ANÁLISIS CLÚSTER

ÍNDICE

ANÁLISIS CLUSTER (Manuel)

1. ANTECEDENTES

Esta es una una tabla de datos realizada por el profesor Ali S. Hadi en la que se estudia la concentración de nitrógeno y los porcentajes de terreno agrícola, forestal, comercial industrial y residencial de 20 ríos de Nueva York.

	CONCENTR	AGRICOLA	FORESTAL	COMIND	RESIDENCIAL
1	1,10	26,00	63,00	,29	1,20
2	1,01	29,00	57,00	,09	,70
3	1,90	54,00	26,00	,58	1,80
4	1,00	2,00	84,00	1,98	1,90
5	1,99	3,00	27,00	3,11	29,40
6	1,42	19,00	61,00	,56	3,40
7	2,04	16,00	60,00	1,11	5,60
8	1,65	40,00	43,00	,24	1,30
9	1,01	28,00	62,00	,15	1,10
10	1,21	26,00	60,00	,23	,90
11	1,33	26,00	53,00	,18	,90
12	,75	15,00	75,00	,16	,70
13	,73	6,00	84,00	,12	,50
14	,80	3,00	81,00	,35	,80
15	,76	2,00	89,00	,35	,70
16	,87	6,00	82,00	,15	,50
17	,80	22,00	70,00	,22	,90
18	,87	4,00	75,00	,18	,40
19	,66	21,00	56,00	,13	,50
20	1,25	40,00	49,00	,13	1,10

2. OBJETIVOS

El objetivo del análisis de clúster es explorar cómo se agrupan los ríos según sus similitudes, identificando posibles patrones o estructuras naturales en los datos. Para ello, se realizará un análisis exploratorio que permita determinar el número óptimo de clústers y evaluar la coherencia de las agrupaciones formadas.

3. VARIABLES EN ESTUDIO

*Hay **5 variables cuantitativas**:*

- *CONCENTR*: Concentración de nitrógeno en el agua
- *AGRICOLA*: Porcentaje de terreno agrícola
- *FORESTAL*: Porcentaje de terreno forestal
- *COMIND:* Porcentaje de terreno comercial industrial
- *RESIDENCIAL:* Porcentaje de terreno residencial

4. MODELO

4.1. Distancias

Dado que se trabaja con variables cuantitativas, se estandarizan. Se aplican principalmente distancias euclidianas.

4.2. Algoritmo Jerárquico

Se usa el método jerárquico aglomerativo, construyendo dendogramas, donde cada observación comienza como un clúster individual y se fusionan sucesivamente según distintos criterios de enlace.

4.3. Algoritmo K-Medias

Fijando el número de clústers, el algoritmo optimiza la asignación de individuos minimizando la distancia entre observaciones y centros de clúster, que son recalculados iterativamente hasta que la diferencia entre iteraciones sea menor que un umbral ϵ.

4.4. ANOVA

Se realiza un ANOVA para determinar si existen diferencias significativas entre las medias y significancias de las variables en los distintos clústeres generados por K-medias. El modelo es: $y_{ij} = \mu + \zeta_i + \epsilon_{ij}$

5. ESTUDIO PILOTO

No procede, debido a que supone que ya se ha realizado un estudio piloto previamente.

6. FORMA Y TAMAÑO DE LA MUESTRA

No procede debido a que no conocemos al autor de este estudio piloto y, por tanto, tampoco podemos saber la forma y el tamaño de la muestra que tomó.

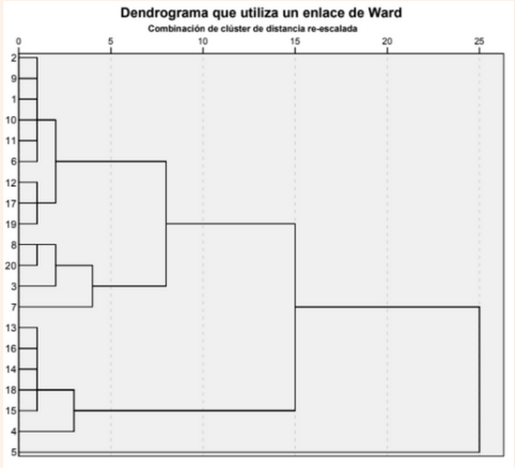
7. ANÁLISIS DE DATOS

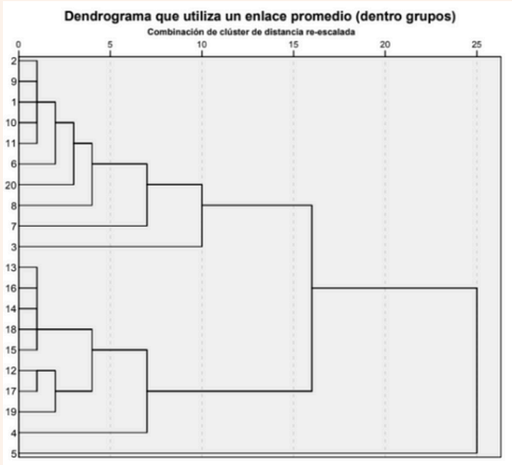
7.1 DENDOGRAMAS

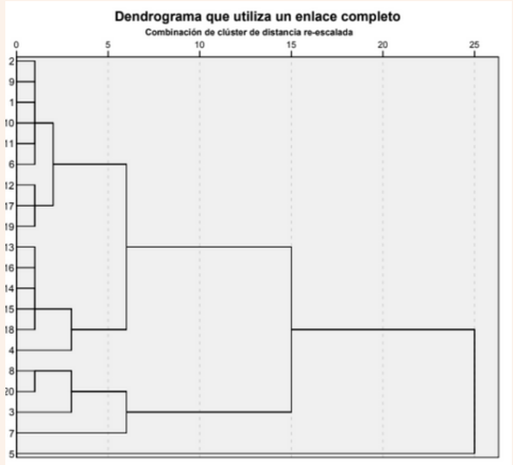
Ward, Centroid y C**omplete-linkage** muestran tres saltos grandes: sugieren 2-3 clusters coherentes antes del enlace final.

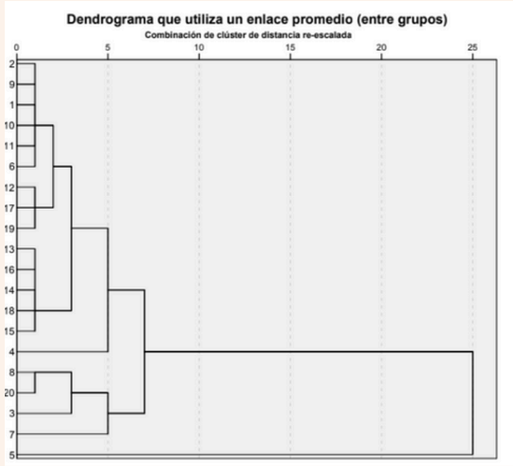
Average-linkage hace un corte muy marcado que separa 2 macrobloques.

Single-linkage encadena casi todos al principio, pero indica 2 clusters.









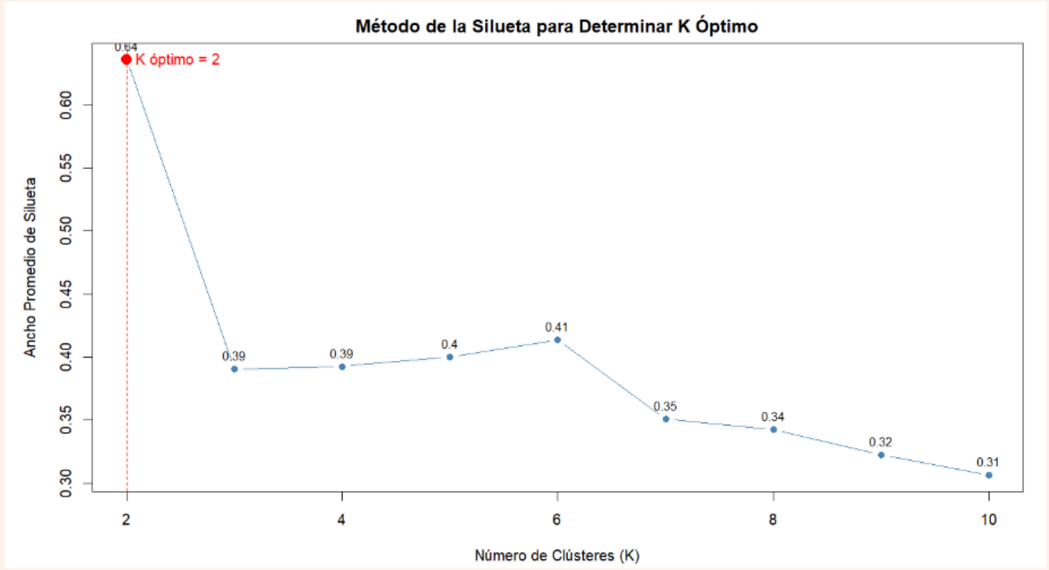
7.2. ELECCIÓN DE NÚMERO DE CLÚSTERS

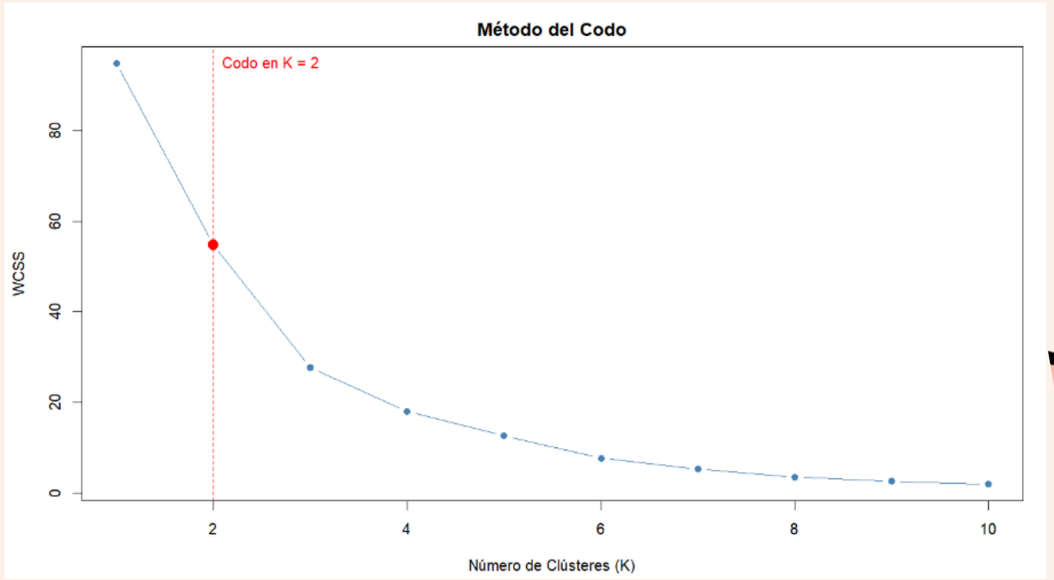
Utilizamos dos métodos para determinar el número de clústeres: el método del codo y el método del índice de silueta.

Silueta: Calculamos el índice de silueta, que mide qué tan bien definidos están los clústeres. El índice fue más alto para 2 clústeres.

Codo: Analizamos cómo cambia la inercia a medida que aumenta el número de clústeres. La gráfica mostró un "codo" en 3 clústeres, pero el algoritmo seleccionó 2 clústeres como la mejor opción.

Ambos métodos coinciden en que 2 clústeres es el número óptimo.





7.3. ALGORITMO DE K-MEANS

Realizamos un análisis **K-means** con dos clústeres para clasificar los ríos.

El resultado muestra un agrupamiento muy **desequilibrado**: el Clúster 1 contiene un solo caso, mientras que el Clúster 2 agrupa los 19 restantes.

El **caso aislado** (Clúster 1) presenta valores extremadamente altos en residencial e industrial y valores muy bajos en áreas agrícolas y forestales, lo que sugiere que representa una zona altamente urbanizada.

Centros de clústeres iniciales

	Clúster 1	Clúster 2
Puntuación Z(CONCENTR)	1,90657	-,97905
Puntuación Z(AGRICOLA)	-1,11333	-,90967
Puntuación Z(FORESTAL)	-2,00928	1,18539
Puntuación Z(COMIND)	3,43499	-,52362
Puntuación Z (RESIDENCIAL)	4,17096	-,34621

Centros de clústeres finales

	Clúster 1	Clúster 2
Puntuación Z(CONCENTR)	1,90657	-,10035
Puntuación Z(AGRICOLA)	-1,11333	,05860
Puntuación Z(FORESTAL)	-2,00928	,10575
Puntuación Z(COMIND)	3,43499	-,18079
Puntuación Z (RESIDENCIAL)	4,17096	-,21952

Número de casos en cada clúster

Clúster	1	1,000
	2	19,000
Válidos		20,000
Perdidos		,000

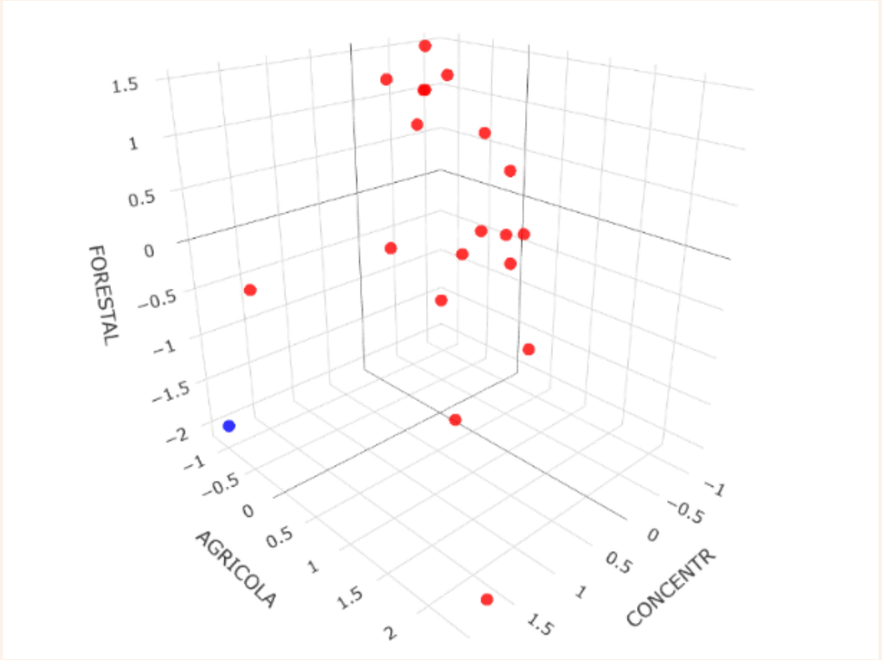
Historial de iteraciones[a]

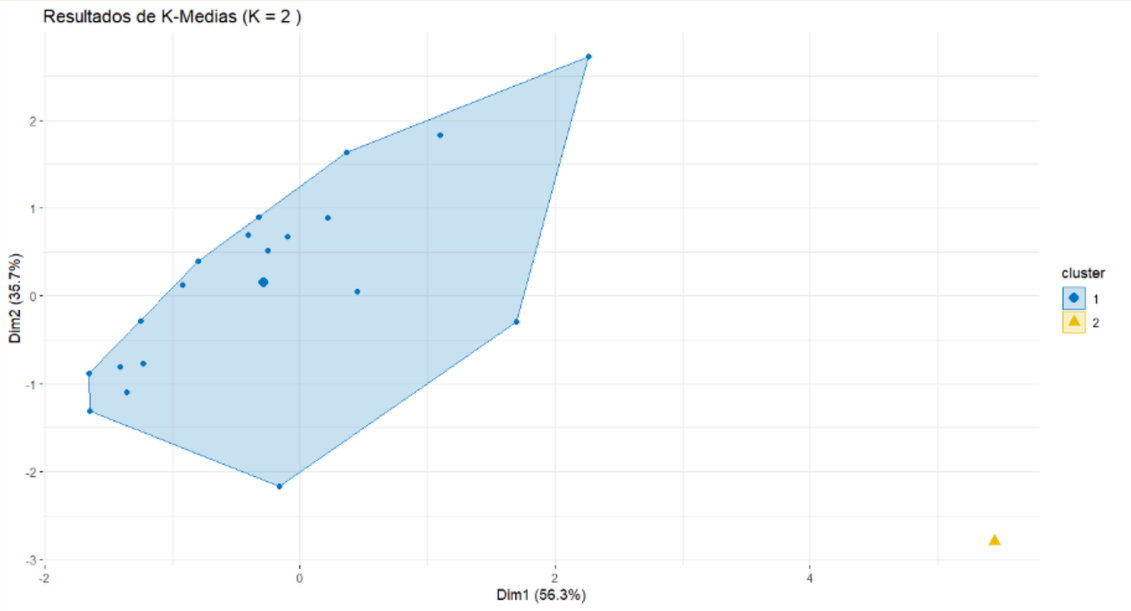
Iteración	Cambiar en centros de clústeres 1	2
1	,000	1,735
2	,000	,000

7.4. REPRESENTACIÓN VISUAL

Representamos visualmente los resultados del algoritmo de K-Medias:

Las gráficas confirman la separación: en el gráfico 3D, el caso atípico aparece aislado en azul, mientras el resto se agrupa en rojo. En el 2D también se ve aislado del resto en la esquina inferior derecha.





7.5. PERTENENCIA A CADA CLÚSTER

Como ya sabíamos, hay un clúster individual y un clúster que contiene a todos los otros ríos. Conseguimos con la tabla de la derecha, determinar que es el río 5 el que se diferencia del resto.

	ZCONCENTR	ZAGRICOLA	ZFORESTAL	ZCOMIND	ZRESIDENCIAL
1	-,13168	,44805	,00841	-,29855	-,23680
2	-,33780	,65171	-,32787	-,56334	-,31495
3	1,70045	2,34886	-2,06533	,08539	-,14302
4	-,36070	-1,18122	1,18539	1,93893	-,12739
5	1,90657	-1,11333	-2,00928	3,43499	4,17096
6	,60117	-,02715	-,10369	,05892	,10707
7	2,02108	-,23081	-,15973	,78709	,45094
8	1,12791	1,39845	-1,11253	-,36475	-,22117
9	-,33780	,58382	-,04764	-,48390	-,25243
10	,12023	,44805	-,15973	-,37799	-,28369
11	,39505	,44805	-,55206	-,44419	-,28369
12	-,93325	-,29870	,68097	-,47066	-,31495
13	-,97905	-,90967	1,18539	-,52362	-,34621
14	-,81874	-1,11333	1,01725	-,21911	-,29932
15	-,91034	-1,18122	1,46562	-,21911	-,31495

Clúster de pertenencia

Número del caso	Clúster	Distancia
1	2	,420
2	2	,867
3	2	3,644
4	2	2,697
5	1	,000
6	2	,841
7	2	2,458
8	2	2,196
9	2	,670
10	2	,560
11	2	,950
12	2	1,116
13	2	1,735
14	2	1,652
15	2	2,013
16	2	1,514
17	2	,816
18	2	1,397
19	2	1,203
20	2	1,667

7.6. ANOVA

Según los resultados del **ANOVA**, las variables **RESIDENCIAL** y **COMIND** son las que más diferencian entre grupos, con una significancia muy alta ($p < 0.001$) y **elevados valores F**, lo que indica una gran capacidad discriminativa. También FORESTAL y CONCENTR muestran diferencias significativas ($p < 0.05$), aunque en menor medida. En cambio, AGRICOLA no presenta diferencias estadísticamente significativas ($p = 0.264$), por lo que su poder discriminativo es limitado. Estos resultados sugieren priorizar las variables RESIDENCIAL y COMIND en futuros análisis.

ANOVA

	Clúster		Error			
	Media cuadrática	gl	Media cuadrática	gl	F	Sig.
Puntuación Z(CONCENTR)	3,826	1	,843	18	4,539	,047
Puntuación Z(AGRICOLA)	1,305	1	,983	18	1,327	,264
Puntuación Z(FORESTAL)	4,250	1	,819	18	5,186	,035
Puntuación Z(COMIND)	12,420	1	,366	18	33,977	<,001
Puntuación Z (RESIDENCIAL)	18,313	1	,038	18	479,503	<,001

8. CONCLUSIONES

Para agrupar los ríos, se realizó un análisis clúster combinando métodos jerárquicos y de partición. Los **dendrogramas** sugirieron la existencia de dos grupos, lo cual fue confirmado por el método del codo y el coeficiente de silueta, ambos indicando que k = 2 es el número óptimo de clústeres. Finalmente, se aplicó **k-medias** para refinar la clasificación.

El resultado del análisis con k-medias mostró que **un único río** quedó clasificado en un clúster independiente, mientras que los 19 ríos restantes se agruparon en el otro. El río aislado se caracteriza principalmente por su mayor proximidad a zonas **residenciales**, característica que mayormente le dista del resto.