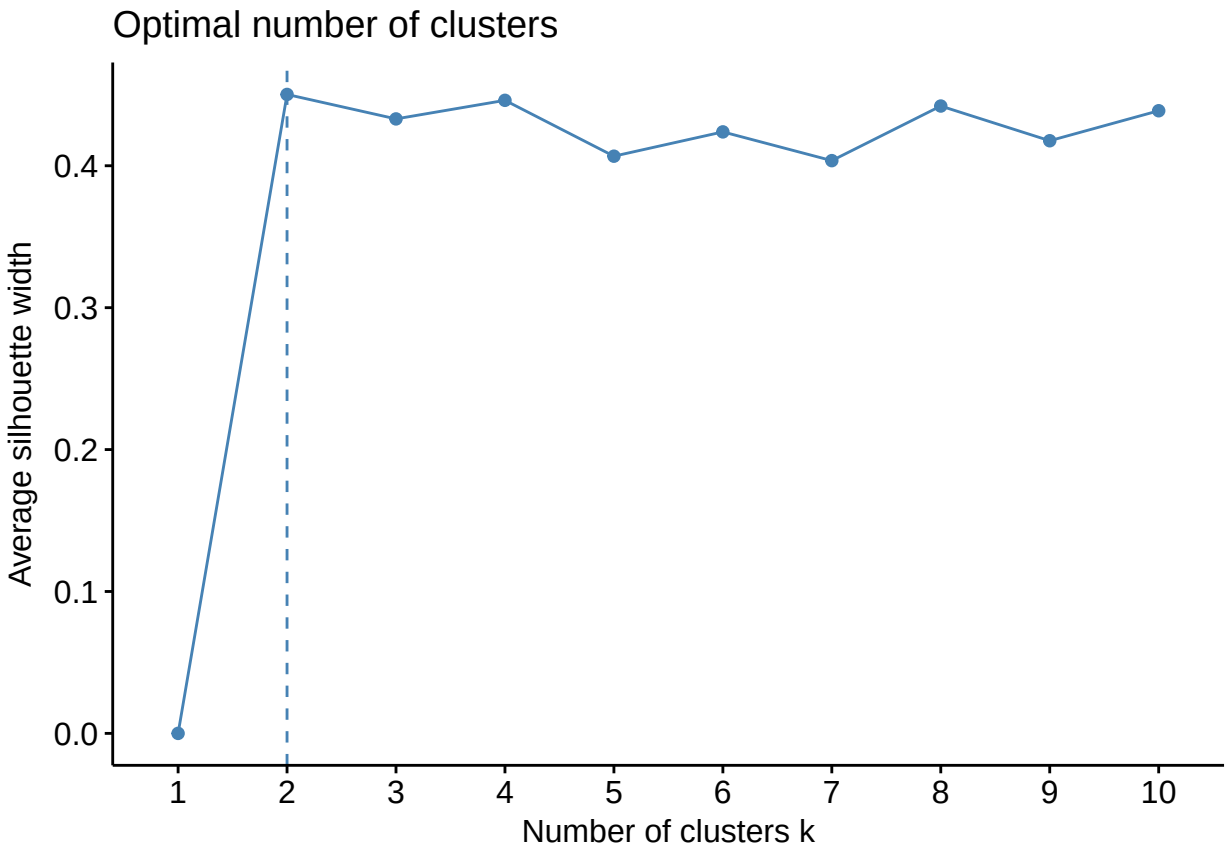

Cluster Dendrogram
Height
20
10
0

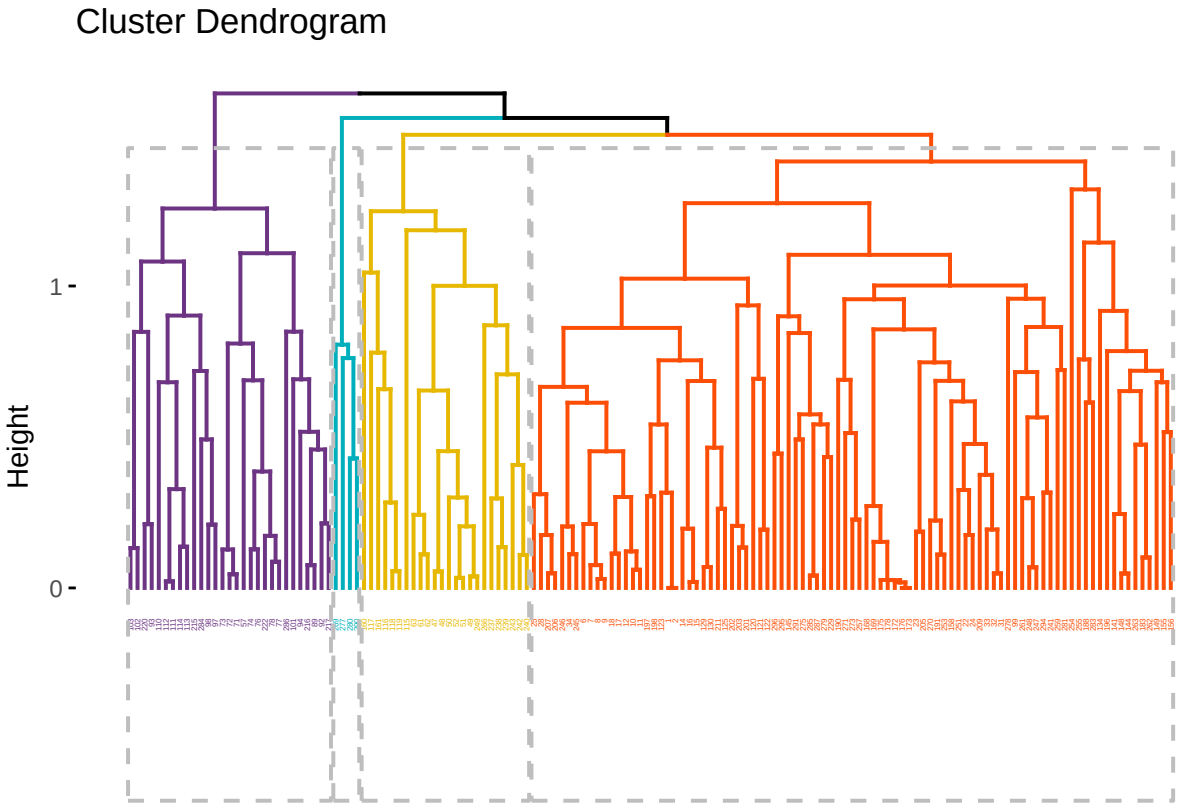


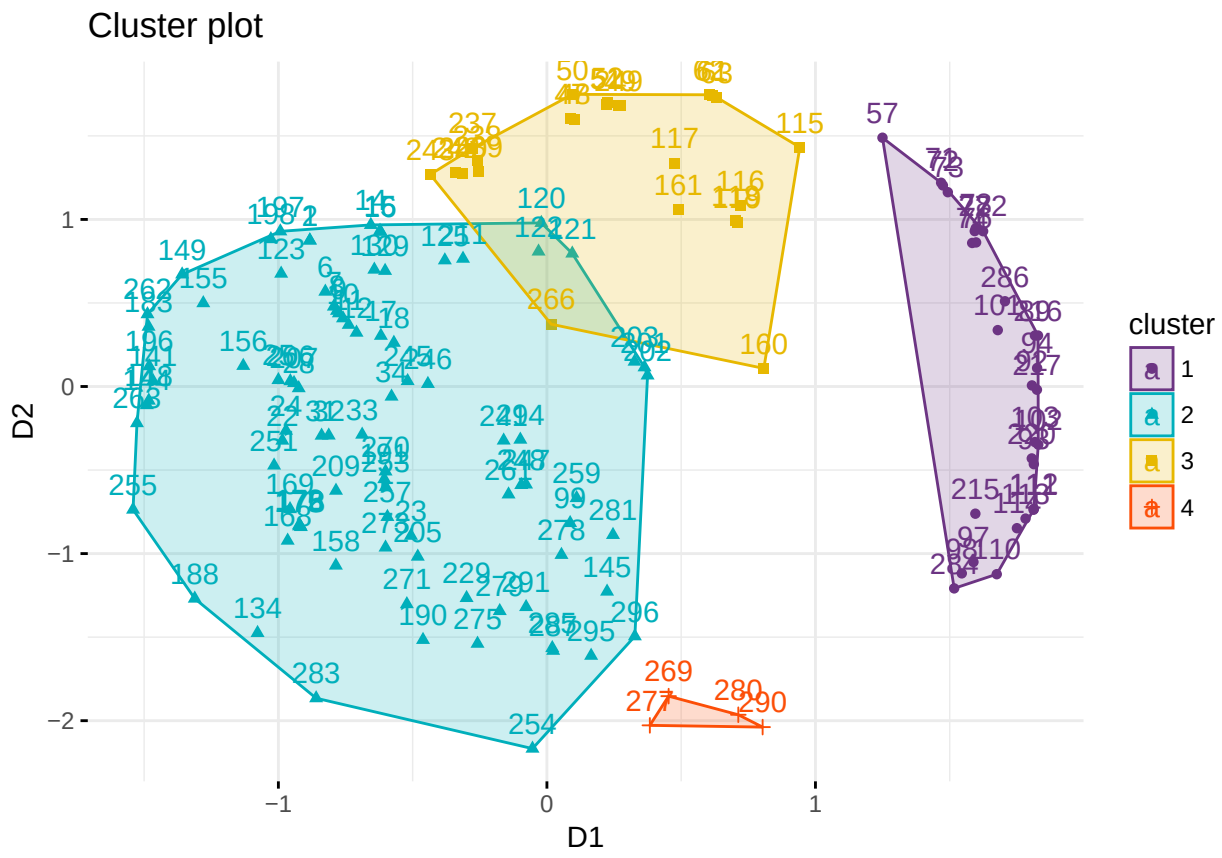

### Disimilitud basada en Correlación

$$d_{cor}(X_t, Y_t) = \sqrt{\left(\frac{1-\rho}{1+\rho}\right)^{\beta}}$$

donde $\rho$ es el coeficiente de correlación de Pearson entre las series, y $\beta$ se define a priori.


Optimal number of clusters
Average silhouette width
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
6
7
8
9
10
Number of clusters k

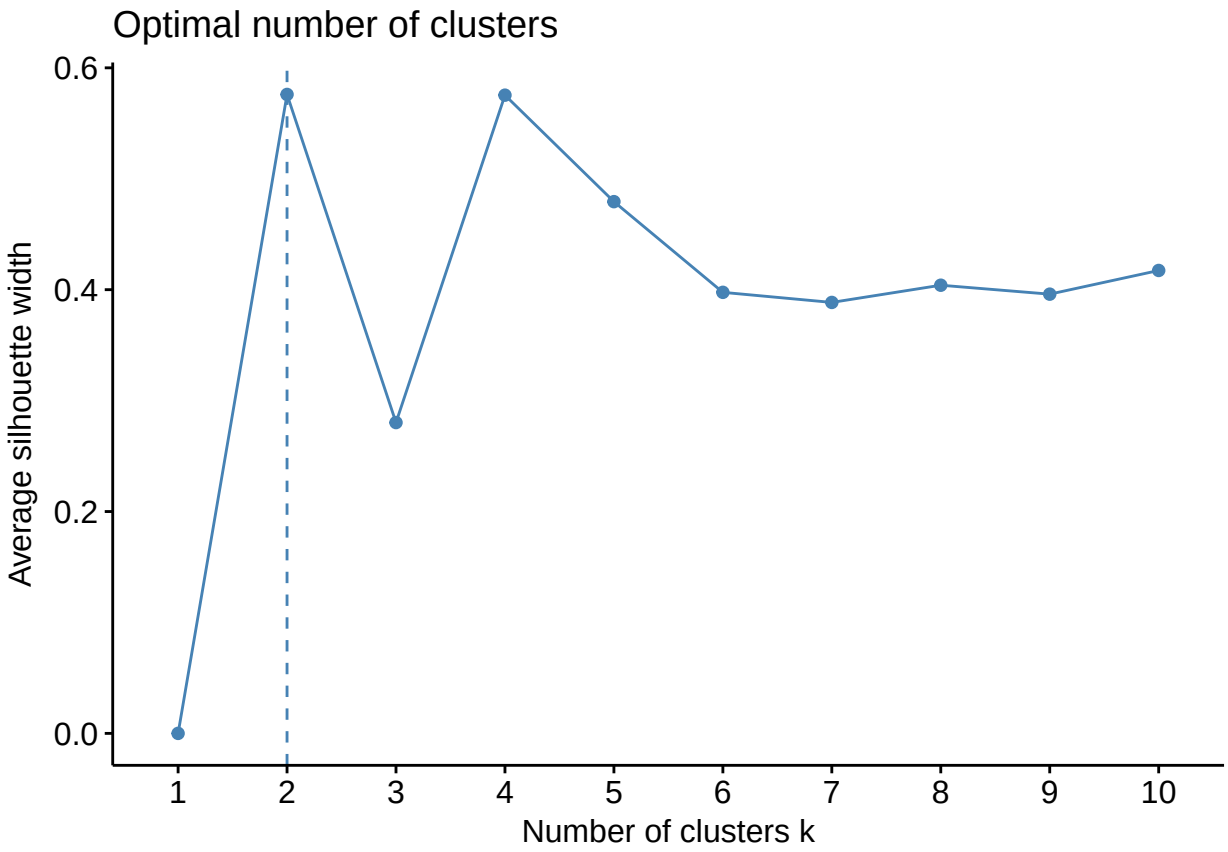

Cluster Dendrogram
Height
1
0

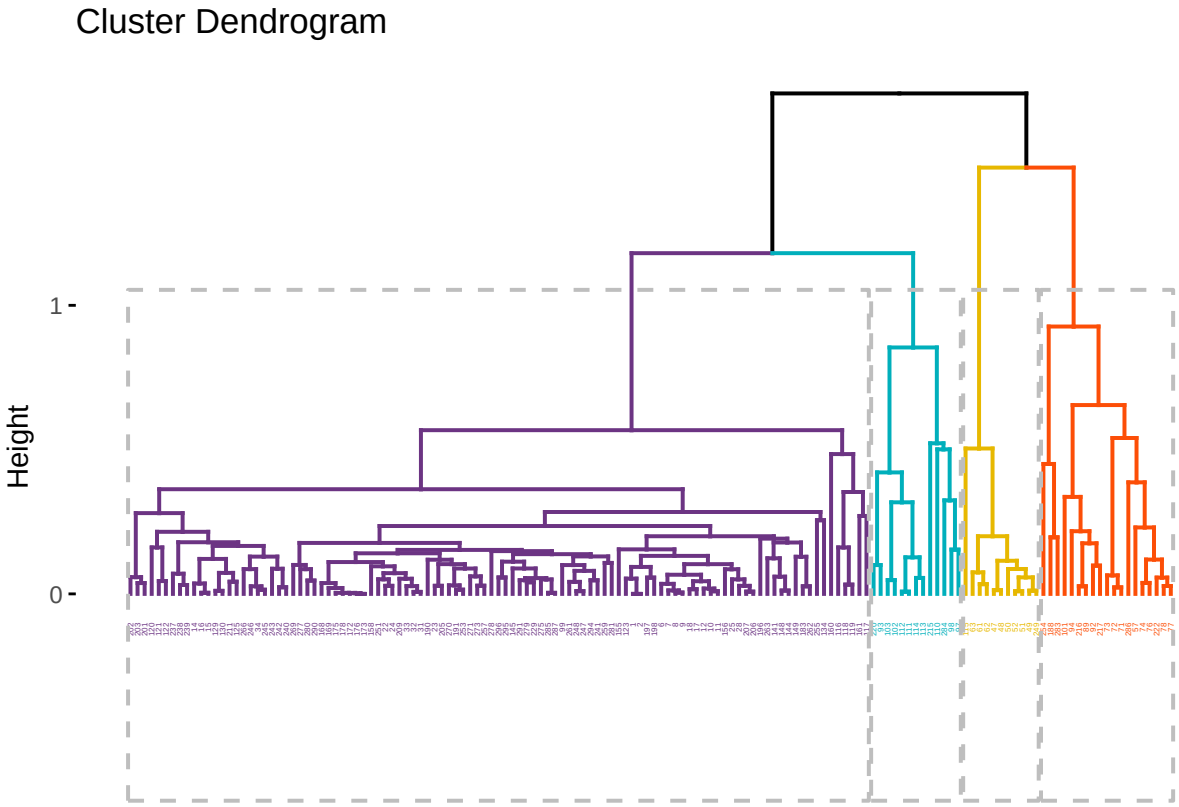


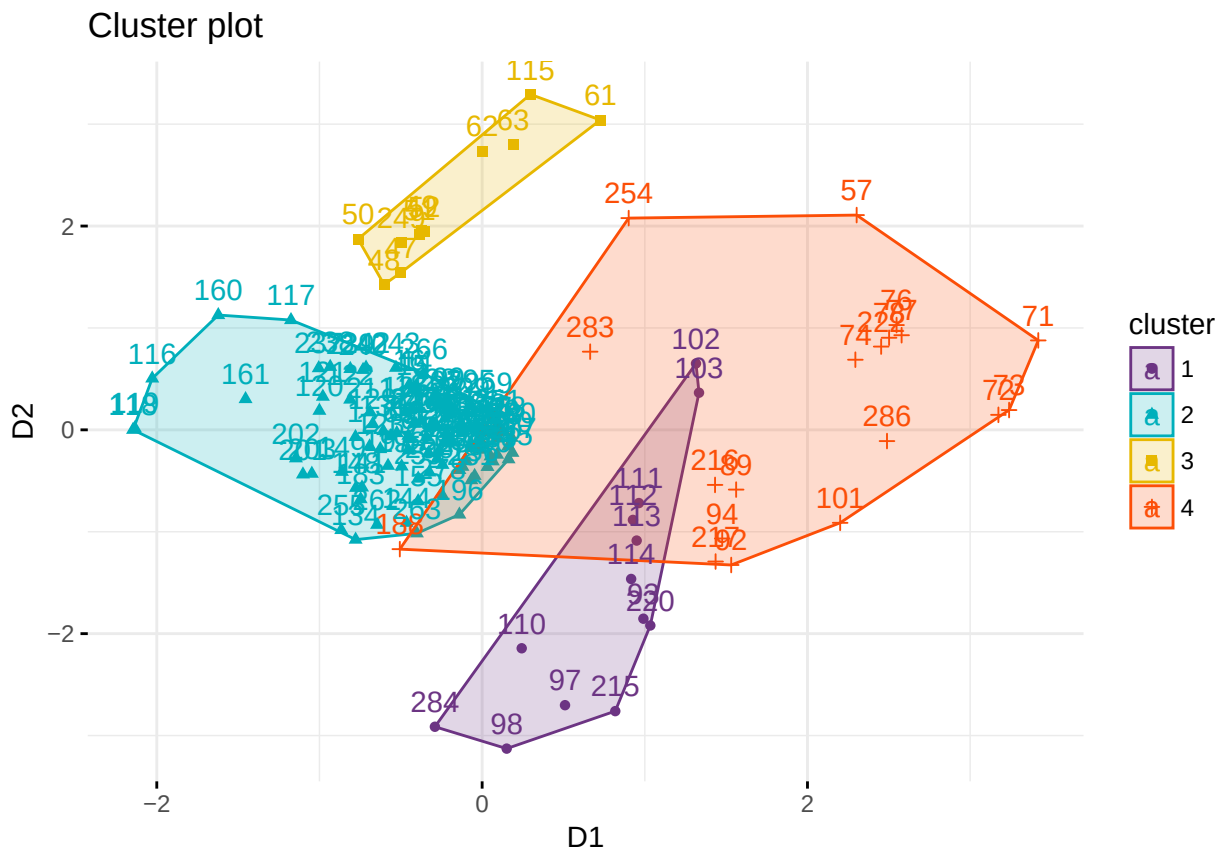

### Distacia basada en la Correlación Cruzada

$$d_{ccor}(X_t, Y_t) = \sqrt{\frac{(1 - CC(x_t, y_t, 0))^2}{\sum_k (1 - CC(x_t, y_t, k))^2}}$$

Donde $CC(x_t, y_t, k)$ es la función de correlación cruzada con $k$ retardos.


Optimal number of clusters
Average silhouette width
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
Number of clusters k


Cluster Dendrogram
Height
1
0




## Observacion 2

Por la naturaleza de los datos, puede que la métrica más adecuada para analizar la similitud entre las series de vazoes, sea la Distancia basada en la Correlación Cruzada, ya que asume que dos puntos son próximos (cercanos) si presentan una alta Correlación Cruzada, esto a su vez se traduce en la posible existencia de una relación de dependencia entre todo par de series que se encuentren cercanas.

## Bibliografía


- Pablo Montero & José A. Vilar (2014). TSclust: An R Package for Time Series Clustering.
- Borg, I., & Groenen, P. J. (2005). Modern multidimensional scaling: Theory and applications. Springer Science & Business Media.