Trabajo Fin de Máster

Revisión de métodos de clustering para datos funcionales

Rikelvi Felipe Fermín González

Máster en Técnicas Estadísticas

Curso 2023-2024

Propuesta de Trabajo Fin de Máster

Título en galego: Revisión dos métodos de clustering de datos funcionais
Título en español: Revisión de métodos de clustering para datos funcionales
English title: Review of clustering methods for functional data
Modalidad: Modalidad A
Autor: Rikelvi Felipe Fermín González, Universidad de Coruña
Directores: Manuel Oviedo de la Fuente, Universidad de Coruña; Manuel Febrero Bande, Universidad de Santiago de Compostela
Breve resumen del trabajo: Los datos funcionales surgen cuando una de las variables de interés en un conjunto de datos se puede ver de forma natural como una función. En el presente trabajo se propone una revisión sistemática de los métodos de clúster en datos funcionales, como el dendograma, el k-medias o el mean shift entre otros. Sin embargo, estos procedimientos requieren conocer o estimar parámetros relacionados a cada algoritmo, como fijar de antemano el número de agrupaciones. Por ello, este TFM plantea estudiar y/o adaptar los procedimientos de selección del número óptimo de clústeres u otros parámetros relacionados en datos multivariantes (como los basados en la silueta) a cuando se dispone de un conjunto (o varios conjuntos) de datos funcionales. Finalmente, se probará el rendimiento de estos procedimientos y los incluidos recientemente en la literatura en escenarios simulados y/o sobre datos reales.
Recomendaciones: Haber cursado o cursar la asignatura “Análisis de Datos Funcionales”, conocimiento de R.
Otras observaciones:

Don Manuel Oviedo de la Fuente, Profesor contratado interino de sustitución de Universidad de Coruña, y don Manuel Febrero Bande, Profesor Catedrático de Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Revisión de métodos de clustering para datos funcionales

fue realizado bajo su dirección por don Rikelvi Felipe Fermín González para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En línea, a 2 de febrero de 2024.

El director:
Don Manuel Oviedo de la Fuente

El director:
Don Manuel Febrero Bande

El autor:
Don Rikelvi Felipe Fermín González

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **el autor declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Quisiera comenzar expresando mi profundo agradecimiento a Jehová Dios por cada aspecto de mi vida. Seguidamente, deseo extender mi gratitud al Ministerio de Educación Superior, Ciencia y Tecnología de la República Dominicana por seleccionarme como becario del máster en Técnicas Estadísticas, impartido en las prestigiosas instituciones académicas gallegas: la Universidad de La Coruña, la Universidad de Santiago de Compostela y la Universidad de Vigo. Estoy enormemente agradecido con los distinguidos directores de mi trabajo final, los señores Manuel Oviedo y Manuel Febrero, por su comprensión, dedicación y paciencia inquebrantable. Por último, pero no menos importante, quiero expresar mi gratitud de manera especial a mi querida esposa, Yesenia, por su apoyo incondicional y su paciencia durante mi período de estudio, así como a mis padres, Carme y Benito, por haberme guiado por caminos de rectitud y bondad.

Índice general

Prefacio

El estudio del análisis de datos funcionales trata los datos de parámetros continuos, como curvas o supeficies, además extiende los métodos estadísticos multivariantes clásicos a escenarios donde se dispone de un gran número de dimensiones, por lo cual, en años recientes y junto al avance tecnológico ha ido creciendo el interés sobre el estudio de estos tipos de datos. En ese sentido, un tema de interés particular es el clustering a datos funcionales que consiste en agrupar los datos según patrones preexistentes en ellos, pero no conocidos. En (Jain et al., 1988) se recogen múltiples métodos de clustering aplicados a datos multivariantes, mientras que en (Ramsay et al., 2005) y (Ferraty et al., 2006) abordan algunos métodos de clustering en datos funcionales, en tanto que el objetivo de la presente memoria es revisar cinco de los métodos más extendidos de clustering a datos funcionales: método jerárquico de enlace mínimo, método jerárquico de enlace máximo, k-medias, cambio medio y DBSCAN. En el Capítulo 2, para llevar a cabo el clustering a datos funcionales se extenderán las herramientas multivariantes a configuraciones de dimensiones continuas en los métodos que permitan tales extensiones y se definirán las herramientas funcionales para los métodos que los requieran, como es el caso del clustering de cambio medio. Sin embargo, antes de realizar el agrupamiento de los datos funcionales, en el Capítulo 1 se introducirá cada método desde un punto de vista génerico. Los datos funcionales a agrupar consisten por un lado en tres simulaciones de datos funcionales, Capítulo 3, con varias subpoblaciones (o subconjuntos) de patrones diferenciados, y por otro lado, un conjunto de datos reales sobre la generación eléctrica fotovoltaica de España en 2022, Capítulo 4. Por último, en el Capítulo 5 se comentan los aspectos más relevantes de los resultados obtenidos.

Capítulo 1

Métodos de Clustering

1.1. Introducción a los Métodos de Clustering

El análisis de grupos o *clustering* es un conjunto de herramientas exploratorias de datos que ayuda a identificar patrones o estructuras dentro de una población o conjunto de elementos. Este consiste en la construcción de grupos de elementos homogéneos en función de las similitudes entre ellos, procurando al mismo tiempo que los grupos formados sean lo más heterogéneos posibles entre sí. Además de tener como objetivo el análisis de datos, otros propósitos de agrupar los datos son: detectar observaciones atípicas, reconocer patrones y formular hipótesis acerca de la composición de la población de origen.

La utilidad del clustering reside en ser una herramienta interdisciplinaria por lo cual hay numerosos ejemplos de su aplicación en diversos campos, *p. ej.*: resulta provechoso agrupar las zonas con grados de criminalidad similares al diseñar políticas de seguridad (Arango González et al., 2016). En campos de la medicina es valioso para clasificar enfermedades según el cuadro clínico de los pacientes (Webster et al., 2021). Un ejemplo clásico de su uso es la segmentación de mercado de consumidores de un bien o servicio económico, en este sentido su uso es pertinente en el mercado eléctrico para identificar patrones de consumo. En este capítulo se describen algunas de los métodos de clustering más utilizados y se reseñaran las ventajas y desventajas de cada uno de ellos.

1.1.1. Conceptos Generales

Antes de introducir los distintos métodos de clustering primero se deben definir los componentes que son necesarios en el clustering. Estos componentes abarcan los objetos sobre los cuales se realiza el clustering y la noción de distancia entre ellos.

Empezamos con las definiciones de los n objetos o elementos sobre el cual se realiza el clustering: $\mathcal{S}_n = \{X_i\}_{i=1}^n$, donde X_i es el i-ésimo elemento. A su vez, en la estadística clásica el clustering se realiza en función de una selección de las p variables o características del espacio de $\mathcal{S}_n$, por tanto, la observación X_{ij} hace referencia al i-ésimo elemento en la j-ésima características como muestra el Cuadro 1.2. Asimismo, existen otros espacios sobre los cuales se puede llevar a cabo el clustering como los espacio continuos que se verá en el Capítulo 2. Por simplicidad, nos referiremos a X de forma genérica como la variable a agrupar, salvo cuando queramos especificar el espacio en el cual se está

trabajando.

Cuadro 1.1: Conjunto de datos $\mathcal{S}_n$ en p-dimensiones.

Elementos	**Dimensiones**					
	1	***2***	$\cdots$	***j***	$\cdots$	***p***
1	X_{11}	X_{12}	$\cdots$	X_{1j}	$\cdots$	X_{1p}
2	X_{21}	X_{22}	$\cdots$	X_{2j}	$\cdots$	X_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	X_{i1}	X_{i2}	$\cdots$	X_{ij}	$\cdots$	X_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
n	X_{n1}	X_{n2}	$\cdots$	X_{nj}	$\cdots$	X_{np}

En segundo lugar definimos una partición de $\mathcal{S}_n$ como una división de $\mathcal{S}_n$ en k subconjuntos $\mathcal{C} = \{C_j\}_{j=1}^{k}$, tal que satisface las siguientes condiciones:

$$C_j \cap C_l = \emptyset \quad \forall\, 1 \leq j \neq l \leq k \tag{1.1}$$

y

$$\bigcup_{j=1}^{k} C_j = \mathcal{S}_n \tag{1.2}$$

es decir, los grupos que componen $\mathcal{C}$ son mutuamente excluyentes y forman un conjunto exhaustivo, por lo que cada elemento X_i pertenece a un, y sólo un, grupo C_j.

Además, los criterios que deben tenerse en cuenta al emplear algún método de clustering varían según la estrategia que estos implementen al formar grupos. Sin embargo, es común a todos los métodos la elección de una medida de distancia (métrica) d entre los objetos de $\mathcal{S}_n$, donde d cumpla las siguientes propiedades:

1. **No negatividad**: $d(X_i, X_j) \geq 0$;
2. **Indiscernibilidad**[1]: $d(X_i, X_j) = 0 \iff X_i = X_j$;
3. **Simetría**: $d(X_i, X_j) = d(X_j, X_i)$;
4. **Desigualdad triangular**: $d(X_i, X_l) = d(X_i, X_j) + d(X_j, X_l)$.

De igual forma, se pide que la distancia entre los objetos tenga una estructura de espacio vectorial para poder realizar operaciones algebraicas sobre ellos.

Por último, un objeto matemático muy útil con base en la distancia, pero que no es genérico en el clustering, es la matriz de distancias $\mathcal{D}$ entre los objetos de $\mathcal{S}_n$ tomados en pares:

$$\mathcal{D} = \{d_{ij} = d(X_i, X_j)\} \quad \forall\, 1 \leq i, j \leq n, \tag{1.3}$$

por tanto, $\mathcal{D}$ es una matriz de dimensión $n \times n$, las entradas de la diagonal principal son iguales a cero y $d_{ij} = d_{ji}$ para cualquier par (i, j). En el Cuadro 1.2 se ilustra la forma genérica de $\mathcal{D}$.

Estos conceptos sirven de base para los métodos de clustering que se mostrarán en el presente estudio, por tanto, se ampliará más sobre ellos en las siguientes secciones.

[1] En los espacios pseudométricos no se cumple la propiedad de indiscernibilidad.

Cuadro 1.2: Matriz de distancias $\mathcal{D}$ genérica.

	X_1	X_2	$\cdots$	X_j	$\cdots$	X_n
X_1	d_{11}	d_{12}	$\cdots$	d_{1j}	$\cdots$	d_{1n}
X_2	d_{21}	d_{22}	$\cdots$	d_{2j}	$\cdots$	d_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
X_i	d_{i1}	d_{i2}	$\cdots$	d_{ij}	$\cdots$	d_{in}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
X_n	d_{n1}	d_{n2}	$\cdots$	d_{nj}	$\cdots$	d_{nn}

1.2. Métodos de Clustering

El clustering se puede llevar a cabo a través de una amplia variedad de métodos que se diferencian tanto en sus objetivos como en los procedimientos que emplean. El objetivo común a todos ellos es el propósito de formar grupos que *a priori* no están definidos. La clasificación más habitual de los métodos de clustering consiste en tres categorías: métodos jerárquicos, métodos de particionamiento y métodos basados en densidad. Como veremos más adelante, en los primeros los grupos se crean a partir del anidamiento de subgrupos, los segundos buscan optimizar algún criterio de distancia y en los terceros los elementos se clasifican de acuerdo con la densidad de la zona en la cual se encuentran. Los métodos de clustering que se utilizan en el presente trabajo son: métodos jerárquicos de enlace mínimo y de enlace máximo, el k-medias, el cambio medio y el DBSCAN. En la siguiente sección se detallará cada método y se ofrece un algoritmo para obtener los grupos.

1.2.1. Métodos Jerárquicos

De acuerdo a (Hastie, Tibshirani et al., 2001), los métodos jerárquicos consisten en construir un árbol de varios niveles a partir de una matriz de distancias $\mathcal{D}$ de los elementos de $\mathcal{S}_n$. En los niveles más bajos se encuentran los grupos de elementos muy próximos o semejantes, mientras que en los niveles superiores se crean grupos como resultado de la agregación de los grupos en niveles inferiores. El árbol formado se denomina dendrograma, el cual es una representación gráfica o diagrama de datos en forma de árbol invertido en el cual los elementos se organizan en subcategorías que se van dividiendo en otros hasta llegar al nivel de detalle deseado. En la Figura 1.1 se muestran dos ejemplos de dendrogramas con datos de distancia de 21 ciudades europeas obtenidas del paquete `dataset` de (R Core Team, 2022).

El dendrograma se interpreta según la distancia en el eje de las ordenadas, y en consecuencia, a medida que se va descendiendo en el árbol, más cercanos son los elementos de $\mathcal{S}_n$. En ese sentido, los grupos se crean al fijar un valor para la distancia (eje de las ordenadas) y trazando una recta horizontal que corte el dendrograma en tal punto, *p. ej.*: en la Figura 1.1 se fijan los valores de altura en 239 km y 2615 km para los dendrogramas de enlace mínimo (izquierda) y de enlace máximo (derecha), respectivamente, y se traza una línea (azul) de corte en ambos dendrogramas, en el primero se crean 17 grupos y en el segundo se crean 4 grupos. Más adelante, se exponen algunas guías sobre la elección del número de grupos k.

En cuanto a la clasificación de los métodos jerárquicos, estos se subdividen en algoritmos aglomerativos

y algoritmos divisivos, según el enfoque que se tome para crear los grupos. Los primeros parten desde los elementos para ir construyendo grupos, con los cuales a su vez se van formando otros grupos. En cambio, los algoritmos divisivos parten del grupo formado por todos los elementos del conjunto para ir formando grupos mediante la división sucesiva de los subgrupos que se van formando. En la práctica los algoritmos aglomerativos se emplean con mayor frecuencia que los algoritmos divisivos y por tanto serán los implementados.

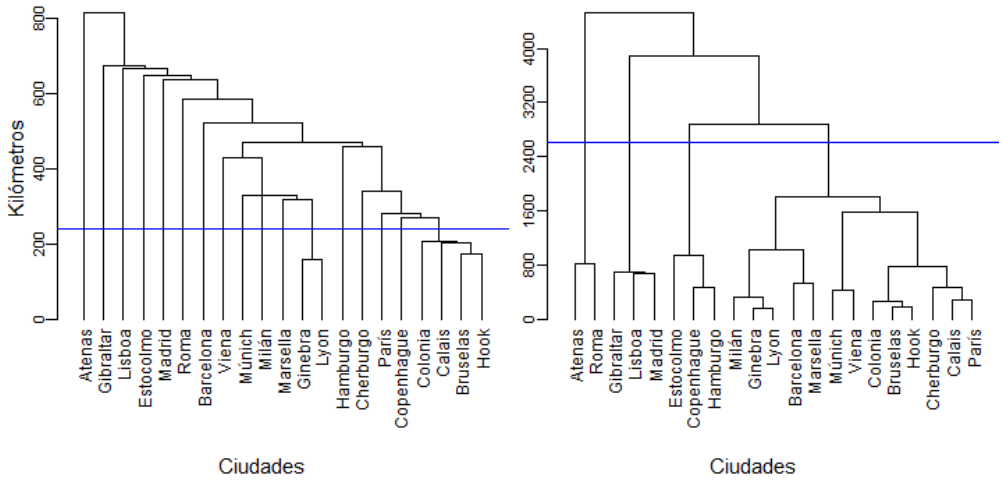

Figura 1.1: Ilustración de dendrogramas con la distancia de 21 ciudades europeas. `Izquierda`: dendrograma de enlace mínimo. `Derecha`: dendrograma de enlace máximo.

Para completar la definición de los métodos jerárquicos hace falta definir el método de enlace entre los subgrupos del dendrograma. Existen múltiples métodos de enlace que en su mayoría pueden implementarse con la fórmula de disimilitud de (Lance et al., 1967):

Definición 1.1 Método de enlace. Dados tres grupos C_i, C_j y C_l definimos el enlace entre la unión de C_i y C_j con C_l como la fórmula de disimilitud de Lance-Williams:

$$d(C_i \cup C_j, C_l) = \alpha_i d(C_i, C_l) + \alpha_j d(C_j, C_l) + \beta d(C_i, C_j) + \gamma |d(C_i, C_l) - d(C_j, C_l)| \tag{1.4}$$

donde los coeficientes α, β y γ son números reales.

Los coeficientes de la Ecuación 1.4 para distintos métodos de enlace se muestran en el Cuadro 1.3.

La elección del método de enlace tiene una incidencia notable en la forma en que se agrupan los datos, *p. ej.*: auxiliándonos de la Figura 1.1 podemos observar que el enlace mínimo tiende a crear dendrogramas con grupos encadenados o de racimos alargados, mientras que el enlace máximo tiende crear dendrogramas más uniformes con forma de pirámide. En el presente trabajo nos centraremos en los enlaces más comunes: mínimo y máximo, por ser los métodos de clustering jerárquico más utilizados y extendidos de la literatura.

Cuadro 1.3: Coeficientes de los métodos de enlace de la la fórmula de disimilitud de Lance-Williams.

Método de enlace	$\boldsymbol{\alpha}_i$	$\boldsymbol{\alpha}_j$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$
Mínimo	1/2	1/2	0	-1/2
Máximo	1/2	1/2	0	1/2
Promedio	$\frac{\lvert C_i\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$\frac{\lvert C_j\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	0	0
Promedio ponderado	1/2	1/2	0	0
Centroide	$\frac{\lvert C_i\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$\frac{\lvert C_j\rvert}{\lvert C_i\rvert+\lvert C_j\rvert}$	$-\frac{\lvert C_i\rvert\lvert C_j\rvert}{(\lvert C_i\rvert+\lvert C_j\rvert)^2}$	0
Centroide ponderado	1/2	1/2	-1/4	0
Varianza mínima	$\frac{\lvert C_i\rvert+\lvert C_l\rvert}{\lvert C_i\rvert+\lvert C_j\rvert+\lvert C_l\rvert}$	$\frac{\lvert C_j\rvert+\lvert C_l\rvert}{\lvert C_i\rvert+\lvert C_j\rvert+\lvert C_l\rvert}$	$-\frac{\lvert C_l\rvert}{\lvert C_i\rvert+\lvert C_j\rvert+\lvert C_l\rvert}$	0

El algoritmo propuesto para crear grupos jerárquicos aglomerativos basado en el método SAHN (Dubes, 1993) es el siguiente:

Algoritmo jerárquico aglomerativo

1. Se inicia tomando cada $X^{(0)}$ de $\mathcal{S}_n$ como un grupo $C^{(0)}$, por tanto se inicia con $k = n$ grupos $\mathcal{C}^{(0)} = \{C_j^{(0)}\}_{j=1}^{k}$, y luego se calcula la matriz de distancias $\mathcal{D}^{(0)}$ entre todos los grupos.

2. Se encuentra el par de grupos más cercano en la matriz de distancias $\mathcal{D}^{(t)}$:

$$d(C_i^{(t)}, C_j^{(t)}) = \text{mín}\{d_{ij}^{(t)}\}, \quad \forall 1 \leq i \neq j \leq k. \tag{1.5}$$

3. Se unen los grupos $C_i^{(t)}$ y $C_j^{(t)}$ del paso 2 formando un nuevo grupo $C_{ij}^{(t)}$:

$$C_{ij}^{(t)} = (C_i^{(t)} \cup C_j^{(t)}). \tag{1.6}$$

4. Se actualiza $\mathcal{D}^{(t)}$ eliminando las filas y columnas correspondientes a los grupos $C_i^{(t)}$ y $C_j^{(t)}$ y se agrega una fila y una columna con los valores del nuevo grupo $C_{ij}^{(t)}$, y en consecuencia el número de grupos se reduce en $k^{(t+1)} = k^{(t)} - 1$. La distancia entre un grupo $C_j^{(t)}$ y el grupo $C_{ij}^{(t)}$ se calcula mediante la Ecuación 1.4 para algún método de enlace en el Cuadro 1.3.

5. Si el número de grupos es uno $k^{(t+1)} = 1$, el algoritmo finaliza. En caso contrario, se vuelve al paso 2.

Elección de k Los métodos jerárquicos no requieren determinar la cantidad de grupos k de manera previa a su realización, por el contrario, se debe determinar k una vez realizado el clustering. En este sentido, para determinar k se suele cortar el dendrograma en los grupos donde se produce la mayor separación en distancia (altura), es así que en la Figura 1.1 se crean 17 y 4 grupos para los métodos de enlace mínimo y máximo, respectivamente. Sin embargo, elegir k según el mayor salto puede producir resultados imprácticos como se ve en el método de enlace mínimo al conjunto de datos `eurodist` ya que produce 17 grupos de los cuales 15 son grupos de un único elemento. Otro enfoque para elegir k consiste en seleccionar el valor de k donde se produzca la mayor caída relativa de la suma total de cuadrados intragrupo, este tema se desarrollará en detalle en la siguiente Subsección 1.2.2. Por último,

se puede elegir k de manera arbitraria o en base a información previa sobre la cantidad de grupos en el conjunto de datos observados.

Ejemplos en $\mathbb{R}^2$ Con el fin de facilitar la compresión de las ventajas y las desventajas de los métodos jerárquicos aglomerativos vistos nos auxiliaremos de cuatro conjuntos de datos en $\mathbb{R}^2$ debido a que se hace más intuitivo la detección de patrones al ojo humano en dos dimensiones que en dimensiones mayores. Los cuatro conjuntos de datos se muestran en la Figura 1.2 y se describen a continuación:

1. Mixtura de tres normales bivariadas con matrices de covarianzas homogéneas, pero con distintos centroides.
2. Mixtura de tres normales bivariadas con matrices de covarianzas heterogéneas y con distintos centroides.
3. Dos anillos circulares concéntricos, donde el radio del anillo interno es la mitad del radio del anillo externo.
4. Dos anillos circulares secantes con el mismo radio.

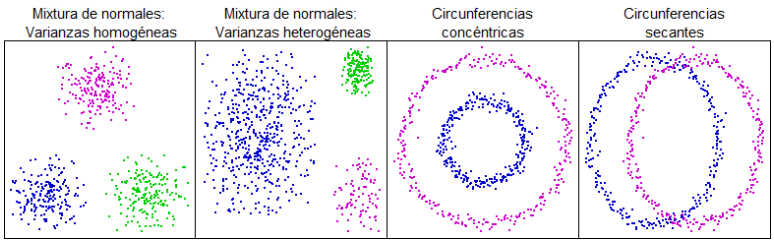

Figura 1.2: Conjuntos de datos en $\mathbb{R}^2$.

Cada color en la Figura 1.2 representa un grupo C o etiqueta de X, con esto se pretende evaluar que "tan bien" el clustering jerárquico clasifica los elementos de $\mathcal{S}_n$.

Empezamos con la Figura 1.3 que muestra los resultados del clustering jerárquico de enlace mínimo. En la primera fila se muestran los dendrogramas, todos con forma de racimos: algo distintivo de este método de enlace. Mientras que en la segunda fila se muestra la clasificación a los distintos conjuntos, se observa que este método enlace trabaja bien tanto con patrones de datos convexos (mixturas de normales) como con patrones de datos no convexos (anillos concéntricos). En los tres primeros conjuntos se logra clasificar los elementos de cada conjunto en el grupo esperado, esto se aprecia en los dendrogramas. Sin embargo, los patrones tienen que estar delimitados y separados para poder asignarlos a grupos distintos lo cual no es el caso de los anillos secantes, donde el enlace mínimo no logra distinguir dos grupos.

Por otro lado, en la Figura 1.4 se muestra el clustering jerárquico de enlace máximo a los cuatro conjuntos de datos en $\mathbb{R}^2$. Se observa que este método trabaja bien con patrones de datos convexos homogéneos que es el caso de las mixturas de tres normales de igual dispersión, donde el método clasificó los elementos en los grupos correspondientes. Por el contrario, si los patrones son convexos y heterogéneos, el método de enlace tendrá un comportamiento inesperado, lo cual es el caso de las mixturas de normales heterogéneas. Asimismo, cuando los patrones de datos son no convexos el enlace máximo no tendrá los resultados esperados, lo cual es el caso de los conjuntos de datos de los anillos circulares. A pesar de las aparentes limitaciones del enlace máximo frente al enlace mínimo, el primero

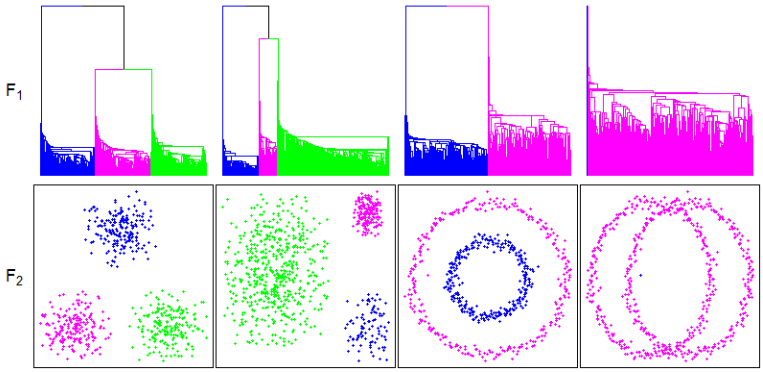

Figura 1.3: Ilustración del clustering jerárquico aglomerativo de enlace mínimo en $\mathbb{R}^2$.

tiende a crear dendrogramas más uniformes y entendibles y es un método menos sensible al ruido, es decir, no se ve igualmente afectado que el enlace mínimo por la proximidad entre los patrones.

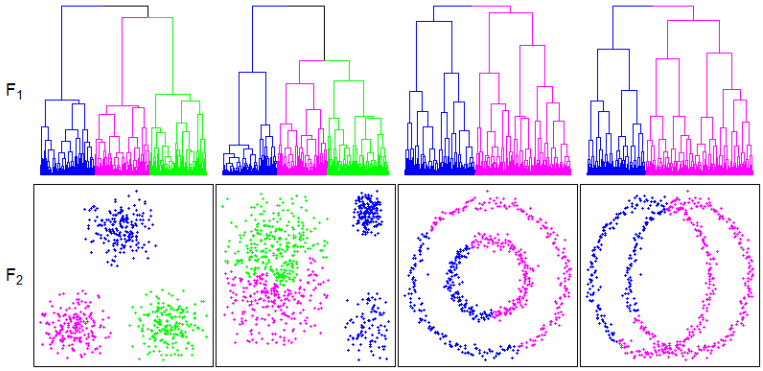

Figura 1.4: Ilustración del clustering jerárquico aglomerativo de enlace máximo en $\mathbb{R}^2$.

Por último, los valores de k donde se producen las mayores separaciones en los dendrogramas de enlace mínimo de los cuatro conjuntos son 2, 3, 3 y 193, respectivamente. Asimismo, los valores de k donde se producen las mayores separaciones en los dendrogramas de enlace máximo de los cuatro conjuntos son 3, 6, 8 y 7, respectivamente. Cabe resaltar que estos resultados de k van acorden con el desempeño de los métodos.

1.2.2. k-Medias

Cuando se va a realizar el particionamiento de un conjunto $\mathcal{S}_n$ es intuitivo agrupar los elementos en subconjuntos de elementos que estén próximos entre sí o que graviten alrededor de un centroide (valor medio). En este sentido, el k-medias es un clustering de particionamiento que consiste en formar grupos minimizando la suma de las distancias cuadráticas entre cada punto X_i al centroide del grupo C_j más cercanos, también llamada suma de cuadrados intragrupo. El k-medias fue propuesto por (MacQueen, 1967) y desde entonces ha sido uno de los métodos de clustering más usados por su eficiencia computacional y sencillez.

Específicamente, el objetivo del k-medias es formar una partición de $\mathcal{S}_n$ que conste k grupos $\mathcal{C} =$

$\{C_j\}_{j=1}^k$ alrededor de k centroides $\{m_j\}_{j=1}^k$ mediante la solución del problema de optimización siguiente:

$$\min_{\{m_j\}_{j=1}^k} \sum_{j=1}^{k} \sum_{X_i \in C_j} d(X_i, m_j)^2 \tag{1.7}$$

sujeto a

$$|C_j| \geq \delta, \quad \forall\, j = 1, \ldots, k \tag{1.8}$$

donde la Restricción 1.8 es opcional e indica la cantidad mínima δ de elementos X_i que debe tener cada grupo C_j.

Aunque el planteamiento del problema es sencillo, encontrar la combinación de $\mathcal{C}$ que haga mínima la Ecuación 1.7 es computacionalmente costoso por la cantidad de combinaciones posibles, *p. ej.*: una partición de $k = 5$ grupos de una muestra de tamaño $n = 20$ obtenemos 749,206,090,500 combinaciones posibles[2]. Entonces, evaluar todas las posibles combinaciones no es computacionalmente viable.

Para evitar evaluar cifras inabordables de combinaciones se estable un criterio de parada al fijar una cota ϵ que finalice el proceso una vez alcanzada o superada. Sin embargo, se debe tener en cuenta que al fijar un criterio de parada no se asegura encontrar el mínimo global de la Ecuación 1.7, y en su defecto converja a un mínimo local. Sin embargo, (Selim et al., 1984) probaron la convergencia del k-medias y de acuerdo a (Jain et al., 1988) el k-medias converge rápidamente con pocas iteraciones.

El algoritmo k-medias se basa en dos pasos fundamentales: asignación de los elementos y actualización de los centroides. A modo de resumen, los elementos X_i se asignan a los centroides más cercanos en una distancia d y luego se recalculan los valores de los centroides mediante un promedio. El algoritmo itera recursivamente estos dos pasos hasta que los centroides no cambian de posición o la diferencia de cada centroide consigo mismo en la iteración previa no es mayor a ϵ. Además, se puede fijar una cantidad máxima de iteraciones, T, para evitar oscilaciones. Entonces, el proceso del algoritmo genérico del k-medias se detalla a continuación:

Algoritmo k-medias

1. Se eligen k elementos en el espacio de $\mathcal{S}_n$ de forma aleatoria como los centroides iniciales $\{m_j^{(0)}\}_{j=1}^k$ y se establece un valor para ϵ. De manera opcional, se puede establecer un tamaño mínimo de grupo δ o un número máximo de iteraciones, T.

2. Se asigna cada elemento al grupo con el centroide más cercano:

$$C_j^{(t)} = \left\{X_i : d\big(X_i, m_j^{(t)}\big)^2 \leq d\big(X_i, m_l^{(t)}\big)^2, \quad \forall\, 1 \leq l \neq j \leq k\right\} \tag{1.9}$$

   donde X_i es asignado a exactamente un grupo. En caso de que X_i se encuentre a la misma distancia de dos o más centroides, se puede adoptar un criterio arbitrario de asignación.

3. Se actualiza la posición de los centroides mediante el promedio:

$$m_j^{(t+1)} = \frac{1}{\left|C_j^{(t)}\right|} \sum_{X_i \in C_j^{(t)}} X_i, \quad \forall\, j = 1, \ldots, k. \tag{1.10}$$

[2]Se utilizó la fórmula de los números de Stirling de segunda especie: $\frac{1}{k!}\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}j^n$.

4. **(Opcional)** Si el tamaño de al menos uno de los grupos es menor al mínimo establecido $\left|C_j\right| < \delta,\ \exists\, j = 1, \ldots, k$, se reinicia el algoritmo, pero se eligen centroides iniciales distintos a los usados.

5. Si la diferencia entre cada centroide con respecto a él mismo en la iteración previa es menor a la cota fijada $d\big(m_j^{(t+1)}, m_j^{(t)}\big) \leq \epsilon,\ \forall\, j = 1, \ldots, k$, u opcionalmente si se alcanzó el máximo de iteraciones $t = T$, entonces se finaliza el algoritmo. En caso contrario, se regresa al paso 2.

Número óptimo de k Luego de ver el algoritmo del k-medias queda por determinar el número óptimo de grupos, lo cual es muy difícil de saber en muchos casos donde no se dispone de información sobre la población. Para solventar este inconveniente lo más usual es realizar un gráfico de sedimentación que consiste en mostrar el progreso de la suma total de cuadrados intragrupo según se va incrementando el número de grupos k. El criterio de selección de k es muy sencillo: se selecciona el valor de k donde se produzca la mayor caída relativa de la suma total de cuadrados intragrupo. El criterio se basa en la ganancia marginal de aumentar k en 1, y a medida que vayamos aumentando k hasta hacerlo igual al tamaño del conjunto n, la suma del error irá disminuyendo, pero cada vez a una taza menor. Este criterio es conocido como el criterio del codo. El procedimiento se ilustra con los datos de distancia de 21 ciudades europeas del paquete `dataset` de (R Core Team, 2022) usados en la Subsección 1.2.1 y se muestra en la Figura 1.5 donde la mayor caída de la suma total de cuadrados intragrupo se produce en $k = 3$. La clasificación de las ciudades en cuanto a la distancia a las demás son: `lejanas` (magentas), `intermedias` (verdes) y `cercanas` (azules).

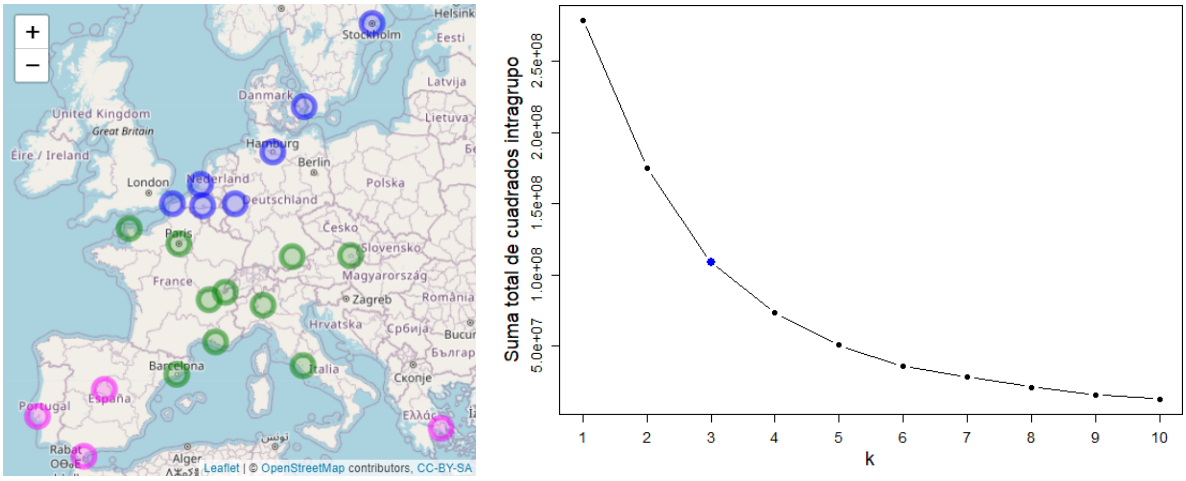

Figura 1.5: Ilustración de la sedimentación de k. `Derecha`: Grupos de ciudades europeas. `Izquierda`: Sedimentación de la suma total de cuadrados intragrupo según el tamaño de k.

Ejemplos en $\mathbb{R}^2$ Para matizar algunas particularidades del k-medias nos auxiliamos de los cuatro conjuntos de datos en $\mathbb{R}^2$ mostrados en la Subsección 1.2.1. Los resultados del clustering de k-medias a los conjuntos de datos bivariados se muestran en la Figura 1.6 donde cada fila de gráficos es un escenario distinto a los demás. En la primera fila se fijó k en cada conjunto de datos según el número esperado de grupos, se observa que el método trabaja bien con patrones de datos convexos (mixturas de normales), en cambio, el método no realiza las asignaciones esperadas en los conjuntos de datos no convexos, sin embargo, crea dos grupos regulares al dividir los datos en dos medialunas. Luego, en la segunda fila se fijó k en cada conjunto de datos según el número esperado de grupos, pero se varió

el conjunto de centroides iniciales con la intención de mostrar el impacto de la selección inicial de los centroides sobre los resultados que distan de los obtenidos en la fila 1. Por otro lado, en la tercera fila se muestran los resultados del clustering fijando k en cada conjunto de datos una unidad mayor al número esperado de grupos, donde la variación de k tiene un impacto notable en los resultados. Por último, se resalta que el k-medias crea fronteras lineales entre los grupos, como se observa en todas las clasificaciones de la Figura 1.6 y por tal razón el método no crea grupos de forma no convexas.

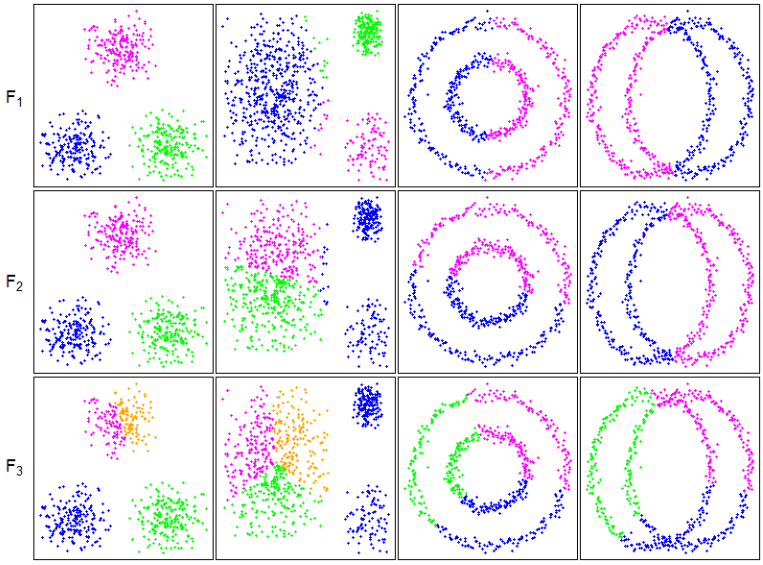

Figura 1.6: Ilustración del clustering k-medias en $\mathbb{R}^2$.

En cuanto a la selección del k óptimo se estimaron los siguientes valores donde la suma total de cuadrados intragrupo tuvo la mayor caída relativa para los cuatro conjuntos de datos son 3, 2, 3 y 3, respectivamente. Si bien los resultados son distintos a los esperados para los conjuntos 2, 3 y 4, las estimaciones de k no distan mucho de la esperadas.

1.2.3. Cambio Medio

El clustering de cambio medio (*mean-shift*) es un método de agrupamiento no paramétrico basado en la densidad kernel (núcleo) de los datos propuesto por (Fukunaga et al., 1975). Este método consiste en asignar cada elemento a los grupos mediante la escala de su gradiente hasta llegar a un máximo local de densidad. Específicamente, para asignar un elemento X_i a un grupo C_j se debe calcular su cambio medio $m(X_i)$, el cual consiste en estimar la densidad de X_i en su vecindario a partir de una función $\mathcal{K}'$, derivada de una función kernel $\mathcal{K}$.

El parámetro fundamental del clustering de cambio medio es h, conocido como parámetro de suavizado o ventana. Este determina el vecindario de un elemento X_i como todos los elementos X_j que se encuentran a una distancia menor o igual a h:

$$N(X_i) = \{X_j : d(X_i, X_j) \leq h, \ \forall 1 \leq i, j \leq n\} \tag{1.11}$$

La ventana h regula la curtosis del vecindario por lo que valores altos de h homogenizan los datos ocultando modas y valores pequeños de h crean patrones espurios por la información altamente localizada, por tanto, se debe buscar un tamaño de ventana que compense ambos escenarios. Mientras

que el kernel $\mathcal{K}$ es una función de pesos que determina la forma alrededor de un elemento X_i, por tanto, se debe elegir una función kernel $\mathcal{K}(X_i)$ que sea positiva, simétrica e integre uno para que el estimador de la densidad herede buenas propiedades analíticas, *p. ej.*: un kernel ampliamente usado que cumple estas condiciones es el gaussiano, mostrado en el Cuadro 1.4 junto a otros kernels típicos. Asimismo, en la parte izquiera de la Figura 1.7 se muestran los cuatro kernels del Cuadro 1.4 y en la parte derecha se muestran tres escenario del tamaño de ventana h: pequeño, moderado y grande, sobre una distribución normal bimodal de distinto centro y dispersión.

Cuadro 1.4: Ejemplos de algunas funciones kernels.

Kernel	$\mathcal{K}(X)$	$\mathcal{K}'(X)$
Gaussiano	$\frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{d(X_i,\, X)^2}{2h^2}\right)$	$-\frac{d(X_i,\, X)}{h^3\sqrt{2\pi}} \exp\left(-\frac{d(X_i,\, X)^2}{2h^2}\right)$
Epanechnikov	$\frac{3}{4} d\big(h,\, X^2\big)\ \mathbb{1}(\lvert X\rvert \leq h)$	$-\frac{3}{2} X\ \mathbb{1}(\lvert X\rvert \leq h)$
Triangular	$d(h,\, \lvert X\rvert)\ \mathbb{1}(\lvert X\rvert \leq h)$	$-\frac{X}{\lvert X\rvert}\ \mathbb{1}(\lvert X\rvert \leq h)$
Uniforme	$\frac{h}{2}\ \mathbb{1}(\lvert X\rvert \leq h)$	$0\ \mathbb{1}(\lvert X\rvert \leq h)$

$\mathbb{1}$ representa una función indicadora, donde 1 es verdadero y 0 en otro caso.

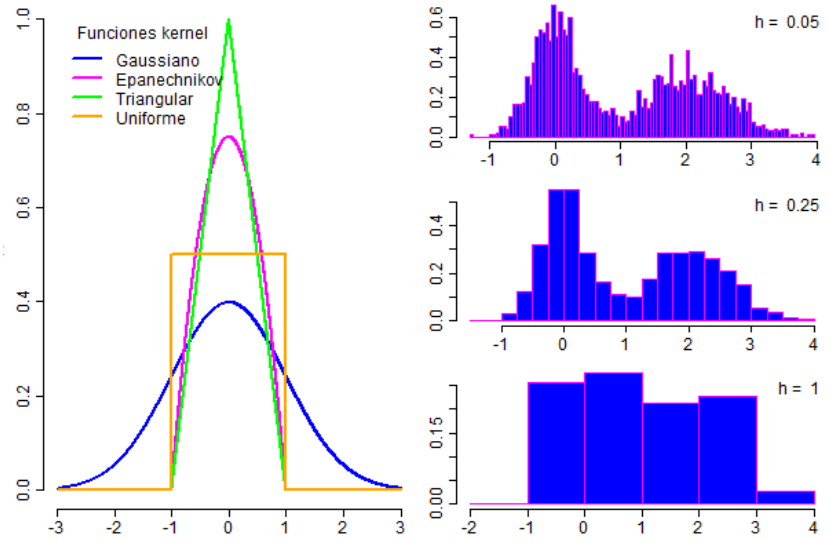

Figura 1.7: Ilustraciones de algunos kernels (lado izquierdo) y del efecto del tamaños de la ventana (lado derecho).

Una vez establecidos $\mathcal{K}$ y h, además de una medida de distancia d, se calcula el gradiente de cambio medio con:

$$m\big(X_i^{(t)}\big) = \frac{\sum_{i=1}^{n} \mathcal{K}'\big(d\big(X_i^{(t)},\, X_i\big)/h\big) X_i}{\sum_{j=1}^{n} \mathcal{K}'\big(d\big(X_i^{(t)},\, X_j\big)/h\big)} \tag{1.12}$$

El algoritmo de cambio medio converge a máximos locales, (Cheng, 1995), al actualizar $X_i^{(t)}$ en cada

iteración, donde $X_i^{(t+1)} = m\big(X_i^{(t)}\big)$. Entonces, se procede recursivamente hasta alcanzar un máximo de iteraciones, $t = T$, o hasta que se alcance un criterio de parada al fijar una cota ϵ que finalice el proceso $\big|m\big(X_i^{(t+1)}\big) - m\big(X_i^{(t)}\big)\big| \leq \epsilon$. En este último paso el proceso tiende a converger a un máximo local.

Con lo anterior expuesto se propone el siguiente algoritmo para el clustering de cambio medio:

Algoritmo de cambio medio

1. Se selecciona la función kernel $\mathcal{K}$, el tamaño de la ventana h y un valor para ϵ. De manera opcional, se puede establecer número máximo de iteraciones, T.
2. Se selecciona un elemento $X_i^{(0)}$ del conjunto $\mathcal{S}_n$.
3. Se calcula el cambio medio de $X_i^{(t)}$ mediante la Ecuación 1.12.
4. Se actualiza la posición de $X_i^{(t)} = m\big(X_i^{(t-1)}\big)$.
5. Los pasos del 2 al 4 se repiten para todo $\mathcal{S}_n$ hasta que cada $X_i^{(t)}$ converja a una moda local o la diferencia con respecto a él mismo en la iteración previa sea menor a la cota fijada $d\big(X_i^{(t)}, X_i^{(t-1)}\big)| \leq \epsilon,\ \forall\, i = 1, \ldots, n$, u opcionalmente hasta que se alcance el máximo de iteraciones $t = T$.
6. Los elementos del conjunto original $\mathcal{S}_n$ se asignan a las modas locales a las que convergieron, donde las modas son los grupos $\mathcal{C} = \{C_j\}_{j=1}$

Un aspecto importante de clustering de cambio medio es que el número de grupos no se puede determinar, sin embargo, existe una relación inversa entre el tamaño de la ventana h y la cantidad de grupos creados: a menor tamaño de h, más grupos, y a mayor tamaño de h, menos grupos.

Elección de $\mathcal{K}$ La elección del kernel $\mathcal{K}$ tiene una importancia secundaria en la estimación de la densidad porque sus resultados tenderán a ser similares siempre y cuando $\mathcal{K}$ cumpla con propiedades deseables: continuidad y simetría.

Tamaño óptimo de h Como se vio en la Figura 1.7 la elección de h tiene el mayor impacto en la estimación de la densidad y, en consecuencia, en el clustering de cambio medio. La estimación del tamaño óptimo de la ventana h no es trivial por lo que se le ha dedicado mucha atención en la literatura. Existen varios métodos para estimar h: selector por la regla del pulgar (Silverman, 1986), selector *plug-in* de (Sheather et al., 1991) o selector por validación cruzada del error global de estimar el parámetro. Sin embargo, y como se explicará más adelante en la Sección 1.3, se opta por evaluar el clustering de cambio medio para una malla de valores de h mediante alguna medida de calidad y se selecciona el valor de h que arroje mejores resultados.

Ejemplos en $\mathbb{R}^2$ Para ilustrar el comportamiento del cambio medio sobre distintos tipos de datos y el impacto del tamaño de la ventana utilizaremos la Figura 1.8. El clustering de cambio medio tiene un buen desempeño en grupos convexos, siempre que se elija un tamaño de ventana adecuado, como se observa en las mixturas de normales de la primera fila, en cambio, tiene un desempeño deficiente en grupos no convexos sin importar la ventana: anillos circulares concentricos y secantes. Por otro lado, se observa en la segunda fila que para un tamaño de h muy pequeño se crean muchos pequeños

grupos, mientras que en la tercera fila se observa que un tamaño alto de h conduce a pocos grupos, es decir, existe una relación inversa entre el tamaño de la ventana y la cantidad grupos formados. Por último, las estimaciones de h por la regla del dedo[3] produjo más grupos de los esperados en los cuatro conjuntos de datos bivariados: 10, 446, 374 y 316, respectivamente.

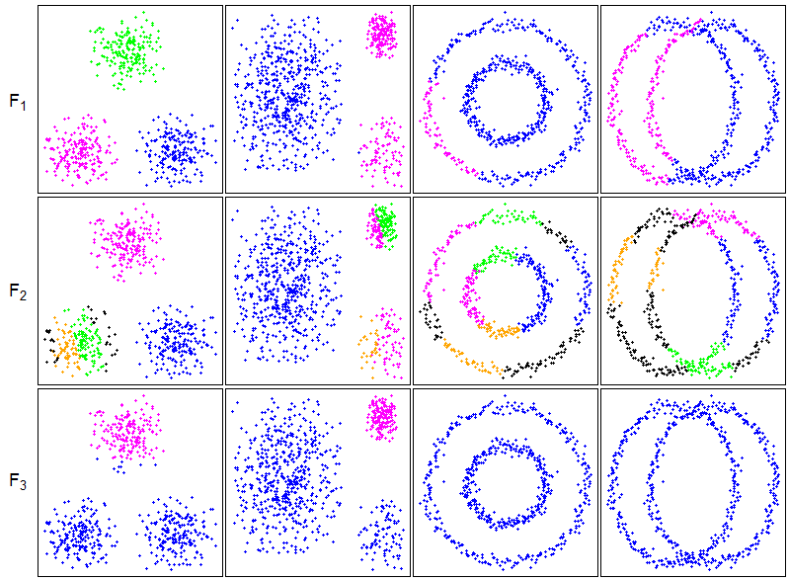

Figura 1.8: Ilustración del clustering de cambio medio en $\mathbb{R}^2$.

1.2.4. DBSCAN

El DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) es un clustering de particionamiento no paramétrico basado en la densidad de los datos, sus siglas en castellano significan *agrupamiento espacial basado en densidad de aplicaciones con ruido*. Este fue propuesto por (Ester et al., 1996) y se ha vuelto muy popular desde entonces por su flexibilidad al formar grupos en patrones de datos no convexos, además de identificar valores atípicos. El objetivo del algoritmo es formar grupos a partir de zonas densas, elementos centrales o núcleos, separadas por zonas de baja densidad, elementos fronterizos o bordes, pero excluyendo de los grupos a los elementos que se encuentran fuera del alcance de los centrales: elementos atípicos o ruido. En este sentido el concepto fundamental es el vecindario de un elemento:

$$N(X_i) = \{X_j : d(X_i, X_j) \leq \epsilon, \ \forall 1 \leq i, j \leq n\} \tag{1.13}$$

donde ϵ define la vecindad alrededor de un elemento, es decir, si la distancia entre dos elementos es menor o igual a ϵ, entonces se consideran vecinos. Por otro lado, se encuentra el parámetro δ que acota la cantidad mínima de elementos en un vecindario para que sea considerado un elemento central.

Dada la naturaleza de los parámetros ϵ y δ cada elemento X_i pertenece a uno, y sólo uno, de los subconjuntos de $\mathcal{S}_n$: central, frontera o ruido (atípicos).

[3] $\boldsymbol{h} = \text{diag}(\boldsymbol{H})$, donde $\boldsymbol{H} = \left(\frac{4}{n(p+2)}\right)^{\frac{2}{(m+4)}} \hat{\Sigma}$.

Definición 1.2 Subconjunto central o núcleo. Es el conjunto compuesto por los elementos que tienen al menos δ vecinos

$$\mathcal{N} = \{X_i : |N(X_i)| \geq \delta\}. \tag{1.14}$$

Definición 1.3 Subconjunto frontera o borde. Es el conjunto compuesto por los elementos que tienen menos de δ vecinos y al menos uno de ellos es un elemento central

$$\mathcal{B} = \{X_i \mid X_i \notin \mathcal{N} \wedge \exists\, X_j \neq X_i : X_j \in \{N(X_i), \mathcal{N}\}\}. \tag{1.15}$$

Definición 1.4 Subconjunto ruido. Es el conjunto compuesto por los elementos que tienen menos de δ vecinos y ninguno de ellos es elemento central

$$\mathcal{R} = \{X_i : X_i \notin \{\mathcal{N}, \mathcal{B}\}\,. \tag{1.16}$$

Los elementos centrales y fronterizos siempre pertenecen a algún grupo C_j, mientras que los elementos atípicos no pertenecen a ninguno. Además, los elementos centrales siempre pertenecen a zonas densas y los elementos fronterizos siempre pertenecen a zonas de baja densidad. Antes de continuar a los pasos del algoritmo, se requieren dos conceptos adicionales para completar la definición de grupo: *accesibilidad* y *conectividad*. La accesibilidad indica si se puede acceder a un elemento desde otro elemento directa o indirectamente, mientras que la conectividad establece si dos elementos pertenecen al mismo grupo o no. Entonces, en términos de accesibilidad y conectividad, se puede hacer referencia a dos elementos en el DBSCAN como:

Definición 1.5 Directamente alcanzable por densidad (dad). Se considera que un elemento X_i está directamente alcanzable por densidad desde un elemento central X_j cuando son vecinos

$$X_i \overset{\text{dad}}{\sim} X_j \implies d(X_i, X_j) \leq \epsilon \wedge X_j \in \mathcal{P}. \tag{1.17}$$

Definición 1.6 Alcanzable por densidad (ad). Se considera que un elemento X_i está alcanzable por densidad desde un elemento central X_j cuando no son vecinos, pero existe una cadena de elementos centrales $\{X_l \in \mathcal{N}\}$ que los une

$$X_i \overset{\text{ad}}{\sim} X_j \implies d(X_i, X_j) > \epsilon \wedge X_i\, X_j \overset{\text{dad}}{\sim} \{X_l \in \mathcal{N}\}. \tag{1.18}$$

Definición 1.7 Conectado por densidad (cd). Se considera que dos elementos fronterizos X_i, X_j están conectados por densidad cuando ambos son alcanzables por densidad desde un elemento central X_l

$$X_i \overset{\text{cd}}{\sim} X_j \implies X_i, X_j \in \mathcal{B} \wedge X_i, X_j \overset{\text{ad}}{\sim} X_l. \tag{1.19}$$

Se debe tener en cuenta que el recíproco no es verdadero en la Definición 1.5, mientras que en las Definiciones 1.6 y 1.7 sí. Con estos conceptos se busca que los elementos de un patrón pertenezcan al mismo grupo cuando exista una cadena de elementos entre ellos. Entonces, se propone el siguiente algoritmo:

Algoritmo DBSCAN

1. Todos los elementos X_i se identifican como no visitados y se calcula la matriz de distancias $\mathcal{D}$ de $\mathcal{S}_n$.
2. Se selecciona aleatoriamente un elemento X_i y, recursivamente, se identifican como elementos centrales pertenecientes al mismo grupo, aquellos vecinos o elementos alcanzables que tengan un vecindario mayor o igual que δ. Los elementos alcanzables con vecindario menor que δ se clasifican como fronterizos.
3. El paso 2 se repite para todos los elementos que no hayan sido visitados.
4. Se finaliza el algoritmo si no hay elementos por visitar. Aquellos elementos que no pertenezcan a ningún grupo son clasificados como ruido: $X_i \in \mathcal{R}$.

Elección de ε La elección del radio ε tiene una importacia primaria en el DBSCAN por lo cual se debe estimar adecuadamente. Un primer acercamiento para resolver el problema puede ser un gráfico de sedimentación porque se puede esperar que en una agrupación de δ elementos la distancia de los elementos centrales y fronterizos estén dentro de cierto rango, mientras que los elementos atípicos pueden tener una distancia mucho mayor, sin embargo, pueden existir varios puntos de inflexión, lo que dificulta la elección del parámetro. La solución al inconveniente de varios puntos de inflexión es promediar las distancias de cada elemento a sus k vecinos más cercanos, donde k es fijado de antemano. Este método consiste en determinar el ε óptimo mediante la estimación del punto de inflexión donde se produce el mayor cambio a lo largo de la curva de distancia de los k-vecinos.

Elección de δ La elección del parámetro δ tiene una importancia secundaria en el DBSCAN y no existe una regla universal para su elección. De todos modos, en espacios multivariantes se suele elegir δ igual al número de dimensiones del conjunto de datos.

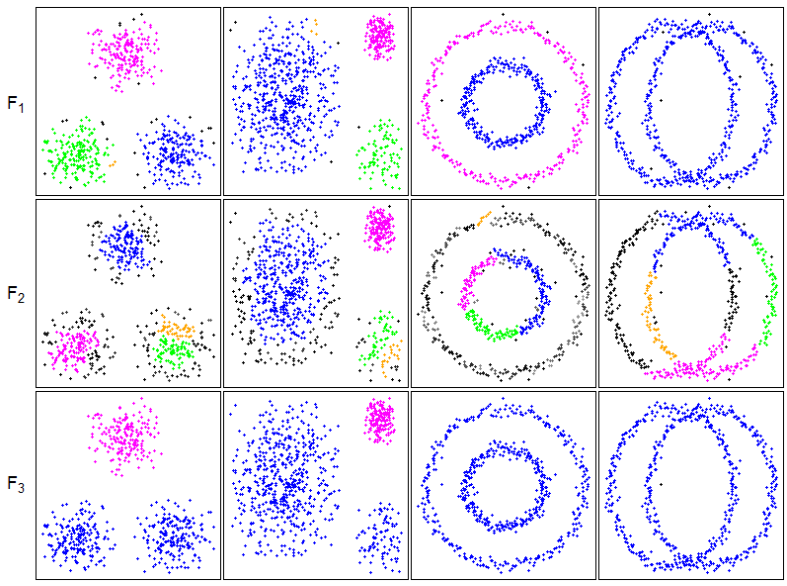

Figura 1.9: Ilustración del DBSCAN en $\mathbb{R}^2$.

Ejemplos en $\mathbb{R}^2$ Para ejemplificar algunas particularidades del DBSCAN nos auxiliamos de los cuatro conjuntos de datos en $\mathbb{R}^2$ mostrados en la Subsección 1.2.1. Los resultados del DBSCAN se muestran en la Figura 1.9. En la primera fila se observa la versatilidad del DBSCAN logrando descifrar tanto patrones convexos como no convexos (mixtura de normales y anillos circulares concéntricos), pero no logra separar los separar los patrones conectados como los anillos secantes. Además, para producir resultados aceptables es necesario fijar un tamaño de ε adecuado. En la segunda fila se muestra el resultado de un valor ε pequeño, lo que produce que muchos elementos sean clasificados como ruido. Mientras, la tercera fila muestra el resultado de un valor ε grande, lo que causa que los elementos sean clasificados en pocos grupos. Si bien no se ofrece una ilustración del parámetro δ, se puede decir que tiene mayor incidencia en el aumento de los elementos clasificados como ruido: un valor δ pequeño tiende a producir poco ruido y un valor δ grande tiende a producir mucho ruido.

Los valores óptimos de ε estimados por los k vecinos más cercanos para los cuatro conjuntos de datos en $\mathbb{R}^2$ son: 0.50, 0.30, 0.9 y 0.9, respectivamente. En la Figura 1.10 se muestran los gráficos de los k vecinos más cercanos para los conjuntos bivariados, donde las líneas discontinuas marcan el valor óptimo de ε: estos se tomaron para realizar el clustering de la fila 1. Mientras que el valor seleccionado de δ corresponde a la cantidad de dimensiones de los conjuntos: 2.

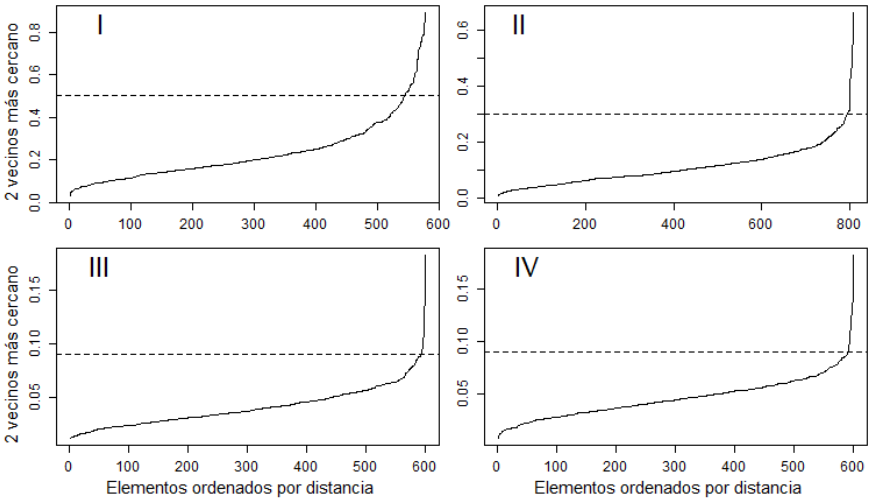

Figura 1.10: 2 vecinos más cercano a los conjuntos de datos en $\mathbb{R}^2$.

1.3. Evaluación del Clustering

Como se ha visto, el principal objetivo del clustering es la asignación de elementos a grupos que no están previamente definidos y de los cuales no se dispone información, por lo cual el clustering enmarca dentro del aprendizaje no supervisado. A esto se añade que pocas veces se dispone de información sobre los parámetros de los métodos de clustering, los cuales determinan la formación de los grupos. Si bien existen técnicas para estimar los parámetros de los métodos de clustering, la estrategia que se adopta en el presente trabajo es determinar los parámetros "*óptimos*" de los métodos de clustering a través de la evaluación de una rejilla de valores de estos y seleccionar aquellos que obtengan mejores resultados de calidad de los grupos formados.

Existen diversas medidas de calidad para evaluar los grupos formados mediante clustering, pero se utilizarán cuatro: índice de silueta, índice de Dunn, índice de Davies-Bouldin e índice de Calinski-Harabasz. En las siguientes subsecciones se expone en qué consiste cada índice.

1.3.1. Índice de Silueta

El índice de silueta es una medida que evalúa la calidad de un método en función de la coherencia de los elementos dentro de los grupos y la separación entre los grupos. Se calcula para cada punto y varía entre -1 y 1, donde un valor cercano a 1 indica que el elemento está bien clasificado en su grupo, mientras que un valor cercano a -1 indica que el punto podría haber sido asignado a un grupo diferente.

Definición 1.8 Índice de Silueta. Dados un conjunto de datos $\mathcal{S}_n$ agrupados en k clústeres $\mathcal{C}$ y una función de distancia d se define el índice de silueta como:

$$IS_k = \frac{1}{n}\sum_{i=1}^{n}\frac{\beta(X_i) - \alpha(X_j)}{\text{máx}\{\alpha(X_i),\, \beta(X_j)\}} \tag{1.20}$$

donde $\alpha(X_i)$ es la distancia promedio entre el elemento X_i y los demás elementos dentro del mismo grupo

$$\alpha(X_i) = \frac{1}{|C_r| - 1}\sum_{\substack{X_j \in C_r \\ i \neq j}} d(X_i,\, X_j), \tag{1.21}$$

mientras que $\beta(X_i)$ es la distancia promedio entre el elemento X_i y los elementos del grupo más cercano diferente al que pertenece

$$\beta(X_i) = \min_{r \neq s}\frac{1}{|C_s|}\sum_{\substack{X_i \in C_r \\ X_j \in C_s}} d(X_i,\, X_j). \tag{1.22}$$

La primera cualidad de este índice es que se encuentra acotado y en segundo lugar su desempeño es bueno en grupos densos y bien separados. Por el contrario, una desventaja del índice de silueta es su bajo desempeño en grupos no convexos como los basados en densidad.

1.3.2. Índice de Dunn

El índice de Dunn es la ratio de la distancia mínima entre los centroides de los grupos y la distancia máxima dentro de los puntos dentro de los grupos: máximo diámetro. Un valor alto del índice indica una mejor separación entre los grupos.

Definición 1.9 Índice de Dunn. Dados un conjunto de datos $\mathcal{S}_n$ agrupados en k clústeres $\mathcal{C}$ con centroides $\{m_1, \ldots, m_k\}$ y una función de distancia d se define el índice de Dunn como:

$$ID_k = \frac{\min_{C_r \neq C_s} d(m_r,\, m_s)}{\max_{C_i = C_j} d(X_i,\, X_j)} \tag{1.23}$$

donde los centroides m_r y m_s pertenecen a los grupos C_r y C_s, respectivamente.

Las ventajas del índice de Dunn son su computo simple y su buen desempeño en grupos bien separados. Por otro lado, su computación es costosa en conjuntos de datos extensos y tiene un bajo desempeño en grupos no convexos.

1.3.3. Índice de Davies-Bouldin

El índice de Davies-Bouldin busca medir la calidad de un clustering en función de la coherencia dentro de los grupos y la separación entre los grupos. Se basa en la relación entre la dispersión dentro de los grupos y la distancia entre los centroides de los grupos. Un valor más bajo del índice de Davies-Bouldin indica una mejor separación y coherencia entre los grupos.

Definición 1.10 Índice de Davies-Bouldin. Dados un conjunto de datos $\mathcal{S}_n$ agrupados en k clústeres $\mathcal{C}$ con centroides $\{m_1, \ldots, m_k\}$ y una función de distancia d se define el índice de Davies-Bouldin como:

$$IDB_k = \frac{1}{k}\sum_{r=1}^{k} \max_{r\neq s}\left\{\frac{\hat{\sigma}_r + \hat{\sigma}_s}{d(m_r, m_s)}\right\} \tag{1.24}$$

donde $\hat{\sigma}_i$ y $\hat{\sigma}_j$ son las desviaciones típicas de los grupos C_r y C_s, respectivamente.

La ventaja del índice de Davies-Bouldin es que se basa únicamente en características inherentes al conjunto de datos, ya que su cálculo sólo utiliza distancias puntuales, además, su cómputo es rápido. Mientras que sus desventajas son su bajo desempeño en grupos no convexos como los basados en densidad y está limitado al espacio euclidiano por el uso de distancia entre centroides.

1.3.4. Índice de Calinski-Harabasz

El índice de Índice de Calinski-Harabasz consiste en maximizar la relación entre la varianza entre grupos y la varianza intragrupo, por lo cual, un valor alto del índice indica una mejor separación entre los grupos y un valor bajo indica que los grupos son muy cercanos.

Definición 1.11 Índice de Calinski-Harabasz. Dados un conjunto de datos $\mathcal{S}_n$ agrupados en k clústeres $\mathcal{C}$ con centroides $\{m_1, \ldots, m_k\}$ y una función de distancia d se define el índice de Davies-Bouldin como:

$$ICH_k = \frac{SS_N(n-k)}{SS_D(k-1)}, \tag{1.25}$$

donde el numerador es la suma de las distancias cuadráticas entre los centroides de los grupos y el centroide del conjunto

$$SS_N = \sum_{r=1}^{k} |C_r|\, d(m_r, m)^2, \tag{1.26}$$

mientras que el denominador es la suma de las distancias cuadráticas dentro de los grupos

$$SS_D = \textstyle\sum_{r=1}^{k} \sum_{X_i \in C_r} d(X_i, m_j)^2. \tag{1.27}$$

Este índice tiene un buen desempeño en grupos densos y bien separados, además, su cómputo es rápido. Por el contrario, el índice de Calinski-Harabasz tiene bajo desempeño para grupos no convexos como los basados en densidad.

1.3.5. Tasa de Aciertos

Cuando se dispone de la afiliación de los elementos X_i a los grupos $\mathcal{C}$ es tentativo evaluar en qué medida los métodos de clustering clasifican correctamente los elementos en sus respectivos grupos, *p. ej.*: datos simulados con grupos de elementos conocidos. Se debe resaltar que este tipo de medidas no

aplica a datos reales, además, cuantificar la asignación correcta de los elementos no es el objetivo del clustering. En cambio, en los métodos de clasificación supervisados, donde se conocen las afiliaciones de los elementos, se estudia la eficiencia mediante una tabla de contingencia llamada matriz de confusión donde las filas representan las clasificaciones esperadas y las columnas las clasificaciones predichas. En el Cuadro 1.5 se muestra la matriz de confusión

Cuadro 1.5: Matriz de confusión.

Esperado	**Predicho**	
	Positivo	**Negativo**
Positivo	Verdaderos Positivos (VP)	Falsos Negativos (FN)
Negativo	Falsos Positivos (FP)	Verdaderos Negativos (VN)

A partir de la matriz de confusión se obtiene la precisión global o tasa de aciertos (TA) de un modelo de clasificación. Esta medida cuantifica lo bueno que es un método asignando los elementos de un conjunto al grupo que pertenece. Además, esta tasa se encuentra acotada en el intervalo [0, 1], donde el valor 0 indica que todos los elementos fueron asignados en grupos distintos a los esperados (mal clasificados) y 1 indica que todos los elementos fueron asignados en los grupos esperados.

Definición 1.12 Tasa de aciertos. Dada una matriz de confusión se define la tasa de aciertos como el cociente de los elementos bien clasificados entre el total de elementos:

$$TA = \frac{VP + VN}{VP + FP + VN + FN}. \tag{1.28}$$

Capítulo 2

Extensión del Clustering a Datos Funcionales

Como se vio en el Capítulo 1 existen diversos métodos para construir grupos en un conjunto de datos. En principio, estos métodos pueden ser extendidos a datos funcionales con algunas consideraciones adicionales relativas a la noción de distancia entre elementos del conjunto de datos. Para extender los métodos de clustering a datos funcionales, primero se ofrecerán los conceptos relativos a las variables y los espacios funcionales. Asimismo, se mostrarán las herramientas estadísticas en estas configuraciones que son necesarias en los procedimientos de clustering.

2.1. Introducción al Análisis de Datos Funcionales

El análisis de datos funcionales extiende los métodos estadísticos multivariantes clásicos a escenarios donde se dispone de un gran número de dimensiones y a los datos que son intrínsecamente curvas (funciones). De acuerdo a (Wang et al., 2015) *el Análisis de Datos Funcionales (ADF) se ocupa del análisis y la teoría de los datos que se encuentran en la forma de funciones, imágenes y formas u objetos más generales*. El ADF ha ido ganando notoriedad en las últimas décadas con los avances tecnológicos en almacenamiento y poder de procesamiento de los ordenadores haciendo posible llevar a cabo procesos intensos en cálculo.

El ADF comparte los principales objetivos con los métodos estadísticos clásicos tales como describir un conjunto de datos a través de sus estadísticos de tendencia central y dispersión, contrastar hipótesis acerca de la población y realizar inferencia sobre la población a partir de una muestra. Sin embargo, la ventaja del ADF es la información adicional que se puede obtener a partir de la configuración funcional de los datos, *p. ej.*: las derivadas. Cabe resaltar que la configuración funcional de los datos requiere implementar técnicas especializadas para obtener información.

2.1.1. Datos Funcionales

De acuerdo a (Ferraty et al., 2006), una variable aleatoria funcional $\mathcal{X}$ es una variable aleatoria que toma valores en un espacio funcional (métrico o pseudométrico) $\mathcal{F}$, por lo cual, un dato funcional es

una realización de un proceso estocástico en un espacio continuo, asimismo una muestra de datos funcionales $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ es una colección de realizaciones del proceso estocástico que son independientes e idénticamente distribuidos (*i.i.d.*). Dos ejemplos de datos funcionales son mostrados en la Figura 2.1, en la parte `A` se muestra el periodograma logarítmico de cinco fonemas del idioma francés[1]. Mientras que en la parte `B` del gráfico se muestran 115 curvas de niveles de NO_X[2] medidos por hora en el barrio de Poblenou, Barcelona, España.

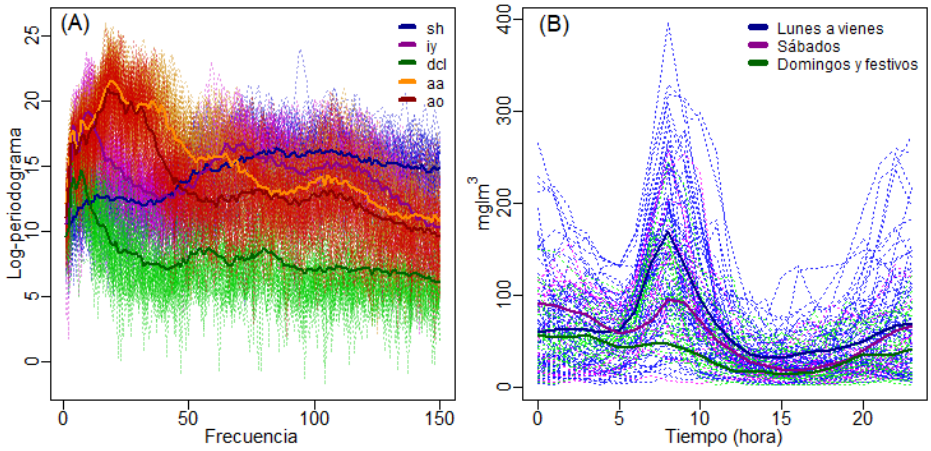

Figura 2.1: Ejemplos de datos funcionales: Phoneme y Poblenou.

En un espacio funcional se pueden definir operaciones internas entre los datos funcionales tales como suma, resta o multiplicación.

Definición 2.1 Espacio funcional. Se dice que una función $d : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$ es un espacio funcional si cumple que es:

1. No negativa: $d(\mathcal{X}, \mathcal{Y}) \geq 0$
2. Indiscernible: $d(\mathcal{X}, \mathcal{Y}) \iff \mathcal{X} = \mathcal{Y}$
3. Simétrica: $d(\mathcal{X}, \mathcal{Y}) = d(\mathcal{Y}, \mathcal{X})$
4. Desigualmente triangular: $d(\mathcal{X}, \mathcal{Z}) = d(\mathcal{X}, \mathcal{Y}) + d(\mathcal{Y}, \mathcal{Z})$

Un espacio funcional ampliamente utilizado en datos funcionales es el espacio de Hilbert, el cual es una generalización del espacio euclidiano. Este espacio métrico será utilizado en la mayoría de los métodos de clustering debido a su propiedades y sencillez.

Definición 2.2 Espacio de Hilbert. Dada una función f definida en el intervalo $\mathcal{S} = [a, b]$ se define el espacio de Hilbert como:

$$\mathcal{L}_2[\mathcal{S}, d] = \left\{ f : \mathcal{S} \longrightarrow \mathbb{R} \text{ tal que } \int_{\mathcal{S}} \left| f(\tau) \right|^2 \mathrm{d}\tau < \infty \right\}. \tag{2.1}$$

[1] Estos datos fueron usados originalmente por (Hastie, Buja et al., 1995) para ilustrar el análisis discriminante lineal a datos funcionales.

[2] El NO_X es una combinación de óxido nítrico (NO) y dióxido de nitrógeno (NO_2). Esta sustancia contaminante tiene efectos negativos en el medio ambiente y en la salud humana. Los datos de NO_X son tomados del paquete `fda.usc` de (Febrero-Bande y Oviedo de la Fuente, 2012).

2.2. Estadísticos Funcionales

2.2.1. Media Funcional

Las medidas de localización y dispersión requeridas en los métodos de clustering y en los índices de calidad requiere una formulación especial en las configuraciones de datos funcionales. Empezamos con la medida más simple de localización la media muestral funcional, la cual se define como la observación que hace mínima la suma de las distancias al cuadrado con respecto a las demás observaciones del conjunto.

Definición 2.3 Media muestral funcional. Sea $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ un conjunto de datos funcionales con $\mathcal{X}_i \sim i.i.d.$ se define la media muestral funcional como:

$$m = \min_{a \in \mathcal{S}_n} \sum_{i=1}^{n} d(\mathcal{X}_i, a)^2. \tag{2.2}$$

La configuración de datos funcionales puede tratarse como datos multivariantes mediante la utilización del espacio de Hilbert. La media muestral funcional en $\mathcal{L}_2$ es más simple que la media muestral funcional general, y es usada en el k-medias y los índices de calidad.

Definición 2.4 Media muestral funcional en $\mathcal{L}_2$. Sea $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ un conjunto de datos funcionales con $\mathcal{X}_i \sim i.i.d.$ se define la media muestral funcional en $\mathcal{L}_2$ como:

$$m = \frac{1}{n} \sum_{i=1}^{n} \mathcal{X}_i. \tag{2.3}$$

2.2.2. Profundidad

En configuraciones de datos funcionales no existe un consenso sobre la definición de la densidad por lo que se utilizan las medidas de profundidad como sucedáneas para tratar problemas que requieran calcular la densidad de datos funcionales. Una definición de profundidad dada por (Febrero-Bande, 2021) es la siguiente:

Definición 2.5 Profundidad. Se dice que una medida de profundidad es un estadístico que ordena de manera decreciente todos los elementos de un conjunto asignando una medida desde el elemento más central $\mathcal{X}_{[n]}$ hasta el elemento más periférico $\mathcal{X}_{[1]}$: los elementos de menor rango son candidatos a ser valores atípicos, mientras que los elementos más rodeados no.

Existen diversas medidas de profundidad para datos funcionales, pero interesa en particular la profundidad modal (Cuevas et al., 2007) que mide la cantidad de elementos en la vecindad y que es utilizada en el clustering de cambio medio en datos funcionales para calcular el vector de cambio entre iteraciones es:

Definición 2.6 Profundidad modal. Dado un conjunto de datos funcionales $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ con $\mathcal{X}_i \sim i.i.d.$ y una medida de distancia d, se define la profundidad modal como:

$$MD(\mathcal{X}) = \sum_{i=1}^{n} \mathcal{K}\left(\frac{d(\mathcal{X}, \mathcal{X}_i)}{h}\right) \tag{2.4}$$

donde $\mathcal{K}$ es una función kernel de pesos y h es el parámetro de suavizado o tamaño de ventana.

Cabe resaltar que la profundidad modal es una medida de cuantos vecinos hay en una h-vecindad, es decir, cuantos elementos están a una distancia h de $\mathcal{X}_i$. Esta medida es muy parecida al estimador no paramétrico de la densidad Rosemblatt-Parzen[3] con la diferencia de que el objetivo de la profundidad modal es proporcionar rangos a los datos.

2.2.3. Dispersión Funcional

Las últimas medidas tratadas corresponden al volumen o tamaño de las curvas medidas a través de la dispersión que estas presenten. Para obtener la dispersión de un conjunto de datos funcionales se utilizan la varianza muestral funcional y la desviación típica funcional.

Definición 2.7 Varianza muestral funcional. Sea $\mathcal{S}_n = \{\mathcal{X}_i\}_{i=1}^n$ un conjunto de datos funcionales con $\mathcal{X}_i \sim i.i.d.$ se define la varianza muestral funcional como:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} d(\mathcal{X}_i,\, m)^2. \tag{2.5}$$

Al igual que en la estadística clásica, la desviación típica muestral funcional se obtiene extrayendo la raíz cuadra positiva de la varianza muestral funcional:

Definición 2.8 Desviación típica muestral funcional. Dada la varianza muestral funcional σ de un conjunto de datos funcional se define la desviación típica muestral funcional como:

$$\sigma = \sqrt{\sigma^2}. \tag{2.6}$$

2.3. Extensión del Clustering

Luego de introducir el análisis de datos funcionales y presentar los estadísticos funcionales requeridos podemos extender los métodos de clustering del Capítulo 1 a los datos funcionales. Extender los métodos de clustering jerárquicos aglomerativos a datos funcionales, en principio, no supone mayor dificultad que utilizar alguna distancia, *p. ej.*: $\mathcal{L}_2$, en el paso 1 del algoritmo que consiste en calcular la matriz de distancias entre los puntos del conjunto de datos. Asimismo, se puede extender el k-medias a datos funcionales usando alguna distancia seleccionada para actualizar los centroides, Ecuación 1.10, en el paso 3 del algoritmo mediante la media muestral funcional u otro estadístico funcional de localización, Ecuación 2.3. En el caso del DBSCAN, se puede usar la distancia $\mathcal{L}_2$ para obtener el vecindario de cada punto, Ecuación 1.13. En cambio, el clustering de cambio medio en datos funcionales requiere sustituir la función de densidad en la Ecuación 1.12 del vector de cambio medio por alguna función sucedánea que pueda calcularse en espacios funcionales como las medidas de profundidad, específicamente la profundidad modal, Ecuación 2.4. Asimismo, en el clustering a datos funcionales se espera un comportamiento similar a los resultados del clustering a datos bivariados vistos en el Capítulo 1.

De igual manera, los índices de silueta, Dunn, Davies-Bouldin y Calinski-Harabasz se pueden calcular usando el espacio de Hilbert $\mathcal{L}_2$. Se sustituyen las medias por la Ecuación 2.3: media muestral funcional, y las medidas de dispersión de varianza y desviación típica por la Ecuación 2.5 y Ecuación 2.6, respectivamente. Con estas sustituciones los métodos de clustering visto queda expandidos a configuraciones de datos funcionales.

[3] El estimador de la densidad Rosenblatt-Parzen en dimensión univariante es $\hat{f}_n = \frac{1}{nh}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h}\right)$.

2.3.1. Implementación en R

Los métodos de clustering son implementados en el lenguaje de análisis estadístico R (R Core Team, 2022). Específicamente, los métodos jerárquicos son implementados mediante la función `hclust()` del paquete `stats` y para llevar a cabo el k-medias a datos funcionales se utiliza la función `kmeans.fd()` del paquete `fda.usc` (Febrero-Bande y Oviedo de la Fuente, 2012). Mientras que se utilizaron las funciones `meanshift.fd()` y `dbscan.fd()` de (Febrero-Bande y Oviedo De La Fuente, 2023) para realizar los clustering de cambio medio y DBSCAN, respectivamente. Asimismo, para calcular el índice de silueta se utilizó la función `silhouette()` del paquete `cluster` (Maechler et al., 2022), la función `dunn()` del paquete `clValid` (Brock et al., 2008) para calcular el índice de Dunn, la función `index.DB()` del paquete `clusterSim` (Walesiak et al., 2020) para calcular el índice de Davies-Bouldin y la función `calinhara()` del paquete `fpc` (Hennig, 2023) para calcular el índice de Calinski-Harabasz.

Capítulo 3

Datos Simulados

3.1. Modelos

Inspirado en (Cuesta-Albertos et al., 2017) tres modelos fueron simulados con el objetivo de comprobar el rendimiento en datos funcionales de los métodos de clustering. Todos los modelos fueron obtenidos de un proceso vexponencial:

$$\mathcal{X} = m_j(\tau) + e_j(\tau) \tag{3.1}$$

donde m_j es la media funcional del grupo $j = 1, 2, 3$ (según el modelo) y e_j es un proceso gaussiano con media cero y matriz de convarianzas:

$$\mathbb{C}ov\big(e_j(\pi),\, e_j(\tau)\big) = \theta_j \exp\left(-\frac{|\pi - \tau|}{0.3}\right). \tag{3.2}$$

Las funciones fueron generadas en el intervalo $\tau \in [0, 1]$ usando una rejilla de 100 puntos de discretización equidistante y en todos los modelos cada población se basa en 50 corridas independientes. A continuación, se describe cada modelo:

Modelo 1 Esta simulación contiene dos poblaciones donde la media y la escala de la población 1 son $P_{11} = 20(1-\tau)\tau^{1.1}$ y $\sigma_{11} = 0.20$, mientras que para la población 2 son $P_{12} = -10(1-\tau)^{1.1}\tau$ y $\sigma_{12} = 0.25$. Al tratarse de dos poblaciones de curvaturas opuestas se forma un ojo, el cual se espera que los métodos clustering puedan separar con facilidad (ver parte izquierda de la Figura 3.1).

Modelo 2 Esta simulación contiene tres poblaciones donde la media y la escala de la población 1 son $P_{21} = 20(1-\tau)\tau^{1.1}$ y $\sigma_{21} = 0.20$; para la población 2 son $P_{22} = -10(1-\tau)^{1.1}\tau$ y $\sigma_{22} = 0.25$; y para la población 3 son $P_{23} = 30(1-\tau)\tau^{1.1}$ y $\sigma_{23} = 0.20$. Este modelo contiene las poblaciones del modelo 1 y una población adicional (ver parte central de la Figura 3.1).

Modelo 3 Esta simulación contiene cuatro poblaciones donde la media y la escala de la población 1 son $P_{31} = 20(1-\tau)\tau^{1.1}$ y $\sigma_{31} = 0.20$; para la población 2 son $P_{32} = -10(1-\tau)^{1.1}\tau$ y $\sigma_{32} = 0.25$; población 3 son $P_{33} = 30(1-\tau)\tau^{1.1}$ y $\sigma_{33} = 0.20$; y para la población 4 son $P_{34} = -20(1-\tau)^{1.1}\tau$ y $\sigma_{34} = 0.25$. Este modelo contiene dos poblaciones internas, P_{31} y P_{32}, y dos poblaciones externas, P_{33} y P_{34} (ver parte derecha de la Figura 3.1).

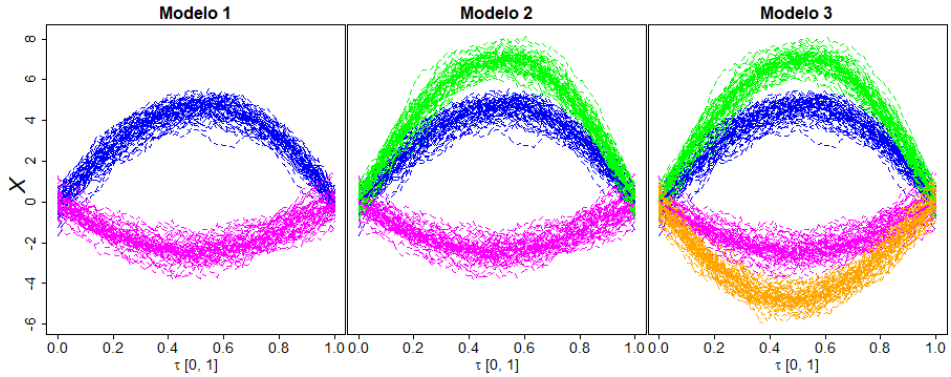

Figura 3.1: Modelos simulados de datos funcionales. Modelo 1: P_{11} curvas azules y P_{12} curvas magentas. Modelo 2: P_{21} curvas azules, P_{22} curvas magentas y P_{23} curvas verdes. Modelo 3: P_{31} curvas azules, P_{32} curvas magentas, P_{33} curvas verdes y P_{34} curvas naranjas.

3.2. Resultados

Las simulaciones consisten en poblaciones de curvas de curvaturas opuestas y de distintas escalas de dispersión, con lo cual se pretende evaluar si los métodos de clustering pueden segmentar tales poblaciones de acuerdo con lo esperado. En el Cuadro 3.1 se muestra la selección de parámetros con la que se llevó a cabo el clustering a los datos funcionales simulados, donde se seleccionó un conjunto de parámetros que produjeran tantos grupos como poblaciones tenga cada modelo debido a que los gráficos de sedimentación de los clustering jerárquicos y k-medias estiman como óptimo $k = 2$, en cuanto a los valores óptimos de h y ε de los métodos de cambio medio y DBSCAN, respectivamente, se crean dos grupos en todos los modelos. En el Apéndice B.1 se muestran los gráficos de sedimentación de los métodos de clustering.

Cuadro 3.1: Selección de parámetros por método de clustering a las simulaciones de datos funcionales según modelo.

Índice de calidad	**Enlace mínimo**	**Enlace máximo**	***k*-Medias**		**Cambio medio**	**DBSCAN**	
	k	k	k	δ	h	ϵ	δ
1	2	2	2	1	0.216	0.55	5
2	3	3	3	1	0.229	0.55	5
3	4	4	4	1	0.231	0.55	5

Por otro lado, la Figura 3.2 muestra las asignaciones de cada clustering, donde cada fila es un modelo ordenados de manera ascendente. A primera vista la mayoría de los métodos clasificaron correctamente las curvas según la población a la que pertenecen. En cambio, se observa que el k-medias tuvo un desempeño deficiente al dividir las poblaciones con la misma curvatura en los modelos 2 y 3. En el Apéndice C se muestran las matrices de confusión.

Antes de continuar, es importante remarcar que la clasificación no es el objetivo principal del clustering, dado que se trata de un método de aprendizaje no supervisado. Sin embargo, se obtuvieron medidas de eficiencia de la clasificación de los métodos para comparar el desempeño que obtuvieron en estas simulaciones. El Cuadro 3.2 muestra la tasa de aciertos por modelo según método. Se resalta el hecho

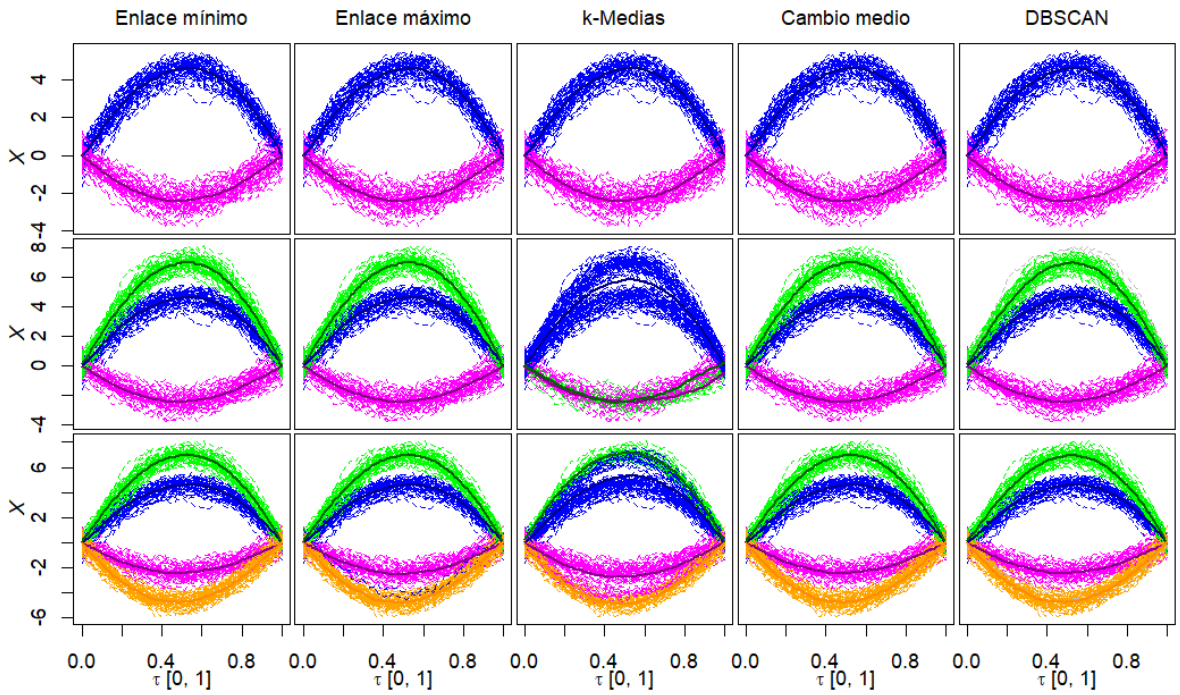

Figura 3.2: Resultados de los métodos de clustering a la simulaciones de datos funcionales.

de que los métodos obtuvieron altas tasas de aciertos en su mayoría, a excepción del k-medias en los modelos 2 y 3, lo que confirma los visto en la Figura 3.2.

Cuadro 3.2: Tasa de aciertos por método de clustering a las simulaciones de datos funcionales según modelo.

Modelo	Métodos de clustering				
	Enlace mínimo	**Enlace máximo**	***k*-Medias**	**Cambio medio**	**DBSCAN**
1	1.000	1.000	1.000	1.000	1.000
2	1.000	1.000	0.447	1.000	0.987
3	1.000	0.990	0.830	1.000	0.990

Para completar el análisis sobre el clustering a las simulaciones se obtuvieron las medidas de calidad de los grupos creados. En el Cuadro 3.2 se muestran los resultados de los índices de calidad por modelo y escenario según el tipo de índice empleado y el método de clustering.

En el Cuadro 3.3 se muestran los resultados de los índices de calidad por modelo según clustering. En el caso del índice de silueta, en el modelo 1 todos los métodos obtuvieron el mismo resultado, el cual es el más alto entre todos los modelo puesto que para los métodos de clustering es más fácil segmentar los puntos en dos grupos, mientras que en los modelos 2 y 3 se obtuvieron resultados más bajos y variados donde los métodos jerárquicos y el cambio medio presentan mejor desempeño, el k-medias y el DBSCAN peores desempeño y, en general, a medida que el número de poblaciones aumenta el índice disminuye. El índice de Dunn presenta mejores resultados en los clustering del modelo 1, mientras que para los demás modelos es mucho más bajo donde los métodos jerárquicos y el cambio medio

Cuadro 3.3: Índices de calidad por método de clustering a las simulaciones de datos funcionales según modelo.

Modelo	Métodos de clustering				
	Enlace mínimo	**Enlace máximo**	***k*-Medias**	**Cambio medio**	**DBSCAN**
			Silueta		
1	0.872	0.872	0.872	0.872	0.872
2	0.714	0.714	0.571	0.714	0.626
3	0.629	0.622	0.419	0.629	0.563
			Dunn		
1	2.148	2.148	2.148	2.148	2.148
2	0.389	0.389	0.066	0.389	0.344
3	0.337	0.243	0.084	0.337	0.344
			Davies-Bouldin		
1	0.115	0.115	0.115	0.115	0.115
2	0.307	0.307	1.842	0.307	0.293
3	0.381	0.390	0.614	0.381	0.372
			Calinski-Harabasz		
1	2726.155	2726.155	2726.155	2726.155	2726.155
2	2672.199	2672.199	822.680	2672.199	1853.358
3	3522.078	3436.969	1769.337	3522.078	2720.971

presentan mejores resultados, el k-medias y el DBSCAN peores desempeño y, en general, a medida que el número de poblaciones aumenta el índice disminuye. Los resultados del índice Davies-Bouldin muestran un mejor desempeño de los métodos de clustering en el modelo 1, mientras que en los modelos 2 y 3 los resultados fueron peores a medida que la cantidad de poblaciones aumentaba en los modelos y en general todos los métodos obtuvieron valores similares a diferencia del k-medias con valores más deficientes. Por último, el índice de Calinski-Harabasz muestra mejores resultados en los modelos 1 y 2, específicamente en el k-medias y el DBSCAN, mientras que los métodos jerárquicos y el cambio medio obtuvieron peores resultados.

Capítulo 4

Datos Reales

4.1. Generación Eléctrica Fotovoltaica

Los datos reales consisten en una serie temporal de generación eléctrica fotovoltaica en España en el año 2021 (Red Eléctrica de España, 2022) donde la observación $\mathcal{X}_i$ es la generación fotovoltaica en MWh[1] en el día i-ésimo por lo cual el conjunto de datos contiene 365 observaciones, $\mathcal{S}_{365} = \{\mathcal{X}_i\}_{i=1}^{365}$. En cuanto a la discretización de las observaciones, se toman 24 puntos que corresponden al inicio de cada hora del día $\tau \in [0, 23]$. En la Figura 4.1 se muestra el conjunto de datos donde se aprecia la concavidad positiva de la generación fotovoltaica: este comportamiento describe la intensidad de energía solar capturada durante el día que se hace máxima posterior al mediodía.

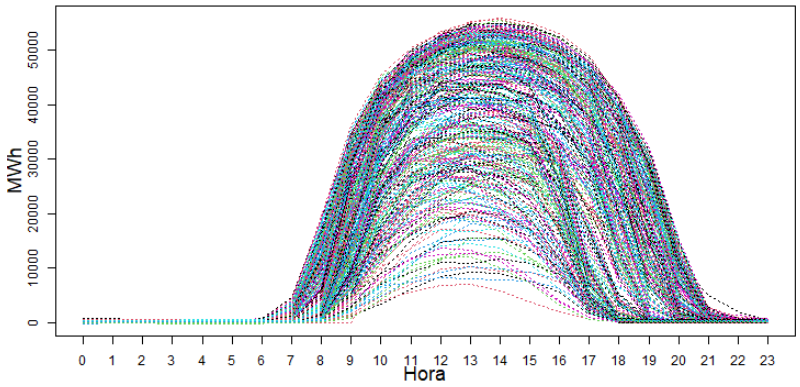

Figura 4.1: Generación eléctrica fotovoltaica en España, 2021.

La distribución de máximos de generación eléctrica fotovoltaica por hora se muestra en el Cuadro 4.1 donde se observa que el rango de la distribución está entre las 11 y las 16 horas, sin embargo, en el intervalo de las 12 a las 15 horas se encuentran más del 95 % de los días. En este punto cabe resaltar que la producción eléctrica mediante esta tecnología está condicionada por el estado del tiempo, específicamente la nubosidad. De todos modos, esta distribución nos brinda una idea sobre el comportamiento de la luminosidad y al mismo tiempo nos sugiere que pueden existir varios grupos de luminosidad

[1] Megavatio hora.

durante el año.

Cuadro 4.1: Distribución de máximos de la generación fotovoltaica por hora.

Frecuencia	**Hora**					
	11	**12**	**13**	**14**	**15**	**16**
Absoluta	12	51	152	131	18	1
Relativa	0.033	0.140	0.416	0.359	0.049	0.003

4.1.1. Modelado

La primera dificultad de trabajar con datos funcionales se debe a que estos tienen lugar en un espacio infinito-dimensional, mientras que sólo podemos observar su trayectoria en un conjunto finito de puntos, sólo disponemos de observaciones discretas X_{ij} de cada generación de energía fotovoltaica $\mathcal{X}_i(\tau_{ij})$ en un conjunto finito de cortes $\{\tau_{ij} : j = 0, \dots 23\}$. El enfoque más común para abordar este problema es reconstruir la forma funcional de los datos mediante un conjunto de bases[2]. El objetivo de modelar los datos es suavizar la trayectoria de cada realización para mejorar la su visualización, ayudando a extraer información.

Definición 4.1 Representación en base. Dado un dato funcional $\{\mathcal{X}(\tau) : \tau \in [\tau_j, \tau_{j+1}], j = 1, 2, \dots, P-1\}$ que admite una expansión en base $\Phi = \{\phi_l\}_{l=1}^{p}$ en $\mathcal{L}_2$ se representa por:

$$\mathcal{X}(\tau) = \sum_{l=1}^{p} \alpha_l \phi_l(\tau) \tag{4.1}$$

donde $p \in \mathbb{N}$ y $\alpha_l \in \mathbb{R}$.

Cada sistema de bases posee características particulares con ventajas y desventajas por lo que su elección depende de las atribuciones que se hagan a los datos funcionales. Usualmente, se elige una base suave de al menos una derivada, además, por razones computacionales se prefiere elegir el menor número p de funciones base como sea posible. Los sistemas de bases más comunes son los fijos: *BSplines*, *Fourier*, *Polinomiales* y *Wavelets*, aunque existen muchos otros. De todos estos sistemas destacamos las BSplines y las bases de Fourier con los cuales suavizaremos los datos de generación fotovoltaica.

Definición 4.2 BSplines. Dado un conjunto de polinomios de orden p definido en el intervalo $\mathcal{T} = [\tau_1, \tau_m]$ de tal manera que los bordes de los subintervalos coincidan hasta la derivada $p-2$, se define el sistema de BSplines como:

$$\mathcal{X}(\tau) = \sum_{i=1}^{p+L-1} c_i \phi_i(\tau, \pi) \tag{4.2}$$

donde $L-1$ es el número de puntos interiores π.

Las BSplines son bases muy flexibles adaptando los cambios en la trayectoria de las curvas, sin embargo, presentan problemas en las fronteras al tratar de interpolar los valores en esas zonas con información poco variable creando un comportamiento espurio. En la Figura 4.2 se muestra la representación de los datos de generación eléctrica fotovoltaica en 10 bases de BSplines, donde se observa el comportamiento antes descrito en horario nocturno y una ampliación del intervalo de mayor captación de energía a medida que el número de horas de luz solar aumenta.

[2]Para una consulta más detallada véase (Ramsay et al., 2005).

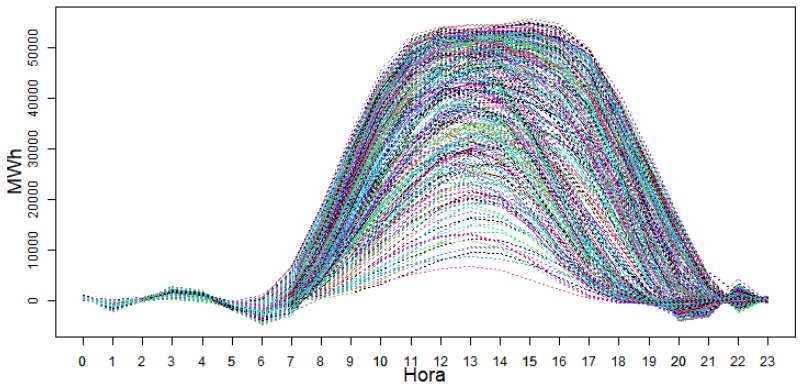

Figura 4.2: Representación en bsplines de la generación eléctrica fotovoltaica en España, 2021.

Definición 4.3 Bases de Fourier. Dado un intervalo $\mathcal{T} = [\tau_1, \tau_m]$, se define el sistema de bases de Fourier como:

$$\left\{ \phi_0 = \frac{1}{\sqrt{\mathcal{T}}}, \left\{ \phi_{2r-1}(\tau) = \frac{\text{sen}(r\omega\tau)}{\sqrt{\mathcal{T}/2}}, \ \phi_{2r}(\tau) = \frac{\cos(r\omega\tau)}{\sqrt{\mathcal{T}/2}} \right\}_{r\in\mathbb{N}} \right\} \tag{4.3}$$

donde $\omega = 2\pi/\mathcal{T}$ (la constante ω describe la amplitud del intervalo).

Las bases de Fourier son idóneas para trabajar con datos periódicos que no presenten picos ni valles muy pronunciados, lo que es ideal para el rango del día donde hay luz solar, pero el sistema de bases introduce valores negativos en el horario nocturno debido a la poca o nula captación de energía que hace que los valores sean constantes o muy similares, escenario que contrasta con el comportamiento sinusoidal de las bases de Fourier. En la Figura 4.2 se muestra la representación de los datos de generación eléctrica fotovoltaica en 5 bases de Fourier, donde se observa un desplazamiento del máximo de generación fotovoltaica a medida que la amplitud de la onda se expande, es decir, los días con pocas horas de luz tiene un máximo más temprano que los días de mayor luz solar.

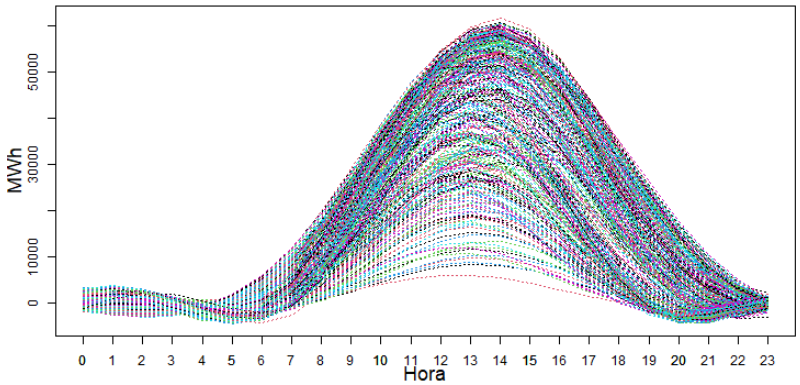

Figura 4.3: Representación en bases de Fourier de la generación eléctrica fotovoltaica en España, 2021.

Como dato adicional, para ambos sistemas de bases se calculó el número óptimo de bases mediante la validación cruzada generalizada (Febrero-Bande y Oviedo de la Fuente, 2012). El número óptimo de bases estimado fue 23 y 12 para los sistemas BSplines y Fourier, respectivamente. Dado el alto número de bases estimados como óptimo se fijó en alrededor del 40 % de estos el número de bases.

4.1.2. Derivadas

Las derivadas son funciones que aportan mucha información analítica sobre el comportamiento de una función, p. ej. la primera derivada indica la velocidad de cambio de una función en un punto, además nos ayuda a obtener los mínimos y máximos de una función. Estas herramientas son muy útiles en la etapa exploratoria previa al clustering. Para estimar la derivada de un dato funcional se dispone de varias opciones, entre las más simple: derivar las bases, Ecuación 4.1, en la cual se representen los datos, o calcular los cocientes de cambios consecutivos. Sin embargo, se descarta la diferenciación de la representación de generación eléctrica fotovoltaica en bases debido al comportamiento que presentan en el horario nocturno, por ende, se obtiene la derivada de la generación eléctrica fotovoltaica por diferenciación directa.

Definición 4.4 Diferenciación directa. Dada una variable funcional $\mathcal{X}$ definida en un intervalo $\mathcal{T} = [\tau_1, \tau_m]$, se define la diferenciación directa como:

$$\mathcal{X}'(\tau_j) = \frac{\mathcal{X}(\tau_j) - \mathcal{X}(\tau_{j-1})}{\tau_j - \tau_{j-1}}, \quad \forall\, j = 2, \ldots, m. \tag{4.4}$$

En la Figura 4.4 se muestra la derivada mediante la Ecuación 4.4 de la generación eléctrica fotovoltaica donde se observa que las horas en las que se hace máxima la generación de energía fotovoltaica no queda clara debido al aglutinamiento de los datos, sin embargo, se observan al menos tres grupos de datos con distintas horas de amanecida y al menos cuatro grupos de datos con distintos horas de anochecida.

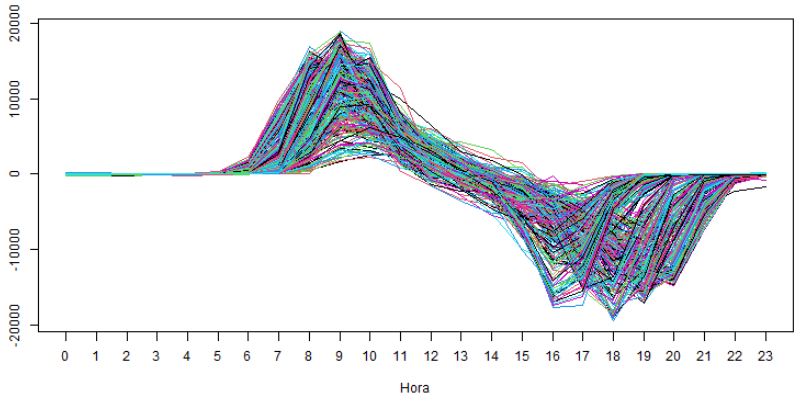

Figura 4.4: Derivada de la generación eléctrica fotovoltaica en España, 2021.

4.2. Resultados

Para realizar el clustering a la generación eléctrica fotovoltaica se evaluó el conjunto de datos en una rejilla de valores de sus parámetros, luego se seleccionaron los valores de los parámetros que mejor desempeño obtuvieron en los índices de calidad. De manera específica, en los métodos que se requiere fijar la cantidad de grupos se evaluaron en $k = 2$, 3 y 4, entendido estos grupos de generación eléctrica fotovoltaica como: {`baja` y `alta`}; {`baja`, `media` y `alta`}; y {`baja`, `intermedia baja`, `intermedia alta` y `alta`}, respectivamente. En cuanto al tamaño mínimo de los grupos en los métodos que se requiera, se evaluó $\delta = 7$, 14, 21 y 28 que representan las semanas que duró cada grupo o período[3].

[3]Se limitó a 28 días con el objetivo de flexibilizar la creación de grupos.

En cambio, el parámetro h del clustering de cambio medio se evaluó en una amplía rejilla de valores debido a que no es evidente la relación de las estimaciones del tamaño de la ventana por la regla del dedo, mientras que se evaluó el DBSCAN en una rejilla de valores para ε entre el óptimo estimado por los 2 vecinos más cercanos (mínimo) y el óptimo estimado entre los 7 vecinos más cercanos (máximo). Para los dos últimos parámetros se procedió de la manera siguiente:

- h se evaluó para 10 valores de MWh equidistantes en el intervalo [7325.5, 8815.3].
- ϵ se evaluó en 80 valores de MWh equidistantes en el intervalo [5000, 12900].

En el Cuadro 4.2 se muestra la selección de parámetros por clustering, mientras que la cantidad de grupos creados por los parámetros que mejoran los índices de Silueta, Dunn, Davies-Bouldin y Calinski-Harabasz en el clustering de cambio medio fueron 2, 3, 2 y 2, respectivamente, mientras que en ese mismo orden la cantidad de grupos creados por el DBSCAN fueron 2, 2, 2, y 1.

Cuadro 4.2: Selección de parámetros por método de clustering a los datos de generación eléctrica fotovoltaica.

Índice de calidad	**Enlace mínimo**	**Enlace máximo**	***k*-Medias**		**Cambio medio**	**DBSCAN**	
	k	k	k	δ	h	ϵ	δ
Silueta	2	2	2	7	8653.804	11500	14
Dunn	2	4	4	21	7652.835	5200	7
Davies-Bouldin	2	3	4	21	8653.804	5600	7
Calinski-Harabasz	3	2	3	28	8653.804	5600	7

Los valores de k donde se produce la mayor separación en los clustering jerárquicos de enlace mínimo y enlace máximo a los datos de generación fotovoltaica son 47 y 3, respectivamente. En ese mismo orden, los valores de k donde se producen las mayores caídas relativas de la suma total de cuadrados intragrupo son 12 y 2, respectivamente. Un resultado similar se obtuvo para el clustering k-medias, donde la mayor caída relativa de la suma total de cuadrados intragrupo se produjo en $k = 2$.

En cuanto a los resultados de los índices de calidad, el Cuadro 4.3 muestra los mejores resultados de las cuatro medidas.

Cuadro 4.3: Índices de calidad por método de clustering a los datos de generación eléctrica fotovoltaica.

Índice de calidad	**Enlace mínimo**	**Enlace máximo**	***k*-Medias**	**Cambio medio**	**DBSCAN**
Silueta	-0.087	0.509	0.462	0.526	0.381
Dunn	0.091	0.078	0.058	0.056	0.272
Davies-Bouldin	1.254	0.700	0.711	0.726	0.111
Calinski-Harabasz	0.627	605.576	586.022	642.055	∞

En la Figura 4.3 se muestra la clasificación de las curvas de generación fotovoltaica a partir de la selección de los parámetros. De los clustering jerárquico, el enlace máximo obtuvo mejores resultados en todos los índices en comparación al método de enlace mínimo, esto se debido a que el enlace máximo creo grupos más balanceados, mientras que el método de enlace mínimo creó grupos desbalanceados:

en el Apéndice A se muestran los dendrogramas de ambos métodos de enlace para cada valor de k. Por otro lado, el clustering de k-medias arrojó buenos resultados en la mayoría de los índices de calidad debido a que separó las curvas por la longitud de su período, es decir, clasificó la generación fotovoltaica en grupos de días de luminosidad solar corta, media y larga, sin embargo el resultado del índice de silueta indica una mala asignación de los elementos debido a que la distancia promedio dentro de los grupos es mayor a la distancia de los elementos al grupo más cercano. El clustering de cambio medio obtuvo resultados intermedios en la mayoría de los índices en comparación con los resultados de los otros clustering con la diferencia de crear grupos de generación fotovoltaica por la amplitud de las curvas, es decir, clasificó la generación en grupos de días de luminosidad corta y larga. Por último se muestran los resultados del DBSCAN, el cual mostró un bajo desempeño comparativo a excepción del índice de Calinski-Harabasz, sin embargo, los resultados de este método quedan en entredicho dado el elevado porcentaje de puntos clasificados como ruido: 27.4 %, 86.8 %, 83.0 %, 90.7 % para los índices del Cuadro 4.3 de arriba a abajo, respectivamente, este problema ocurre como consecuencia de evitar que las curvas se clasifiquen en el mismo grupo debido a la cercanía entre ellas, es decir, los datos son muy sensibles al tamaño de ϵ y cuando este es pequeño la mayoría de días se clasifica como ruido: sólo para los parámetros seleccionados por el índice de silueta se obtuvieron dos grupos bien diferenciados según la intensidad lumínica: corta y larga.

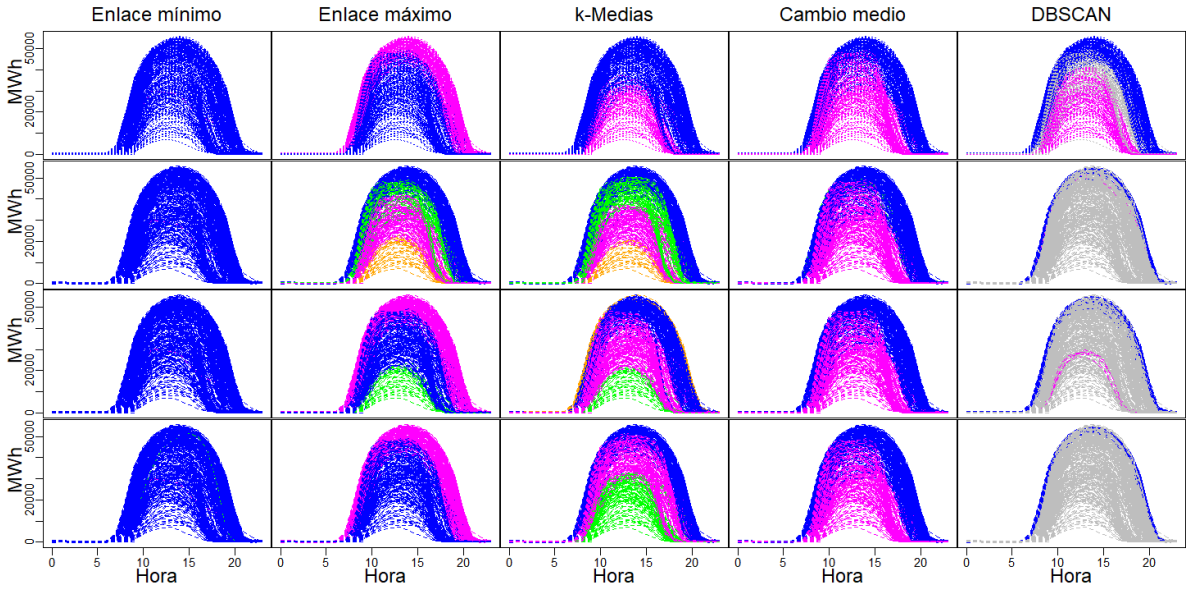

Figura 4.5: Clasificación de los métodos de clustering a la generación de eléctrica fotovoltaica según los parámetros que mejoran los resultados de los índices de calidad. En la primera fila (arriba) según el índice de silueta; en la segunda fila según el índice de Dunn; en la tercera fila según el índice de Davies-Bouldin; y en la cuarta fila (abajo) según el índice de Calinski-Harabasz.

Capítulo 5

Discusión

Como se ha visto a lo largo del presente trabajo, para llevar a cabo los métodos de clustering sobre datos funcionales se utilizaron en gran medida la extensión de las herramientas multivariantes, por tanto, en los escenarios tratados se interpretan los resultados como en la estadística clásica. Lo primero a destacar es la naturaleza de los métodos de clustering, que para formar grupos toman en cuenta distintos enfoques que determinan su rendimiento y sus resultados, además el desempeño de cada método en términos de la calidad de los grupos creados estuvo relacionado con los patrones en los cuales se disponen los datos, siendo algunos más ventajosos que otros en determinados escenarios y viceversa, *p. ej.*: se vio que los métodos de clustering jerárquico de enlace máximo, k-medias y cambio medio trabajan bien sobre patrones de datos convexos, mientras que los métodos de clustering jerárquico de enlace mínimo y el DBSCAN trabajan bien tanto con datos de patrones convexos como con datos de patrones no convexos siempre que los patrones estén bien delimitados.

Los puntos destacados de cada método se listan a continuación:

- El clustering jerárquico de enlace mínimo arrojó buenos resultados en los tres conjuntos de datos funcionales simulados. Sin embargo, obtuvo resultados precarios aplicado al conjunto de datos de generación eléctrica fotovoltaica debido a la proximidad entre los datos. Este enlace maneja bien los patrones no convexos, pero no así los conjuntos de datos poco delimitados.
- El clustering jerárquico de enlace máximo logró buenos resultados en los tres conjuntos de datos funcionales simulados. Asimismo, el método obtuvo buenos resultados para los datos de generación eléctrica fotovoltaica. Aunque este enlace no maneja bien los patrones no convexos, no es sensible al ruido como el enlace mínimo.
- El clustering de k-medias obtuvo un desempeño aceptable en los modelos de datos funcionales simulados con excepción del modelo 2, lo que muestra que el método no funciona adecuadamente con datos no convexos. En cambio, si bien el método obtuvo resultados de calidad regulares en los datos de generación eléctrica fotovoltaica, la agrupación que hizo fue consistente en delimitar los grupos por longitud del período.
- El clustering de cambio medio se desempeñó bien en los escenarios simulados logrando agrupar los datos en las poblaciones correspondientes. Mientras que en el escenario real tuvo un desempeño aceptable formando dos grupos según la profundidad de los datos. Caber resaltar el rol que jugó

una selección del parámetro de ventana adecuado en el diseño de los grupos: no muy pequeña para evitar la formación de muchos grupos locales, ni muy grande para evitar el agrupamiento de todos los datos en un único grupo.

- El DBSCAN obtuvo buenos resultados para los datos funcionales simulados debido a que las poblaciones en cada modelo se encontraban delimitadas. En cambio, no ofreció los resultados esperados en el escenario real debido a que los datos de generación fotovoltaica se encontraban muy agrupados, derivando en la clasificación de una proporción importante de los datos como ruido. De todos modos, en un escenario pudo separa la luminicidad por día en dos: corta y larga.

Por último, se recomienda para investigaciones futuras determinar los parámetros de los métodos de clustering mediante la combinación de la evaluación de la calidad de los grupos con múltiples técnicas de estimación de los parámetros de los clustering.

Apéndice A

Dendrogramas

Dendrogramas de los clustering jerárquicos a datos funcionales simulados y datos de generación eléctrica fotovoltaica.

A.1. Datos Simulados

Figura A.1: Dendrogramas de los métodos de clustering jerárquicos a los datos funcionales simulados.

A.2. Generación Eléctrica Fotovoltaica

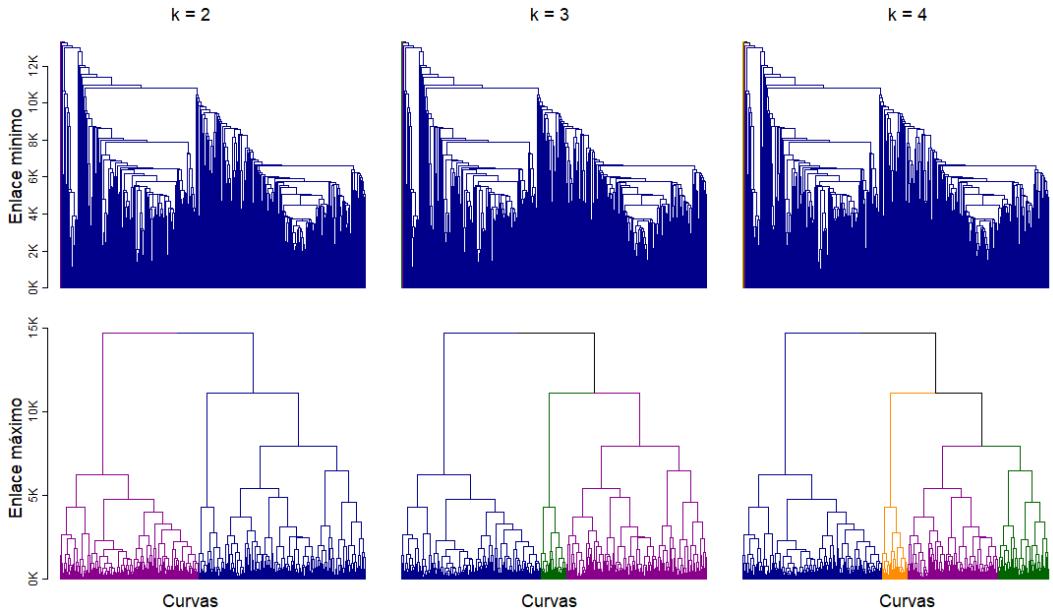

Figura A.2: Dendrogramas de los métodos de clustering jerárquicos a datos de generación eléctrica fotovoltaica.

Apéndice B

Gráficos de Sedimentación

Gráficos de sedimentación de los clustering jerárquicos y k-medias y graficos de los k vecinos más cercanos a datos funcionales simulados y datos de generación eléctrica fotovoltaica.

B.1. Datos Simulados

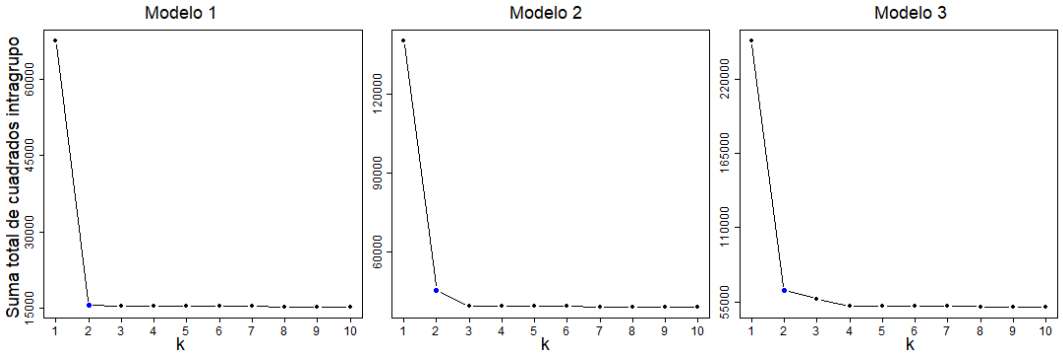

Figura B.1: Gráfico de sedimentación del clustering jerárquico de enlace mínimo a los datos funcionales simulados.

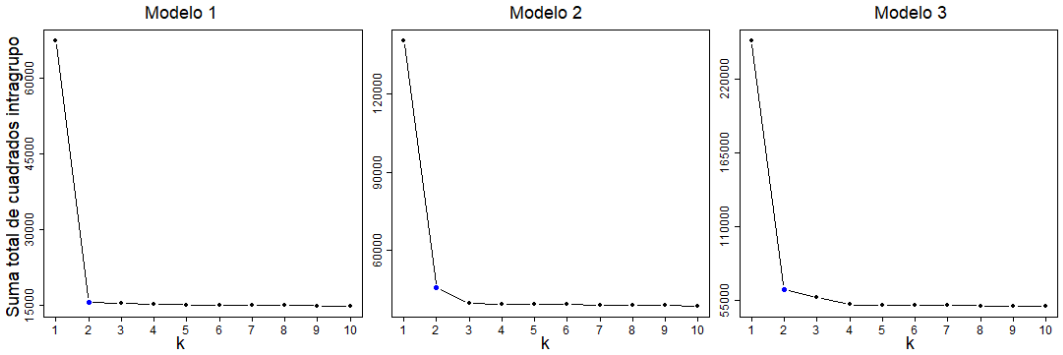

Figura B.2: Gráfico de sedimentación del clustering jerárquico de enlace máximo a los datos funcionales simulados.

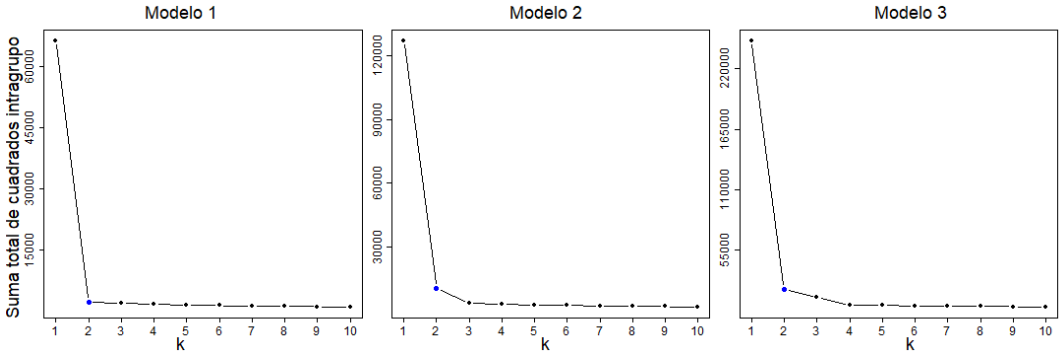

Figura B.3: Gráfico de sedimentación del clustering k-medias a los datos funcionales simulados.

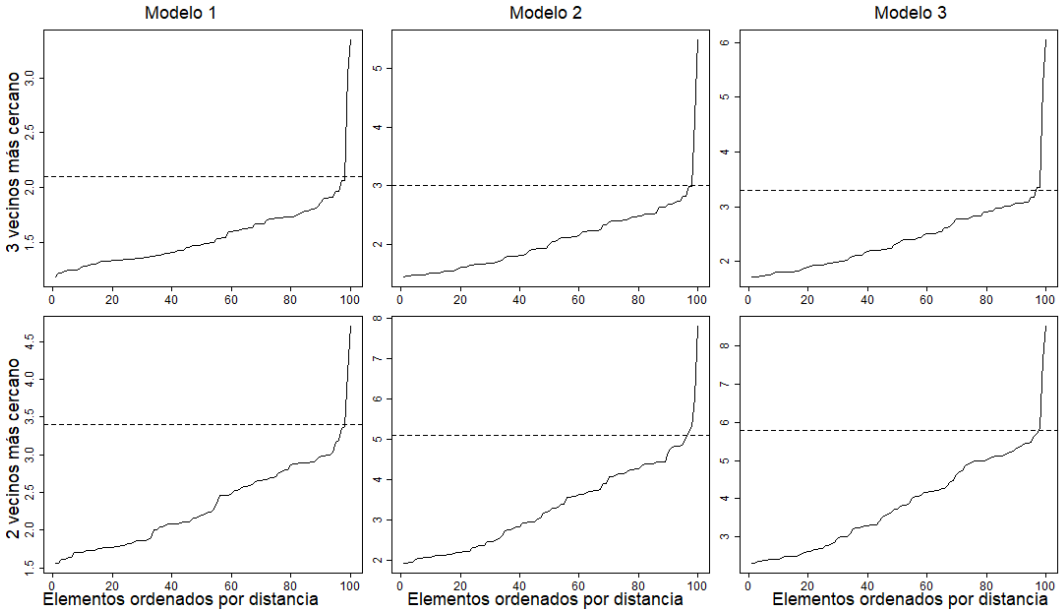

Figura B.4: Gráfico de k vecinos más cercanos del DBSCAN a los datos funcionales simulados.

B.2. Generación Eléctrica Fotovoltaica

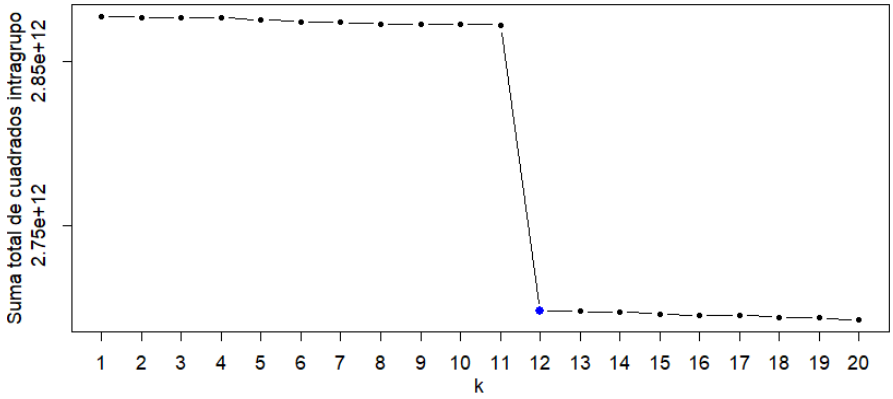

Figura B.5: Gráfico de sedimentación del clustering jerárquico de enlace mínimo a la generación eléctrica fotovoltaica.

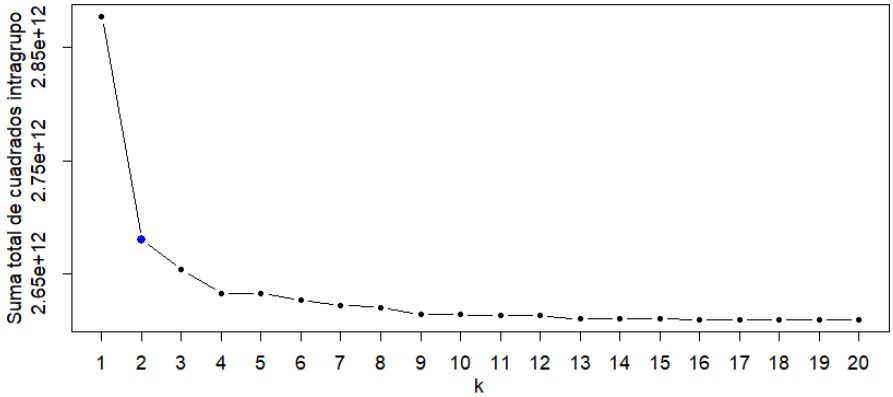

Figura B.6: Gráfico de sedimentación del clustering jerárquico de enlace máximo a la generación eléctrica fotovoltaica.

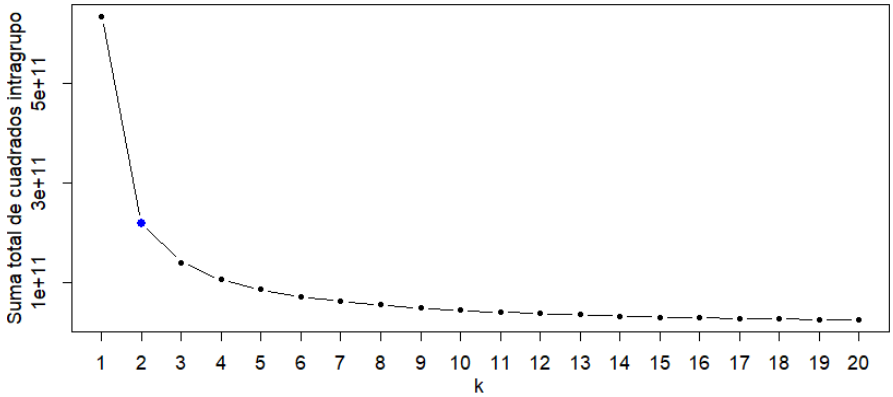

Figura B.7: Gráfico de sedimentación del clustering k-medias a la generación eléctrica fotovoltaica.

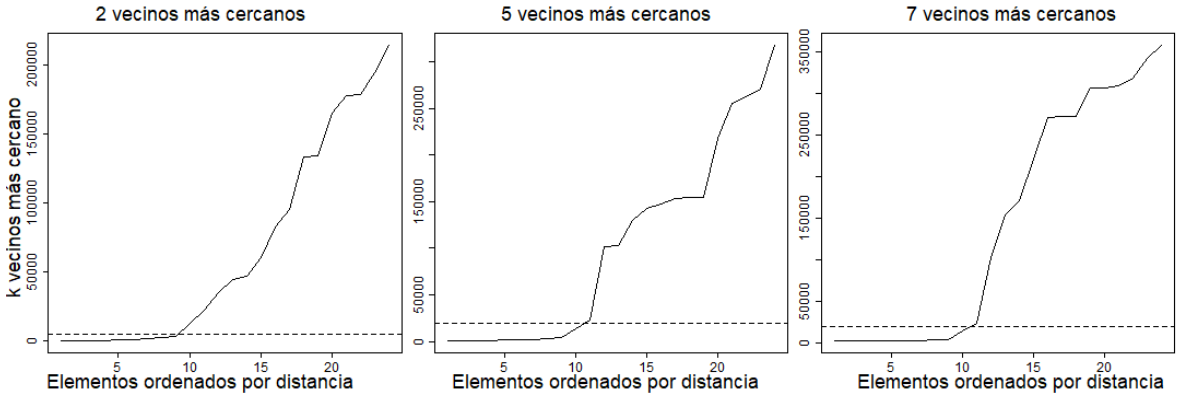

Figura B.8: Gráfico de k vecinos más cercanos del DBSCAN a la generación eléctrica fotovoltaica.

Apéndice C

Matrices de Confusión

Matrices de confusión de los resultados de los métodos de clustering a datos funcionales simulados por modelos.

Cuadro C.1: Matrices de confusión de los resultados del clustering a la simulación 1 por método.

	Enlace mínimo		**Enlace máximo**		***k*-Medias**		**Cambio medio**		**DBSCAN**		
	1	**2**	**1**	**2**	**1**	**2**	**1**	**2**	**0**	**1**	**2**
1	50	0	50	0	50	0	50	0	0	50	0
2	0	50	0	50	0	50	0	50	0	0	50

Cuadro C.2: Matrices de confusión de los resultados del clustering a la simulación 2 por método.

	Enlace mínimo			**Enlace máximo**			***k*-Medias**			**Cambio medio**			**DBSCAN**			
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**0**	**1**	**2**	**3**
1	50	0	0	50	0	0	50	0	0	50	0	0	0	50	0	0
2	0	50	0	0	50	0	0	17	33	0	50	0	0	0	50	0
3	0	0	50	0	0	50	50	0	0	0	0	50	2	0	0	48

Cuadro C.3: Matrices de confusión de los resultados del clustering a la simulación 3 por método.

	Enlace mínimo				**Enlace máximo**				***k*-Medias**				**Cambio medio**				**DBSCAN**				
	1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**0**	**1**	**2**	**3**	**4**
1	50	0	0	0	50	0	0	0	50	0	0	0	50	0	0	0	0	50	0	0	0
2	0	50	0	0	0	50	0	0	0	50	0	0	0	50	0	0	0	0	50	0	0
3	0	0	50	0	0	0	50	0	23	0	27	0	0	0	50	0	2	0	0	48	0
4	0	0	0	50	0	2	0	48	0	11	0	39	0	0	0	50	0	0	0	0	50

Bibliografía

Arango González, María Alejandra, Juan Diego Jaramillo Morales y Lucas Jaramillo Escobar (2016). "Técnicas de clustering para detectar patrones espaciales de criminalidad en jóvenes y adultos en Medellín. Octubre del 2013 a noviembre del 2014". En: *Revista Criminalidad* 58, págs. 25-45.

Brock, Guy, Vasyl Pihur, Susmita Datta y Somnath Datta (2008). "clValid: An R Package for Cluster Validation". En: *Journal of Statistical Software* 25.4, págs. 1-22. URL: https://www.jstatsoft.org/v25/i04/.

Cheng, Yizong (1995). "Mean shift, mode seeking, and clustering". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.8, págs. 790-799.

Cuesta-Albertos, J.A., M. Febrero-Bande y M. Oviedo de la Fuente (2017). "The DDG-classifier in the functional setting". En: *TEST* 26, págs. 119-142.

Cuevas, Antonio, Manuel Febrero y Ricardo Fraiman (2007). "Robust estimation and classification for functional data via projection-based depth notions". En: *Computational Statistics* 22.3, págs. 481-496.

Dubes, Ricard (1993). "Cluster Analysis and Related Issues". En: *Pattern Recognition and Computer Vision*. Lansing, Michigan, págs. 3-32.

Ester, Martin, Hans-Peter Kriegel, Jorg Sander, Xiaowei Xu et al. (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise." En: *kdd*. Vol. 96. 34, págs. 226-231.

Febrero-Bande, Manuel (2021). *Transparencia de Análisis Exploratorio de Datos Funcionales*. Universidad de Santiago de Compostela. La Coruña, España. URL: https://www.usc.gal/es/departamento/estadistica-analisis-matematico-optimizacion/directorio/manuel-febrero-bande-291.

Febrero-Bande, Manuel y Manuel Oviedo De La Fuente (2023). *Funciones de clustering Mean-Shift y DBSCAN para datos funcionales*. Comunicación personal.

Febrero-Bande, Manuel y Manuel Oviedo de la Fuente (2012). "Statistical Computing in Functional Data Analysis: The R Package fda.usc". En: *Journal of Statistical Software* 51.4, págs. 1-28. URL: https://www.jstatsoft.org/v51/i04/.

Ferraty, Frédéric y Philippe Vieu (2006). *Nonparametric Functional Data Analysis*. New York, NY: Springer.

Fukunaga, Keinosuke y Larry D. Hostetler (1975). "The estimation of the gradient of a density function, with applications in pattern recognition". En: *IEEE Trans. Inf. Theory* 21, págs. 32-40.

Hastie, Trevor, Andreas Buja y Robert Tibshirani (1995). "Penalized Discriminant Analysis". En: *The Annals of Statistics* 23.1, págs. 73-102.

Hastie, Trevor, Robert Tibshirani y Jerome Friedman (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc.

Hennig, Christian (2023). *fpc: Flexible Procedures for Clustering*. R package version 2.2-10. URL: https://CRAN.R-project.org/package=fpc.

Jain, Anil K. y Richard C. Dubes (1988). *Algorithms for Clustering Data*. New Jersey: Prentice Hall.

Lance, G.N. y W.T. Williams (1967). "A general theory of classificatory sorting strategies ii. clustering systems". En: *The computer journal* 10.3, págs. 271-277.

MacQueen, J. B. (1967). "Some Methods for Classification and Analysis of MultiVariate Observations". En: *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*. Ed. por L. M. Le Cam y J. Neyman. Vol. 1. University of California Press, págs. 281-297.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, Mia Hubert y Kurt Hornik (2022). *cluster: Cluster Analysis Basics and Extensions*. URL: https://CRAN.R-project.org/package=cluster.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Ramsay, J. O. y B. W. Silverman (2005). *Functional Data Analysis*. New York, NY: Springer.

Red Eléctrica de España, S. A. U. (2022). *ESIOS Red Eléctrica*. URL: https://www.esios.ree.es/es.

Selim, Shlomo Z. y Mohamed A. Ismail (1984). "K-Means Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality". En: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, págs. 81-87.

Sheather, Simon y M. Jones (ene. de 1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". En: *Journal of the Royal Statistical Society. Series B. Methodological* 53, págs. 683-690.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. Monographs on Statistics and Applied Probability. London: Chapman & Hall/CRC.

Walesiak, Marek y Andrzej Dudek (2020). "The Choice of Variable Normalization Method in Cluster Analysis". En: *Education Excellence and Innovation Management: A 2025 Vision to Sustain Economic Development During Global Challenges*. Ed. por Khalid S. Soliman. International Business Information Management Association (IBIMA), págs. 325-340.

Wang, Jane-Ling, Jeng-Min Chiou y Hans-Georg Müller (2015). "Functional data analysis". En: *Proceedings of the National Academy of Sciences*, págs. 1-41.

Webster, AJ, K Gaitskell, I Turnbull, BJ Cairns y R Clarke (2021). "Characterisation, identification, clustering, and classification of disease". En: *Sci Rep* 11, pág. 5405.

Índice de cuadros

Índice de figuras