Trabajo Fin de Máster

Estimación no paramétrica de conjuntos de nivel

Andrea González López

Máster en Técnicas Estadísticas

Curso 2025-2026

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación no paramétrica de conxuntos de nivel
Título en español: Estimación no paramétrica de conjuntos de nivel
English title: Nonparametric estimation of level sets
Modalidad: Modalidad A
Autor/a: Andrea González López, Universidad de Santiago de Compostela
Director/a: Paula Saavedra Nieves, Universidad de Santiago de Compostela; Alberto Rodríguez Casal, Universidad de Santiago de Compostela
Breve resumen del trabajo: En este trabajo abordaremos el problema de estimación no paramétrica de conjuntos de nivel revisando distintas metodologías propuestas en la literatura. En este contexto, los clusters se definen como componentes conexas de los conjuntos de nivel y uno de los objetivos será estimar el número de componentes conexas a partir de las estimaciones de conjuntos. Para analizar el comportamiento de los estimadores, se realiza un estudio de simulación y, finalmente, se presenta una aplicación sobre datos reales de seísmos en Grecia.
Recomendaciones: Cursar previamente Estadística No Paramétrica
Otras observaciones:

Don/doña Paula Saavedra Nieves, categoría 1 de la Universidad de Santiago de Compostela, don/doña Alberto Rodríguez Casal, categoría 2 de la Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Estimación no paramétrica de conjuntos de nivel

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

En Santiago de Compostela, a 16 de julio de 2026.

SAAVEDRA NIEVES PAULA - 44847364D
Firmado digitalmente por SAAVEDRA NIEVES PAULA - 44847364D
Fecha: 2026.07.16 11:54:10 +02'00'

El/la director/a:
Don/doña Paula Saavedra Nieves

RODRIGUEZ CASAL, ALBERTO (AUTENTICACIÓN)
Firmado digitalmente por RODRIGUEZ CASAL, ALBERTO (AUTENTICACIÓN)
Nombre de reconocimiento (DN): c=ES, serialNumber=34896829X, sn=RODRIGUEZ, givenName=ALBERTO, cn=RODRIGUEZ CASAL, ALBERTO (AUTENTICACIÓN)
Fecha: 2026.07.16 11:56:30 +02'00'

El/la director/a:
Don/doña Alberto Rodríguez Casal

GONZALEZ LOPEZ ANDREA - 45161256C
Firmado digitalmente por GONZALEZ LOPEZ ANDREA - 45161256C
Fecha: 2026.07.15 22:22:39 +02'00'

El/la autor/a:
Don/doña Andrea González López

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/la autor/a 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

Agradezco a mis tutores, Paula y Alberto, por dedicarme su tiempo durante estos dos años, entender las circunstancias y ayudarme en todo momento confiando en que podíamos sacar esto adelante de la mejor manera posible.

A mi familia por animarme a terminar el máster y no dejar a medias este trabajo. Una cosa más hecha.

A mis amigos y amigas que tantas veces me han preguntado si este trabajo ya estaba terminado. ¡Por fin voy a poder decir que sí!

A Martín porque sin su ordenador funcionando 24 horas al día las simulaciones no habrían sido posibles. Te lo devolveré pronto.

A Álvaro por acompañarme tantas tardes en Madrid después de trabajar estos últimos meses, aguantar mis agobios y darme siempre un poco de perspectiva: las cosas nunca son tan horribles como las pinto.

Índice general

Resumen xi

1. Introducción 1

1.1. Presentación de los datos . . . 4

1.2. Preliminares matemáticos . . . 6

2. Estimación no paramétrica de conjuntos de nivel 11

2.1. Plug-in . . . 13

2.1.1. Efecto del parámetro de suavización . . . 14

2.1.2. Estimación del número de componentes conexas . . . 16

2.2. Exceso de masa . . . 22

2.3. Métodos híbridos . . . 23

2.3.1. Método de la envoltura convexa . . . 24

2.3.2. Método de la envoltura $r-$convexa . . . 24

2.3.3. Método del suavizado granulométrico . . . 30

3. Estudio de simulación 35

4. Aplicación a datos reales 45

4.1. CFF . . . 46

4.2. pdfCluster . . . 49

4.3. Método de la envoltura r-convexa . . . 51

4.4. Método del suavizado granulométrico . . . 53

4.5. Conclusiones . . . 55

A. Funciones Kernel unidimensionales y multidimensionales 57

Bibliografía 59

Resumen

Resumen en español

En este trabajo abordaremos el problema de estimación no paramétrica de conjuntos de nivel desde diferentes perspectivas. Comenzaremos con una revisión bibliográfica de algunas de las metodologías propuestas en la literatura para la reconstrucción de conjuntos de nivel partiendo de una muestra aleatoria simple generada a partir de una función de densidad desconocida. En este contexto, los clusters son definidos como las diferentes componentes conexas que forman los conjuntos de nivel y uno de los objetivos principales del trabajo será estimar el número de componentes conexas de los conjuntos de nivel poblacionales a través del número de componentes conexas de las estimaciones obtenidas a partir de los métodos revisados. Posteriormente, se realizará un estudio de simulación sobre distintos modelos de datos con distribución conocida para analizar el comportamiento de las estimaciones del número de componentes conexas. Finalmente, se presenta una aplicación de los diferentes métodos de estimación sobre dos conjuntos de datos reales de seísmos que han tenido lugar en Grecia.

English abstract

In this paper, we will address the problem of non-parametric estimation of level sets from different perspectives. We will begin with a literature review of some of the methodologies proposed in the literature for the reconstruction of level sets based on a simple random sample generated from an unknown density function. In this context, clusters are defined as the different connected components that make up the level sets, and one of the main objectives of this paper will be to estimate the number of connected components of the population level sets based on the number of connected components in the estimates obtained using the methods reviewed. Subsequently, a simulation study will be carried out on various data models with known distributions to analyse the behaviour of the estimates of the number of connected components. Finally, an application of the different estimation methods is presented using two real-world datasets of earthquakes that have occurred in Greece.

Capítulo 1

Introducción

Hay muchas ocasiones donde nos interesa tener una idea de cómo se distribuye un cierto fenómeno en el espacio de cara a hacer prevención o para saber como repartir unos ciertos recursos, como por ejemplo dónde se dan las mayores lluvias o en qué zonas se detectan el mayor número de temblores terrestres. Por eso es relevante saber las zonas más afectadas o, visto de otro modo, con mayor densidad de sucesos mediante estimaciones de conjuntos de nivel. Un ámbito donde es importante la detección de zonas de alta densidad es para ver los puntos más afectados por terremotos ya que construir sin tomar medidas preventivas en zonas con alta actividad sísmica podría dañar infraestructuras y poner en peligro muchas vidas.

Un terremoto o seísmo es un temblor en la tierra que se produce como resultado de un movimiento global de las placas tectónicas. Todo el subsuelo terrestre está dividido en diferentes placas tectónicas tal como se comenta en Bird (2003) y se puede observar en la Figura 1.1. Estas se encuentran en constante desplazamiento, por lo general de forma lenta, lo cual da lugar a choques suaves. Pero no siempre ocurre de esta forma. En ocasiones, cuando dos placas chocan, una de ellas se desliza por encima de la otra de forma que se produce un terremoto. Debido a ello es esperable que las zonas con mayor concentración de seísmos sean aquellas donde limitan dos placas téctonicas. Es decir, si nos fijamos en la Figura 1.1, serían los lugares en los que aparece una línea negra. Otras zonas con abundantes terremotos se encuentran donde existe una falla, esto es, en lugares en los que haya tenido lugar una rotura de las placas. Esta situación es otra de las causas de los seísmos, ya que al haber una división de placas pueden ocurrir desplazamientos como los comentados. Se puede consultar Kanamori y Brodsky (2004) para ampliar la información sobre esta temática.

La Geología es la ciencia encargada de estudiar los terremotos, siendo estos uno de los campos de investigación más interesantes debido a que ser capaces de predecirlos ayudaría a que los efectos negativos de los seísmos fueran menores. Como se ha comentado previamente, las zonas más afectadas son aquellas que se encuentran en el límite de dos placas como es el caso de Grecia, país en el que se centrará el estudio de los terremotos en este trabajo.

Grecia es un país formado por multitud de islas y, debido a la particularidad de su situación tectónica, se han producido un gran número de seísmos de diferentes intensidades a lo largo de los años. Por ello, Grecia se sitúa como el país europeo y el sexto a nivel mundial que más daños sufre debido a terremotos. Tal como se ve en la Figura 1.2[1], este país se encuentra sobre dos placas tectónicas:

[1] Las Figuras 1.1 y 1.2 se han elaborado a partir de bases de datos que se encuentran en distintas librerías del software libre R. Concretamente se ha empleado la base `worldMapEnv` del paquete `maps` (Becker et al, 2023) para el mapa mundial

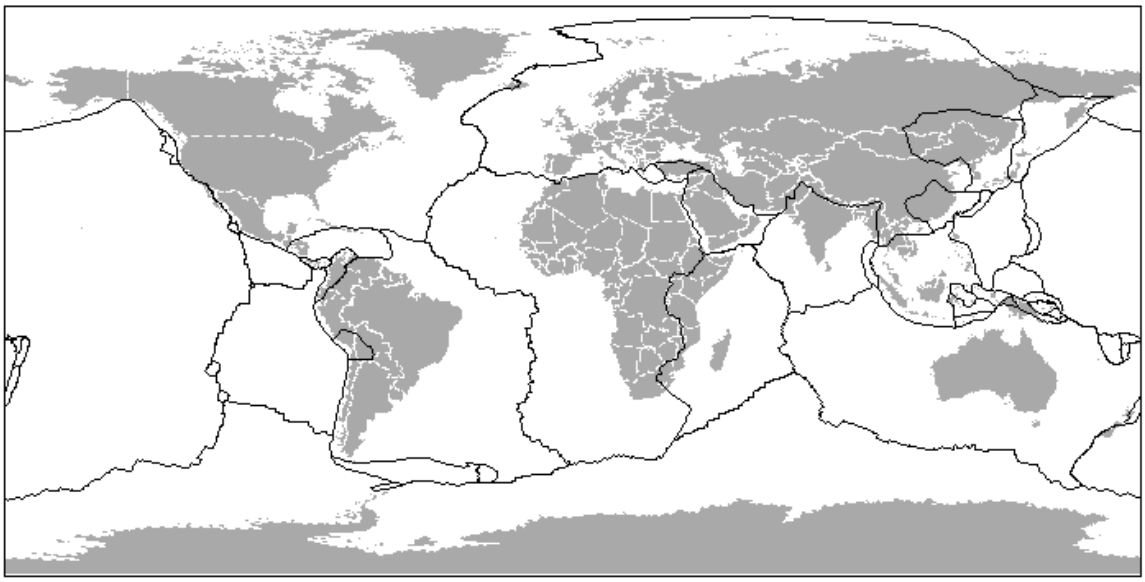

Figura 1.1: División en placas tectónicas (negro) de la superficie terrestre.

la euroasiática en la parte superior y la del mar Egeo cubriendo la mayoría de las islas que forman el país, tal como se detalla en Bird (2003). Así, se puede observar que el país se encuentra muy cercano a cuatro límites entre placas: la euroasiática con la del mar Egeo en el norte, la euroasiática con la africana en el noroeste, la africana con la del mar Egeo en el sur y la del mar Egeo con la Anatolia en el este. De esta forma se tiene una justificación geológica para la numerosa cantidad de seísmos que ocurren en este país en base a lo comentado anteriormente.

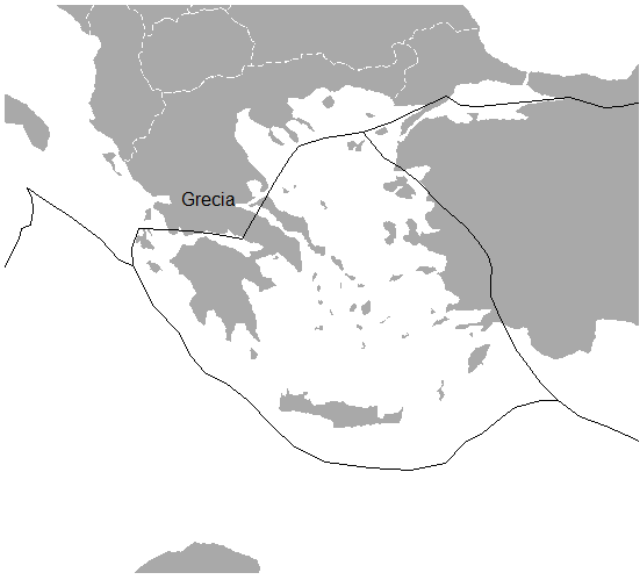

Figura 1.2: División en placas tectónicas (negro) en la superficie de Grecia y alrededores.

y `platesf` del paquete `ks` (Duong, 2024) para la reconstrucción de las placas tectónicas.

En el presente trabajo se obtendrán estimaciones de los conjuntos de nivel para distintos umbrales de la función de densidad de los terremotos de magnitud en escala de Richter mayor que cuatro que han tenido lugar desde 1972 hasta 1992 (observaciones históricas) y desde el 2003 hasta 2023 (observaciones actuales), donde todos los años mencionados están incluidos. El objetivo será contabilizar las zonas de alta concentración de terremotos tras la reconstrucción de los conjuntos de nivel con distintos métodos y observar si el patrón es estable a lo largo del tiempo empleando los dos periodos temporales en estudio comentados. Los datos que se van a utilizar se presentan y detallan en la siguiente sección.

Dada una función de densidad f de un vector aleatorio X con valores en $\mathbb{R}^d$, se define un conjunto de nivel de un cierto nivel $t > 0$ como un subconjunto del dominio donde la densidad supera o iguala ese valor específico t. Esto es, dado un nivel $t > 0$ definimos el conjunto de nivel como:

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}.$$

Elegir el valor de t puede resultar complicado en la práctica. Como alternativa podemos fijar un contenido en probabilidad mínimo $1-\tau$ con $\tau \in (0,1)$ que debe tener el conjunto de nivel con respecto a la distribución inducida por f, evitando de esta forma el problema de qué umbral seleccionar.

La estimación de conjuntos de nivel es una pequeña parte dentro del estudio de la reconstrucción de conjuntos a partir de una muestra de una variable o vector aleatorios. En particular, tomaremos una muestra de observaciones independientes e idénticamente distribuidas $\mathcal{X}_n$ de una variable aleatoria $d-$dimensional X absolutamente continua con función de densidad f. La geometría tiene un papel importante en este problema, el cual se irá desarrollando en capítulos posteriores. En particular, a lo largo de este trabajo se abordará el problema de forma no paramétrica ya que no habrá ninguna suposición sobre la función de densidad del vector aleatorio de partida. En función de las hipótesis que se hagan sobre la forma de los conjuntos de nivel las estimaciones se pueden obtener empleando tres metodologías distintas en las que se profundizará en el Capítulo 2.

Uno de los objetivos generales de la estimación no paramétrica de conjuntos de nivel es la búsqueda de grupos, que se denotarán por clústers, los cuales se identifican con las zonas de alta densidad de los datos. Estos grupos coinciden con las componentes conexas del conjunto de nivel tal como se comenta en Hartigan (1975). Esto relaciona el problema presentado con el aprendizaje no supervisado como se menciona en Biau et al (2007) siendo de gran utilidad en multitud de ámbitos como el estudio de enfermedades (Saavedra, 2014), fenómenos naturales o el comportamiento de especies animales (Saavedra y Crujeiras, 2022).

Los contenidos del trabajo se distribuirán de la siguiente manera. Al comienzo del Capítulo 1 se ha motivado el trabajo añadiendo una pequeña introducción de lo que se va a realizar y los objetivos de este. En la sección siguiente se presentará una base de datos con localizaciones de una especie de planta característica de Australia que se empleará para ilustrar los diferentes métodos y un segundo conjunto de datos con información de terremotos en Grecia para hacer estimaciones no paramétricas de datos reales. La última sección del presente capítulo contiene una explicación de los diferentes conceptos matemáticos que serán necesarios para el desarrollo del Capítulo 3, incluyendo las definiciones de condiciones geométricas y distancias entre conjuntos.

En el Capítulo 2 se realiza una revisión bibliográfica de las distintos métodos que se han desarrollado en la literatura para la estimación no paramétrica de los conjuntos de nivel. Se presentarán las tres formas de estimación principales con sus suposiciones y dentro de ellas algunos de los métodos de cálculo de componentes conexas más destacados y utilizados.

En el Capítulo 3 se harán simulaciones con distintos tamaños de muestra de cuatro modelos en los cuales conocemos el número de componentes conexas que hay para conjuntos de nivel con distintos

niveles. Así podremos testear el funcionamiento de la metodología revisada para la estimación del número de grupos comprobando así el rendimiento de los métodos en diferentes escenarios

Finalmente, en el Capítulo 4 se aplicarán los diferentes métodos sobre datos reales como son las coordenadas donde tienen lugar los seísmos más fuertes en Grecia en dos intervalos temporales. Con eso se verá su evolución con el paso del tiempo mediante la comparación de los conjuntos de nivel.

1.1. Presentación de los datos

En lo que resta de capítulo y el Capítulo 2 se introducen diferentes definiciones y se presentan las metodologías de la estimación no paramétrica de conjuntos de nivel además de las bases de datos que mencionamos en la sección anterior. Para ilustrar y aclarar los conceptos y resultados se empleará una base de datos denominada `grevillea` extraída de la librería `ks` de R (Duong, 2024). Esta contiene 222 observaciones de una planta conocida como *Grevillea uncinulata*, una especie propia del suroeste de Australia. En la base de datos se incluyen la longitud y latitud de las distintas plantas observadas medidas en coordenadas decimales simples en el año 2016. Una primera visualización de los datos se muestra en la Figura 1.3 donde se muestra la localización de cada observación en todo el territorio de Australia (izquierda), y restringidos únicamente al suroeste del país (derecha).

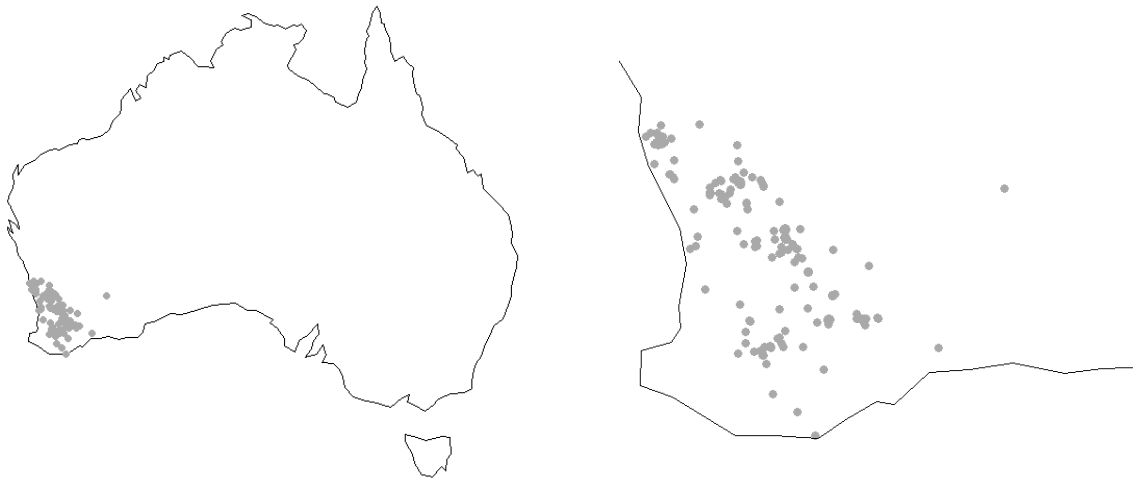

Figura 1.3: Mapa de Australia con las observaciones de *Grevillea uncinulata* (izquierda) y zoom sobre la zona propia (derecha).

Además de los datos anteriores, en el Capítulo 4 se tratará de estimar el número de componentes conexas de conjuntos de nivel para distintos a través de la reconstrucción de las curvas de nivel de dos muestras de datos reales relativos a terremotos en la zona de Grecia empleando diferentes métodos y con la ayuda del software libre R. La base elegida se ha encontrado en *kaggle*[2] cuyos datos han sido actualizados en Junio de 2024 y la versión descargada es de Octubre de 2024. Este conjunto de datos recoge como observaciones los 285.118 seísmos de diferentes magnitudes que han ocurrido en Grecia desde 1965 hasta 2023 (ambos incluidos) y para cada uno de ellos se tiene registrada la siguiente información:

[2]Se puede acceder a ella mediante el siguiente enlace https://www.kaggle.com/datasets/nickdoulos/greeces-earthquakes.

- `DATETIME`: Hora, día, mes y año en que ha sucedido.
- `LAT` y `LONG`: Latitud y longitud en coordenadas decimales simples del lugar donde ha tenido lugar.
- `DEPTH`: Profundidad (en kilómetros) a la que ha ocurrido.
- `MAGNITUDE`: Magnitud medida en escala de Richer del terremoto.

Cada observación se trazará sobre el plano empleando la latitud, representada en el eje de abscisas, y la longitud, representada en el eje de ordenadas. Así, se trabajará sobre un espacio bidimensional euclídeo donde los conjuntos de nivel se denotarán por curvas de nivel. Para tener una idea más realista de las ubicaciones de los terremotos, las observaciones se mostrarán sobre un mapa con la frontera de Grecia, el cual se presenta en la parte derecha de la Figura 1.4. Se ve la coincidencia de estos bordes del país con el mapa real de Grecia que se recoge a la izquierda de la Figura 1.4[3]. Para dibujar la frontera de Grecia se ha extraído una base de datos en formato *gjson* del sitio web `https://gisdata.mapog.com/` y con R se ha trazado el mapa que se usará a lo largo de todo el trabajo.

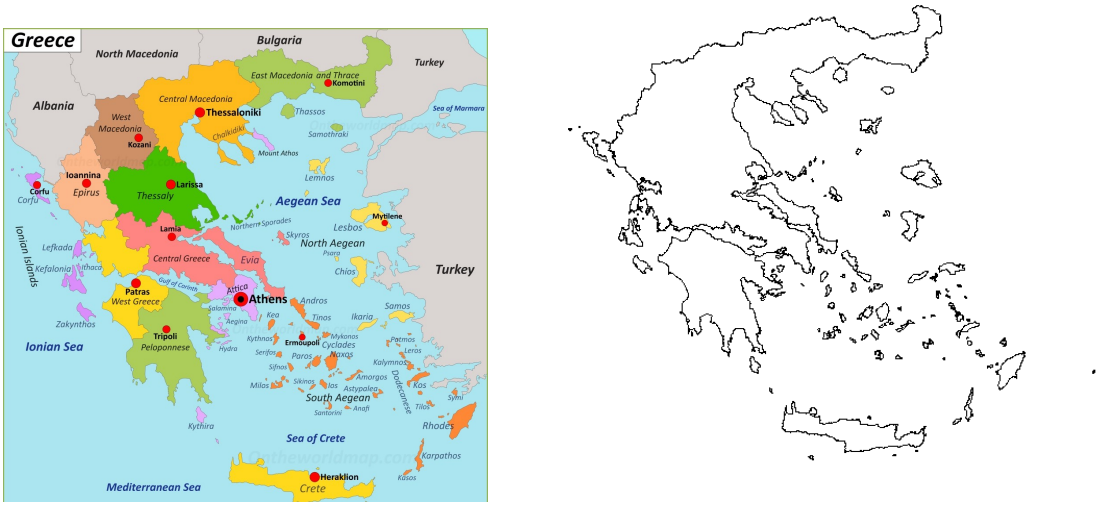

Figura 1.4: Mapa político (izquierda) y frontera (derecha) de Grecia.

La base de datos que se ha introducido contiene observaciones de seísmos de todas las magnitudes posibles pero en este trabajo se restringirá el análisis a aquellos terremotos de magnitud mayor que cuatro. Se ha seleccionado ese límite inferior porque son los seísmos de más de cuatro en escala Richter aquellos que pueden generar daños y son sentidos por las personas. El conjunto de datos resultante se dividirá, como dijimos, en dos subconjuntos: el primero tomando las observaciones que han tenido lugar entre 1972 y 1992 y el segundo con los terremotos ocurridos entre 2003 y 2023. De esta forma, pretendemos chequear si hay un cambio de patrón en la situación de los terremotos entre finales del siglo pasado y las dos primeras décadas de este. En la Figura 1.5 se observan los seísmos del conjunto de datos históricos, en azul, y del conjunto de datos actuales, en gris. Las zonas más afectadas en ambos casos parecen las islas de la parte oeste del país, conocidas como islas jónicas y las del sureste de Grecia.

[3]La imagen de la parte izquierda se ha extraído en la página web `https://ontheworldmap.com/greece/greece-regions-and-capitals-map.html` con último acceso a 18 de octubre de 2024.

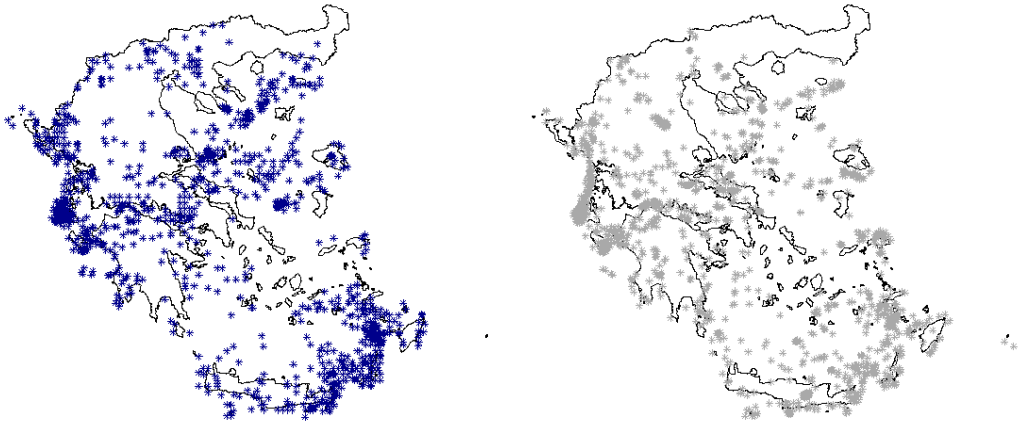

Figura 1.5: Observaciones de terremotos de magnitud mayor que cuatro entre 1972 y 1992 (izquierda) y entre 2003 y 2023 (derecha).

Las estimaciones de las curvas de nivel con distintos umbrales junto con un mayor análisis de los datos se realizarán en el Capítulo 4. Se tomarán valores de $\tau = 0.3, 0.5, 0.8$ para tener una representación de niveles más altos y bajos y ver la evolución de la distribución de los seísmos en los dos períodos mencionados.

1.2. Preliminares matemáticos

En los procedimientos de estimación de conjuntos es natural suponer condiciones geométricas. Algunas de las restricciones de forma más empleadas se presentan en esta sección para su utilización en los capítulos posteriores. Se define también el concepto de componente conexa ya que la estimación del número de estas será la finalidad tras la construcción de los conjuntos de nivel en los Capítulos 3 y 4. Se introduce además en cada definición la notación que se empleará a lo largo del trabajo para hacer referencia a cada concepto. Además, dado un conjunto S, denotaremos a la frontera del conjunto por ∂S, a su interior por $\text{Int}(S)$ y a su complementario por S^c.

Definición 1.1 *Se define una bola abierta de centro a y radio r en el espacio $\mathbb{R}^d$ como el conjunto de puntos que están a una distancia de a menor que el radio. Es decir, $B(a,r) = \{x \in \mathbb{R}^d / d(x,a) < r\}$, donde $d(\cdot,\cdot)$ es una distancia definida sobre el espacio $\mathbb{R}^d$. Una bola cerrada se define de manera análoga con la única diferencia que los puntos que pertenecen a ella son aquellos que están a distancia menor o igual que el radio. Esto es, $B[a,r] = \{x \in \mathbb{R}^d / d(x,a) \leq r\}$.*

A lo largo de este trabajo se empleará la distancia euclídea sobre $\mathbb{R}^2$ que se define a continuación:

Definición 1.2 *Dados dos puntos $x = (x_1, x_2)$ e $y = (y_1, y_2)$ en $\mathbb{R}^2$, se define la distancia euclídea entre los dos puntos en el plano como:*

$$D_2(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

Introducimos también la distancia Hausdorff entre dos conjuntos que se empleará en el Capítulo 2 para el cálculo del parámetro óptimo en el método de suavizado granulométrico.

Definición 1.3 *Sean dos conjuntos $S_1, S_2 \in \mathbb{R}^d$ compactos y no vacíos, se define la distancia Hausdorff como:*

$$d_H(S_1, S_2) = \text{máx}\{\sup_{s_1 \in S_1} d(s_1, S_2), \sup_{s_2 \in S_2} d(s_2, S_1)\}$$

donde dado un punto $s_1 \in \mathbb{R}^d$ y un subconjunto no vacío $S_2 \subset \mathbb{R}^d$, $d(s_1, S_2) = \text{ínf}\{D_2(s_1, s_2), s_2 \in S_2\}$.

Una interpretación de esta distancia es como la máxima distancia euclídea que existe entre un punto de un conjunto para desplazarse al otro conjunto. Notemos así que la distancia Hausdorff entre dos conjuntos compactos siempre será mayor que cero a no ser que los dos conjuntos sean exactamente el mismo.

Además, también necesitaremos para el método del suavizado granulométrico la suma de Minkowski que presentamos a continuación:

Definición 1.4 *Dado un subconjunto $A \subset S$ y compactos y no vacíos, se define la suma de Minkowski se define como:*

$$A \oplus rB = \bigcup_{x \in A} B[x, r] = \{x \in S : B[x, r] \not\subset A^c\}$$

Una de las condiciones geométricas que se pedirán sobre los conjuntos de nivel en los siguientes capítulos es la convexidad definida a continuación:

Definición 1.5 *Un conjunto $S \subseteq \mathbb{R}^d$ se dice que es convexo si para todo $\lambda \in [0, 1]$ y para cada par de puntos $\mathbf{x}, \mathbf{y} \in S$ se cumple que $\lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$ también pertenece a S.*

Esto es, todas las combinaciones convexas de dos puntos de S pertenecen al propio conjunto o, visto de otra forma, si se unen dos puntos cualesquiera de S con una línea recta, esta se encuentra contenida en su totalidad dentro de S. En la Figura 1.6 se muestran ejemplos de conjuntos convexos (los dos primeros) y no convexos (los dos últimos).

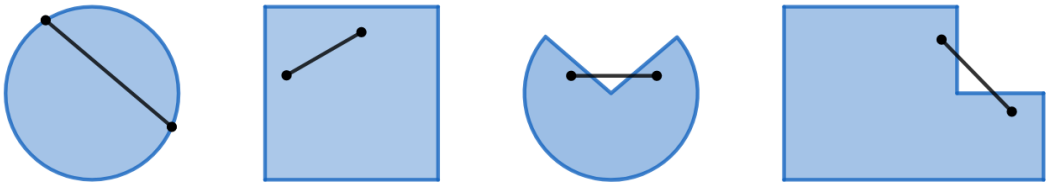

Figura 1.6: Ejemplos de conjuntos convexos (parte izquierda) y de conjuntos no convexos (parte derecha).

Definición 1.6 *Dado un conjunto $S \subseteq \mathbb{R}^d$, se define la envoltura convexa de S como el menor conjunto convexo que contiene a S y se denota por conv(S).*

Siempre se verifica que $S \subseteq \text{conv}(S)$ cumpliéndose únicamente que $S = \text{conv}(S)$ cuando S es convexo. En la Figura 1.7 se muestra las envolturas convexas de los conjuntos no convexos de la Figura 1.6 donde se ha añadido la zona rayada al propio conjunto. Para los conjuntos convexos, la envoltura convexa es el propio conjunto.

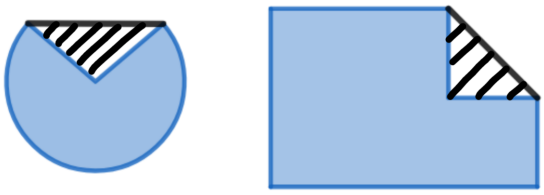

Figura 1.7: Envoltura convexa de los conjuntos no convexos de la Figura 1.6 donde se ha añadido la zona rayada en cada caso.

La condición de convexidad es difícil que se cumpla en muchas situaciones reales, especialmente si esperamos que haya varios grupos de observaciones. La $r-$convexidad, definida a continuación, es una condición más flexible y aplicable.

Definición 1.7 *Dado un conjunto cerrado $S \subseteq \mathbb{R}^d$ y un radio $r > 0$, definimos la envoltura $r-$convexa de S como:*

$$C_r(S) = \bigcap_{\{B(x,r):B(x,r)\cap S=\emptyset\}} B(x,r)^c,$$

donde recordemos que $B(x,r)$ es la bola abierta de centro x y radio r y $B(x,r)^c$ su complementario.

Esta definición no es más que una intersección de complementarios de bolas abiertas de un radio dado $r > 0$. En $\mathbb{R}^2$ los bordes de la envoltura están formados por arcos de ciruferencias de radio r donde los centros de estas se encuentran fuera de la envoltura.

Definición 1.8 *Un conjunto cerrado $S \subseteq \mathbb{R}^d$ se dice que es $r-$convexo para un cierto radio $r > 0$ si el propio conjunto coindice con su envoltura $r-$convexa, esto es: $S = C_r(S)$.*

Notemos que un conjunto convexo es también $r-$convexo para cualquier $r > 0$ pero no al contrario. En la Figura 1.8 se muestra un conjunto que es $r-$convexo para valores de r menores que r' pero que sin embargo no es convexo.

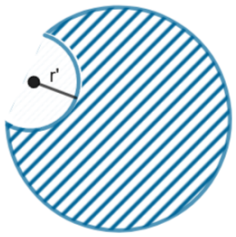

Figura 1.8: Ejemplo de un conjunto no convexo pero sí $r-$convexo para valores de r más pequeños que r'.

En  se usa la librería `alphahull` (Pateiro-López y Rodríguez-Casal, 2022) para el cálculo de la envoltura $r-$convexa, pudiendo elegir cualquier valor del radio. la Figura 1.9 podemos ver en la

izquierda la envoltura convexa de `grevillea` quitado un outlier en la parte superior derecha, y en la parte derecha de la figura, la envoltura $r-$convexa para tres valores de r: $r = 1.8$ en rojo, $r = 3$ en verde y $r = 8$ en azul. Así vemos que a mayor valor de r, más se parece la envoltura $r-$convexa a la envoltura convexa.

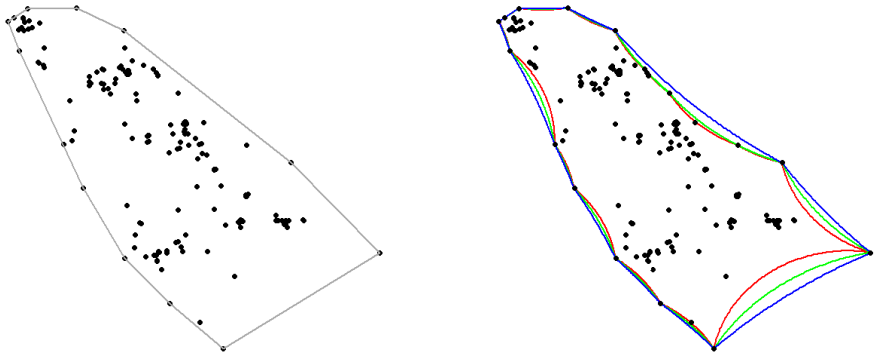

Figura 1.9: Envoltura convexa de la base de datos de `grevillea` (izquierda) y envoltura $r-$convexa con valores de $r = 1.8$ (rojo), $r = 3$ (verde) y $r = 8$ (azul) .

La última condición geométrica que se va a emplear en los métodos híbridos es la siguiente:

Definición 1.9 *Sea $S \subset \mathbb{R}^d$ un conjunto cerrado y un valor $r > 0$. Se dice que una bola de radio r rueda libremente en S si para cada punto frontera $b \in \partial S$ existe un $x \in \mathbb{R}^d$ tal que $b \in B[x, r] \subset S$.*

En la Figura 1.10 tenemos un ejemplo de un conjunto que verifica la condición de que existe una bola que rueda libre a la izquierda, mientras que a la derecha tenemos un conjunto para el que no se cumple esta propiedad debido a que los bordes no son suaves. Por lo tanto cuando pidamos esta condición a un conjunto no estaremos suponiendo otra cosa distinta a que el conjunto tenga bordes suaves.

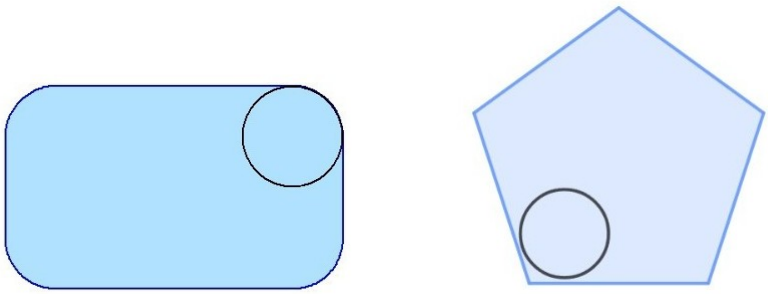

Figura 1.10: En la parte izquierda un conjunto para el que se verifica la condición de bola de radio r que rueda libre, tanto para r como para los radios menores que él y a la derecha un conjunto en el que no se cumple la condición .

Uno de nuestros objetivos será ver cuantos grupos forman el conjunto de nivel y para ello los conceptos que se definen a continuación son de gran relevancia.

Definición 1.10 *Un conjunto S es conexo si, siempre que se pueda escribir como la unión de dos conjuntos abiertos disjuntos, uno de ellos debe ser vacío.*

De forma intuitiva, un conjunto conexo es aquel que no se puede dividir en partes separadas. Ejemplos de distintos conjuntos conexos y no conexos están en la Figura 1.11: las dos de la izquierda son conexos; además, el de la parte superior es convexo mientras que el de la inferior no lo es. Por otro lado, los conjuntos de la derecha no lo son, sino que están formados a su vez por subconjuntos que definimos a continuación.

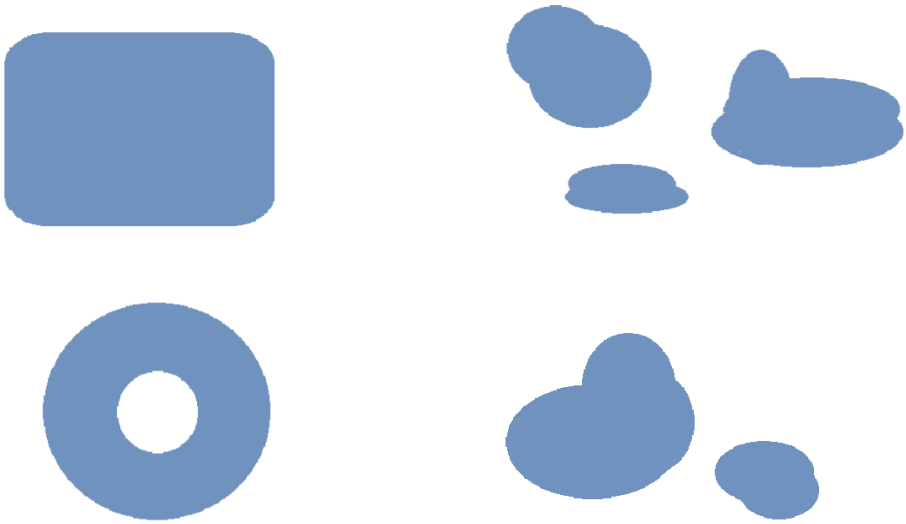

Figura 1.11: En la parte izquierda dos conjuntos conexos mientras que en la derecha dos conjuntos no conexos.

Definición 1.11 *En un conjunto S, se define una componente conexa como un conjunto maximal conexo. Esto es, si $C \subset S$ es una componente conexa, se cumple que:*

- *C es conexa.*
- *Si C_2 es conexo y $C \subset C_2$ entonces $C = C_2$*

Notemos que un conjunto es conexo si y solo si está formado por una única componente conexa. En la Figura 1.11 ya mencionamos que los dos conjuntos presentados a la izquierda son conexos y por lo tanto están formados por una componente conexa. En la parte derecha, tenemos en la zona superior un conjunto formado por tres componentes conexas y en la inferior por dos.

La obtención del número de componentes conexas está muy relacionada con la obtención de clusters y la detección de modas por lo que es de gran interés la estimación de estas en numerosos ámbitos. Por ejemplo, si sabemos que hay una única componente conexa, podemos evitar el uso de metodologías computacionalmente costosas y emplear un camino más sencillo que nos proporcione buenos resultados. Debido a la importancia de la obtención del número de componentes conexas, en este trabajo se proporcionan varias formas de estimarlas.

Capítulo 2

Estimación no paramétrica de conjuntos de nivel

En este capítulo se revisarán distintos métodos que se han desarrollado en la literatura para la estimación no paramétrica de conjuntos de nivel. En cada sección se profundiza en una de las tres metodologías existentes para resolver el problema mencionado. Para ilustrar los conceptos y métodos que se incluyen se empleará la base de datos `grevillea` presentados en el capítulo anterior. Las referencias principales que se han seguido para el desarrollo de este capítulo son Saavedra (2014), Rodríguez-Casal y Saavedra (2022), Cuevas et al (2000), Azzalini y Menardi (2014) y Bolón et al (2024).

Para la estimación no paramétrica de conjuntos de nivel se parte de un conjunto de observaciones independientes e idénticamente distribuídas (iid) $\mathcal{X}_n = \{X_1, ...X_n\}$ de una variable aleatoria $d-$dimensional X absolutamente continua con función de densidad f. En ocasiones, hay zonas del soporte de X que apenas tiene observaciones por lo que estimar este puede no tener gran interés. Por ello, el objeto de estudio será la estimación de conjuntos de nivel para reconstruir aquellas zonas donde la densidad de datos es substancial. Algunas referencias de la estimación del soporte son Rényi y Sulanke (1963), Rényi y Sulanke (1964) y Devroye y Wise (1980) ya que esta temática no se tratará en este trabajo.

Recordemos el concepto de conjunto de nivel. Dado un nivel $t > 0$ se define el conjunto de nivel t como

$$G(t) = \{x \in \mathbb{R}^d : f(x) \geq t\}. \tag{2.1}$$

Es decir, el conjunto de nivel t de una función de densidad f es un subconjunto del dominio donde f supera el valor específico t.

Como el rango de valores razonables de t no siempre es conocido se presenta una definición alternativa en la cual se fija un contenido en probabilidad mínimo $1 - \tau$ con $\tau \in (0, 1)$ que debe tener el conjunto de nivel. Así, tal como se menciona en Hyndman (1996), dado un valor $\tau \in (0, 1)$, se define la región de alta densidad como:

$$\mathcal{L}(\tau) = \{x \in \mathbb{R}^d : f(x) \geq f_\tau\}. \tag{2.2}$$

donde

$$f_\tau = \sup\left\{ y \in (0,\infty) : \int_{G(y)} f(t)dt \geq 1-\tau \right\}.$$

Entonces f_τ es el mayor valor no negativo que verifica $\mathbb{P}(X \in \mathcal{L}(\tau)) \geq 1-\tau$ siendo esta última la probabilidad inducida por la función de densidad f. Nótese que dado un valor $\tau \in (0,1)$, $\mathcal{L}(\tau)$ se puede reescribir como $G(f_\tau)$.

Claramente los conjuntos de nivel serán diferentes en función del valor de t (de igual manera sucede con τ): a mayor valor de t, equivalentemente a mayor valor de τ, el conjunto de nivel será más pequeño ya que el contenido en probabilidad requerido será menor. Lo contrario ocurre para valores pequeños de t y τ. Esto se ilustra en la Figura 2.1 con los datos de `grevillea` donde se muestran estimaciones de conjuntos de diferente nivel. A la izquierda se ha tomado el valor $\tau = 0.1$, en el centro $\tau = 0.5$ y a la derecha $\tau = 0.8$. Se ve de esta forma que al aumentar el valor de τ se obtienen conjuntos de nivel más pequeños. Por tanto, para valores más altos de τ se estiman las modas más relevantes; con τ más baja, estamos más cerca al soporte efectivo.

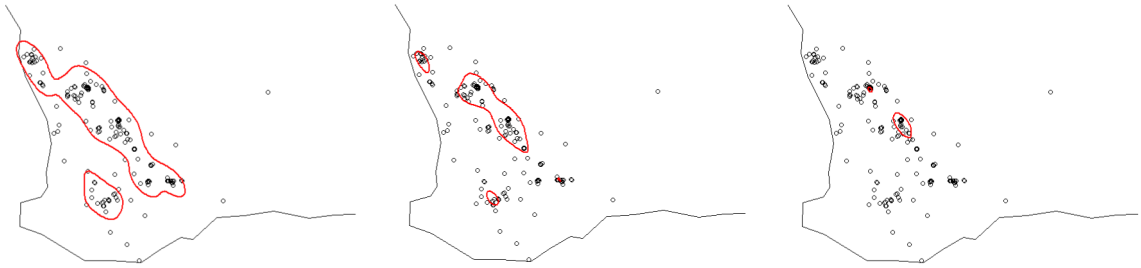

Figura 2.1: Estimación de los conjuntos de nivel para los datos de `grevillea` para distintos valores de τ: $\tau = 0.1$ (izquierda), $\tau = 0.5$ (centro) y $\tau = 0.8$ (derecha).

El número de componentes conexas también varía en función del nivel pero en este caso no existe una relación monótona entre el valor de τ y el número de componentes conexas en el sentido de que a mayor τ estas aumenten o disminuyan. Puede verse esto de nuevo en las Figuras 2.1 y 2.2. En la primera de ellas se muestra que para $\tau = 0.1$ y $\tau = 0.8$ se tienen dos componentes conexas mientras que para $\tau = 0.5$ hay cuatro. Una forma más sencilla de ilustrarlo es con una función de densidad unidimensional de una mixtura de dos normales como la de la Figura 2.2. Para el menor valor de τ se está considerando que no hay dos modas separadas apareciendo una única componente conexa; para $\tau = 0.5$ se consiguen ver las dos modas existentes (dos componentes conexas) y para el valor más alto de τ se capta un único grupo de la moda principal. Por tanto, se observa que no existe la relación de monotonía que se comentaba.

En la práctica solo se dispone de una muestra $\mathcal{X}_n$ y la función de densidad f es desconocida, por lo que es necesario realizar estimaciones. Para realizarlas se emplean tres metodologías que se presentan a continuación y en las que se profundiza en las diferentes secciones del presente capítulo:

- **Plug-in:** No se incluyen asunciones geométricas sobre el conjunto de nivel. Se trata de estimar f usando un estimador no paramétrico f_n, generalmente tipo kernel, que depende de una matriz definida positiva H que sea la matriz de suavizado. Así, se sustituye f por su estimador en la definición de conjunto de nivel en (2.1).

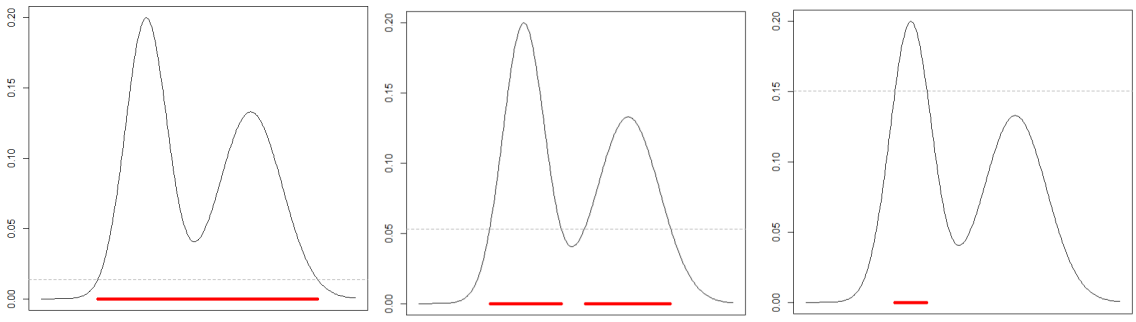

Figura 2.2: Conjuntos de nivel para la mixtura de dos normales para distintos valores de τ: $\tau = 0.3$ (izquierda), $\tau = 0.5$ (centro) y $\tau = 0.9$ (derecha).

- **Exceso de masa:** A priori se asume alguna condición geométrica del conjunto de nivel que se va a estimar y se incluye en el procedimiento de estimación, que consiste en maximizar el exceso de masa empírico (ver Sección 2.2). Aquí no hay estimación no paramétrica de la función de densidad.
- **Híbridos:** Se estima f de forma no paramétrica (como en la primera metodología) y se asume alguna condición geométrica sobre el conjunto de nivel (como en la segunda). Dentro de este grupo, en función de las restricciones geométricas que se añadan, surgen diferentes métodos.

2.1. Plug-in

Los métodos *plug-in* se emplean cuando no se tiene ninguna información a priori de la geometría del conjunto que se pretende estimar. Una idea natural para la estimación de los conjuntos de nivel es a partir de una estimación no paramétrica de la función de densidad. Esta se denotará por f_n y se sustituirá por f en las ecuaciones (2.1) o (2.2), resultando:

$$\hat{\mathcal{L}}(\tau) = \{x \in \mathbb{R}^d : f_n(x) \geq \hat{f}_\tau\}$$

donde $\hat{f}_\tau$ es un estimador del umbral a superar f_τ. En Hyndman (1996) se comentan dos formas de obtener $\hat{f}_\tau$:

1. Mediante integración numérica tras la obtención del estimador f_n resolviendo la ecuación en t

$$\int_{\{f_n \geq t\}} f_n(x)dx = 1 - \tau$$

   El coste computacional de este método se incrementa a medida que aumenta d, la dimensión del espacio.

2. Obtener $\hat{f}_\tau$ como el cuantil τ de la distribución empírica de $f_n(X_1), ..., f_n(X_n)$ tras construir el estimador no paramétrico de la densidad. Aquí el coste computacional es mucho menor.

Por estas razones, será el segundo método, propuesto en Hyndman (1996) el que se empleará en los Capítulos 3 y 4 para obtener $\hat{f}_\tau$ a la hora de estimar los conjuntos de nivel.

La estimación de f se suele realizar mediante la construcción de un estimador tipo núcleo de la siguiente manera:

$$f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_H(x - X_i)$$

donde

$$K_H(u) = |H^{-\frac{1}{2}}|K(H^{-\frac{1}{2}}u)$$

es el kernel reescalado, siendo H una matriz simétrica, definida positiva de dimensión $d \times d$ que se denota por matriz ventana y $K : \mathbb{R}^d \to \mathbb{R}$ es una función de densidad d-dimensional denominada función kernel[1].

Como mostraremos en la subsección siguiente, la elección de la matriz de suavizado H es muy relevante, mientras que el kernel apenas tiene efecto (véase Wanli (2020)). En el Apéndice A se muestran algunos de los kernels unidimensionales y multidimensionales más empleados. En el caso de emplear datos en dimensiones mayores de uno, en este trabajo siempre vamos a usar el kernel normal multidimensional ya que en R la función principal que empleamos es `kde`, la cual no permite cambiar el núcleo. Esto no supone ningún problema ya que a continuación veremos que apenas hay cambios al hacer estimaciones con otro kernel. Para ilustrar el escaso efecto de la elección del núcleo tenemos la Figura 2.3 donde se muestra la diferencia entre las estimaciones usando los kernels unidimensionales presentados en el Apéndice A. Se pretende estimar la densidad unidimensional exacta que se muestra en la Figura 2.2 usando $n = 1000$ puntos y el valor de la ventana por defecto en la función `density`. Se observa que, a excepción del núcleo uniforme, todos proporcionan una estimación bastante similar. La diferencia del núcleo uniforme se debe a la forma de construir con segmentos de rectas el estimador siendo el resultado una estimación discontinua. A la vista de este ejemplo parece claro que podemos concluir que la elección del kernel no será del todo importante mientras sea uno distinto al uniforme. Para ver con más detalle la influencia del kernel en el error de estimación pueden consultarse Silverman (1986) o Hall y Marron (1987).

El cálculo del parámetro de suavizado se puede realizar como en la estimación no paramétrica de la densidad de manera que se minimice algún error de la estimación por ejemplo mediante plug-in o validación cruzada. En Saavedra (2014) se mencionan algunos métodos de obtención de H específicos para la estimación de conjuntos de nivel como los presentados en Baíllo y Cuevas (2006) o Samworth y Wand (2010) que pueden ser de interés pero no se emplearán en este trabajo. Otras referencias posteriores son Chacón y Duong (2018) y Qiao (2020).

2.1.1. Efecto del parámetro de suavización

Tal como se ha comprobado anteriormente, el núcleo usado no tiene gran efecto a la hora de hacer las estimaciones de la densidad o de los conjuntos, sin embargo el valor de la matriz ventana H sí que es muy relevante, tanto para la estimación de los conjuntos, como para el cálculo de las componentes conexas. Para ver cómo afecta el valor de H se tiene la Figura 2.4 donde de nuevo se usa la base de datos `grevillea` para estimar los conjuntos de nivel para los valores de τ de 0.9, 0.6 y 0.1 con dos matrices H distintas.

[1] En el caso unidimensional se define el estimador como $f_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i)$, siendo $K_h(u) = h^{-1}K(\frac{u}{h})$ una función kernel $K : \mathbb{R} \to \mathbb{R}$ reescalada por h. Recordemos que h es un número real positivo que determina el grado de suavización, así valores inapropiadamente grandes de h nos llevan a estimadores sobresuavizados con mucho sesgo y poca varianza mientras que valores demasiado pequeños resultan en estimadores infrasuavizados con poco sesgo y mucha varianza. La matriz H sería en este caso $H = h^2$

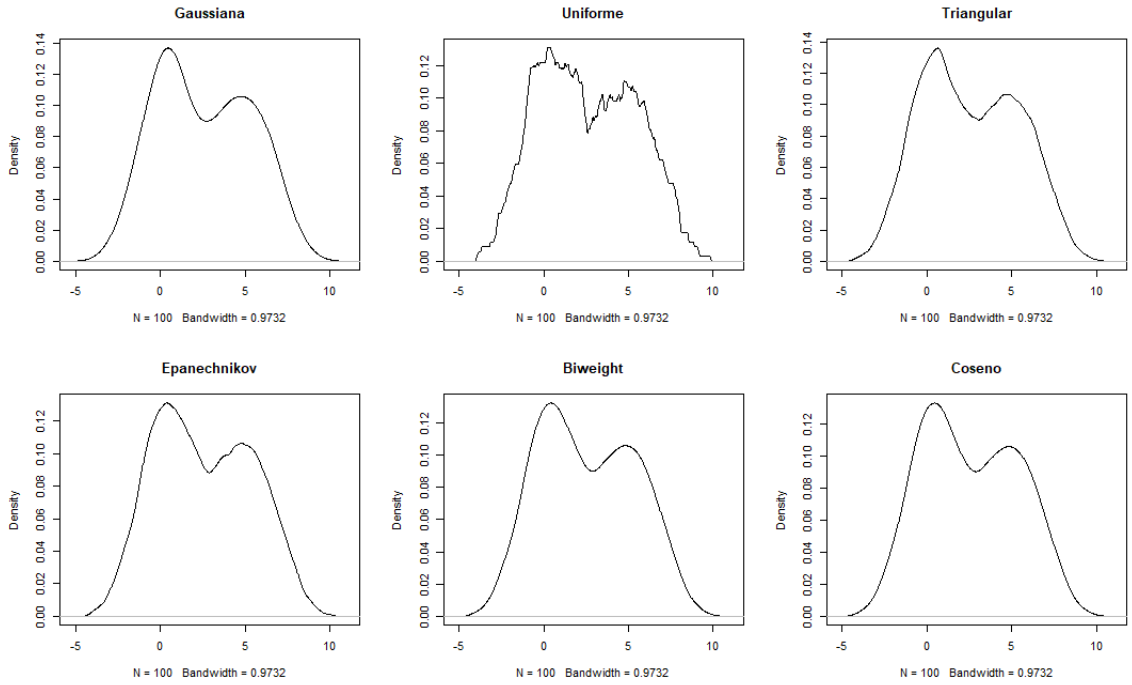

Figura 2.3: Estimaciones de la densidad teórica representada en la Figura 2.2 empleando los seis kernels mencionados en el Apéndice A con n=1000 puntos.

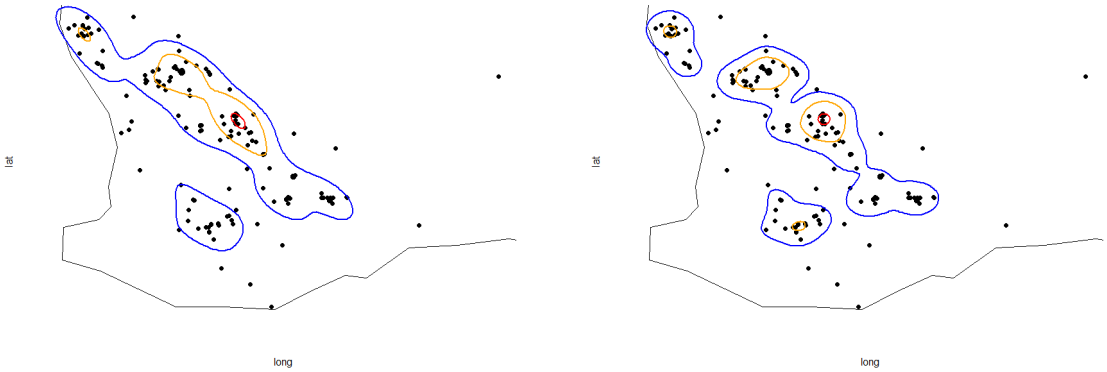

Figura 2.4: Estimación de los conjuntos de nivel para $\tau = 0.9$ (rojo), $\tau = 0.6$ (naranja) y $\tau = 0.1$ (azul) de los datos de `grevillea` usando la matriz de suavizado H (izquierda) y la matriz de suavizado H_2 (derecha).

A la izquierda se han empleado la matriz ventana obtenida mediante plug-in, véase Duong y Hazelton (2003), que nos proporciona por defecto la función `kde` de R y a la derecha se ha tomado como matriz de suavizado la obtenida por plug-in restringida a que sea diagonal, es decir, que cualquier valor que no se encuentre en la diagonal de la matriz sea cero. Como resultado se obtienen las matrices de suavización H_1 y H_2 siguientes:

$$H_1 = \begin{pmatrix} 0.0585 & -0.0446 \\ -0.0446 & 0.0791 \end{pmatrix} \qquad H_2 = \begin{pmatrix} 0.0349 & 0 \\ 0 & 0.0399 \end{pmatrix}$$

El resultado es que el número de componentes conexas varía en función de la matriz H para los distintos valores de τ. En la parte izquierda de la Figura 2.4, para $\tau = 0.1$ se tienen dos componentes conexas; con el mismo contenido en probabilidad en la parte derecha el número de componentes conexas asciende a tres aunque vemos que la superficie es bastante similar en ambos casos. Al tomar $\tau = 0.6$, para la matriz H_1, hay dos componentes conexas, ambas en el interior de la mayor de las modas de $\tau = 0.1$. Por otro lado, para la matriz H_2 aparecen cuatro componentes conexas, cubriendo una superficie diferente que en el caso de la matriz H_1. Por último, para $\tau = 0.9$ tenemos una moda para ambas matrices y además ambas ocupando una superficie muy similar. Por tanto vemos que la elección de H no solo afecta al número de componentes sino también a la área geográfica que estas ocupan.

La matriz de suavizado que se va a obtener para las estimaciones de los conjuntos de nivel es siempre la que nos proporciona por defecto la función `kde` de R obtenida mediante plug-in.

2.1.2. Estimación del número de componentes conexas

Una vez hemos estimado $\hat{\mathcal{L}}(\tau)$, el conjunto de nivel para un cierto valor de τ, mediante métodos plug-in, nos interesa determinar el número de componentes conexas existentes. Esto no es tan sencillo porque la geometría de $\hat{\mathcal{L}}(\tau)$ no es obvia a partir de su definición. Las componentes conexas constituyen regiones de alta densidad de observaciones para los diferentes valores de τ. Cada parte separada del conjunto de nivel estimado se identifica así con una componente conexa. En este trabajo se presentan dos métodos para obtener el número de componentes conexas. El primero de ellos lo denotaremos por 'CFF', siglas obtenidas a partir de los autores Cuevas, Febrero y Fraiman de Cuevas et al (2000) para estimar el número de clústers a partir de distancias entre las observaciones de $\hat{\mathcal{L}}(\tau)$; el segundo método será haciendo uso de la librería de R, `pdfcluster`, que usa las observaciones de $\hat{\mathcal{L}}(\tau)$ junto con técnicas de geometría computacional.

CFF

El primer método, como hemos comentado, se propone y desarrolla en Cuevas et al (2000). Partimos de un conjunto de nivel $\mathcal{L}(\tau)$ y tenemos como objetivo estimar el número de componentes conexas distintas que lo forman, parámetro al cual denotaremos por $T(\mathcal{L})$. Como el conjunto de nivel hemos tenido que estimarlo como explicamos a lo largo de esta sección, realmente tenemos el conjunto estimado no paramétricamente $\hat{\mathcal{L}}(\tau)$. En vez de emplear este directamente para estimar el número de componentes que lo forman, vamos a considerar las observaciones de la muestra $\mathcal{X}_n$ pertenecientes al conjunto de nivel estimado $X_i \in \hat{\mathcal{L}}(\tau), i = 1, ..., k$ donde suponemos, sin pérdida de generalidad, que estas observaciones son las primeras y definiremos un estimador significativamente más simple del conjunto de nivel:

$$\hat{S} = \bigcup_{i=1}^{k} B[X_i, \varepsilon],$$

donde ε se considera un parámetro de suavizado[2]. La razón de usar $\hat{S}$ en vez de la estimación no

[2]Cumpliendo ciertas condiciones, cuanto menor sea el valor que toma ε, más componentes conexas distintas se calcularán, aunque no puede ser arbitrariamente grande ya que para estimar bien el número de componentes conexas

paramétrica del conjunto de nivel directamente es, tal como explican en Cuevas et al (2000), debido a la simplicidad de calcular el número de componentes conexas en una unión finita de bolas cerradas. En la Figura 2.5 se puede observar el conjunto $\hat{S}$ para una muestra trimodal de 500 observacionesy $\tau = 0.5$ con tres valores distintos de $\varepsilon = 2.5, 1, 0.1$. Podemos ver la función de parámetro de suavizado de ε: a la izquierda, para el mayor valor tenemos una única componente; al tomar un ε más bajo se muestran tres componentes distintas como vemos en el centro. Por último si tomamos un valor excesivamente pequeño como en la parte derecha de la Figura 2.5 vemos que se crean muchos grupos pequeños.

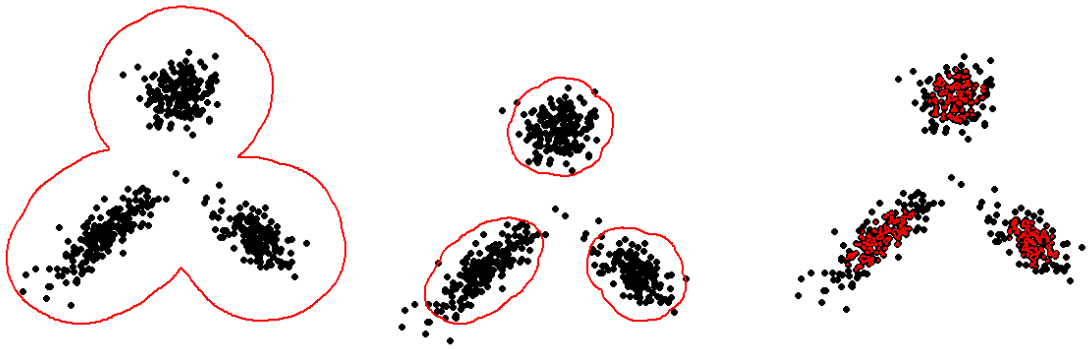

Figura 2.5: En negro, puntos de una muestra trimodal y en rojo, representación del conjunto $\hat{S}$ en rojo para la muestra, $\tau = 0.5$ y distintos valores de ε, a la izquierda $\varepsilon = 2.5$, en el medio $\varepsilon = 1$ y a la derecha $\varepsilon = 0.1$.

Para obtener el estimador $T = T(\hat{S})$ se emplea un algoritmo que presenta a continuación. Se basa en la idea de que dos observaciones $X_i, X_j \in \hat{\mathcal{L}}(\tau)$ pertenecen a la misma componente si y solo si la distancia euclídea entre ellas es menor o igual que dos veces ε. El procedimiento se puede ver en el Algoritmo 1.

Una vez presentado el algoritmo que se va a emplear, solo nos falta estimar el parámetro de suavización ε. Llegar a un valor óptimo de este es complejo ya que hay que encontrar un equilibrio entre un valor lo suficientemente bajo como para que sea capaz de separar componentes conexas pero que sea lo bastante grande como para que los puntos cercanos estén en el mismo clúster. De nuevo Cuevas et al (2000) propone un método para obtener valores de ε para los que se consiguen buenos resultados a la hora de estimar componentes conexas tal como veremos en el capítulo 4. También vamos a proponer otra opción mencionada en Cuevas et al (2000) cuyo funcionamiento veremos que es muy mejorable.

Para comenzar con el método de obtención de ε, tenemos que dividir la muestra de $\hat{\mathcal{L}}(\tau)$ en clusters usando el método de las distancia mínima de manera que cada grupo contenga al menos el 5 % del total de observaciones de $\hat{\mathcal{L}}(\tau)$. Una vez los tenemos, vamos a calcular dos tipos de distancias: por un lado la distancia interna que denotaremos por $\triangle_w$ y por otro la distancia externa, $\triangle_b$. La interna es el percentil $1 - \beta$ de todas las distancias euclídeas $D_2(X_i, X_j)$ de los X_i y X_j que pertenezcan al mismo grupo generado, donde β es un parámetro entre 0 y 1 que podremos elegir. La externa es la mínima distancia entre clusters, es decir, mín $D_2(X_i, X_j)$ cuando X_i y X_j pertenecen a distintos grupos; en el caso de que el número de clusters sea solo uno, se toma $\triangle_b = \infty$. Los valores de β que se van a tomar son $\beta = 0.05, 0.1, 0.2$ que nos indican cuantas distancias entre observaciones del mismo grupo se van a

debe ser menor que la mitad de la distancia mínima entre las componentes conexas del conjunto de nivel.

Algoritmo 1 Estimación del número de componentes conexas CFF

Entrada: Observaciones $X_i \in \hat{\mathcal{L}}(\tau)$, parámetro ε
Salida: T

$T \leftarrow 0$
Marcar todos los puntos X_i como no seleccionados
while existan puntos no seleccionados en $\hat{\mathcal{L}}$ **do**
  PASO 1: Elegir un centro $X_1 \in \hat{\mathcal{L}}$ no seleccionado previamente
  Calcular $r_1 = \min_{X_j \in \hat{\mathcal{L}},\, j \neq 1} D_2(X_1, X_j)$
  Marcar X_1, X_2 como seleccionados
  $C \leftarrow \{X_1, X_2\}$
  $m \leftarrow 1$
  PASO 2:
  if $r_1 > 2\varepsilon$ **then**
    $T \leftarrow T + 1$
    volvemos al PASO 1
  si no
    $X_3 \leftarrow \arg\min_{X_j \notin C} \min_{X_i \in C} D_2(X_i, X_j)$
    $r_2 \leftarrow \min\{D_2(X_1, X_3),\, D_2(X_2, X_3)\}$
    Marcar X_3 como seleccionado
    $C \leftarrow C \cup \{X_3\}$
    $m \leftarrow 2$
  end if
  PASO 3:
  while $r_m \leq 2\varepsilon$ **do**
    $X_{m+1} \leftarrow \arg\min_{X_j \notin C} \min_{X_i \in C} D_2(X_i, X_j)$
    $r_m \leftarrow \min\{D_2(X_{m+1}, X_i) : i = 1, \ldots, m\}$
    if $r_m \leq 2\varepsilon$ **then**
      Marcar X_{m+1} como seleccionado
      $C \leftarrow C \cup \{X_{m+1}\}$
      $m \leftarrow m + 1$
    end if
  end while
  $T \leftarrow T + 1$
end while
devolver T

tener en cuenta para el cálculo de la distancia interna. A mayor valor de β, menor es el de ε y, por tanto, el número de componentes conexas aumenta. Una vez tenemos las distancias interna y externa calculadas, obtenemos el valor de ε como se indica en Cuevas et al (2000):

$$\varepsilon = \max\left\{\triangle_w + \frac{\triangle_b - \triangle_w}{3}, 0.95\left(\triangle_b - \frac{\triangle_w}{2}\right)\right\}, \triangle_b < \infty \qquad \varepsilon = \frac{3\triangle_w}{2}, \triangle_b = \infty.$$

A partir de este punto se emplea la notación siguiente: se denotará por ε_1 cuando $\beta = 0.05$, ε_2 cuando $\beta = 0.1$ y ε_3 cuando $\beta = 0.2$.

Una alternativa que se ofrece pero cuyo funcionamiento no es demasiado bueno en la práctica, tal como veremos a continuación es tomar ε como la menor distancia que conecta cada X_i con su vecino más cercano, esto es:

$$\varepsilon = \max_i \min_{j \neq i} D_2(X_i, X_j)/2$$

Se denotará por ε_4 a este último parámetro y por ε_5 a el doble del valor ε_4 ya que el primero de ellos suele ser demasiado pequeño y genera demasiadas componentes conexas por no ser capaz de agrupar correctamente las observaciones.

Vamos a aplicar esta metodología sobre los datos de `grevillea`. Para ello vamos a considerar el valor de $\tau = 0.2$ y tras obtener los cinco valores de ε en cada uno de los casos, veremos el número de componentes conexas que se obtienen en cada caso. Estos se muestran en la Tabla 2.1. Los grupos resultantes se pueden ver gráficamente en la Figura 2.6.

Parámetros ε	0.6612	0.6152	0.5045	0.1915	0.3830
Nº componentes conexas estimadas	4	4	4	12	5

Tabla 2.1: Valores del parámetro ε y el número de componentes conexas estimadas mediante la metodología CFF para $\tau = 0.2$.

En la parte superior izquierda de la Figura 2.6 se muestran en rojo todos los puntos cuya densidad estimada está por encima de f_τ para $\tau = 0.2$; el resto de representaciones tienen en distintos colores los puntos que pertenecen a la misma componente conexa. Vemos así que para los tres primeros valores de ε el número de componentes conexas es el mismo aunque los grupos formados no sean exactamente iguales. Sin embargo ε_4 nos genera un número de componentes conexas mucho más elevado debido a ser un valor mucho más conservador y demasiado pequeño. Esto se arregla ligeramente con ε_5 pero aun así vemos que la distancia ε sigue siendo pequeña. Tras ver estos resultados, en los dos capítulos restantes únicamente se empleará ε_2 ya que las diferencias no son significativas en las estimaciones con los parámetros ε_1 y ε_3 y ε_4 y ε_5 no tienen un buen funcionamiento en la práctica.

PdfCluster

Como segundo método para el cálculo de componentes conexas una vez hemos estimado $\hat{\mathcal{L}}(\tau)$ se presenta la metodología que denominaremos 'pdfcluster' debido al nombre del paquete de R en el que está implementada. Las ideas de esto se encuentran en Azzalini y Menardi (2014), siendo este un resumen de las dos referencias principales Azzalini y Torelli (2007) y Menardi y Azzalini (2014) y se basan en técnicas de geometría computacional. En particular, se va a seguir lo presentado en Azzalini y Torelli (2007) e implementado en R donde se emplean diagramas de Voronoi y la triangulación de Delaunay. Posteriormente, en Menardi y Azzalini (2014), ya no se hace uso de la triangulación de Delaunay sino que se busca una alternativa que no se trata en este trabajo.

Tras elegir un valor $\tau \in (0, 1)$, y dada una muestra $\mathcal{X}_n$, obtenemos, tal como se ha explicado previamente, $\hat{\mathcal{L}}(\tau) = \{x \in \mathcal{X} : f_n(x) \geq \hat{f}_\tau\}$ y nos vamos a fijar, como antes, en aquellas observaciones de este conjunto muestral. Sobre la muestra completa $\mathcal{X}_n$ se calcula el diagrama de Voronoi para, a partir de ella, obtener la triangulación de Delaunay. Este diagrama, en $\mathbb{R}^2$, se define como una subdivisión del espacio en tantas regiones como observaciones hay en la muestra. La forma de hacer las divisiones es empleando la distancia euclídea de la Definición 1.2 de manera que cada una de las regiones están formadas por aquellos puntos del plano con menor distancia euclídea a cada observación. Notemos que las fronteras de las regiones están formadas por aquellos puntos que se encuentran a igual distancia de las dos observaciones de la muestra incluidas en las regiones que comparten lado. A partir de este diagrama obtenemos la triangulación de Delaunay como la unión mediante segmentos de aquellas observaciones que comparten un lado en el diagrama de Voronoi. Para el cálculo en R de

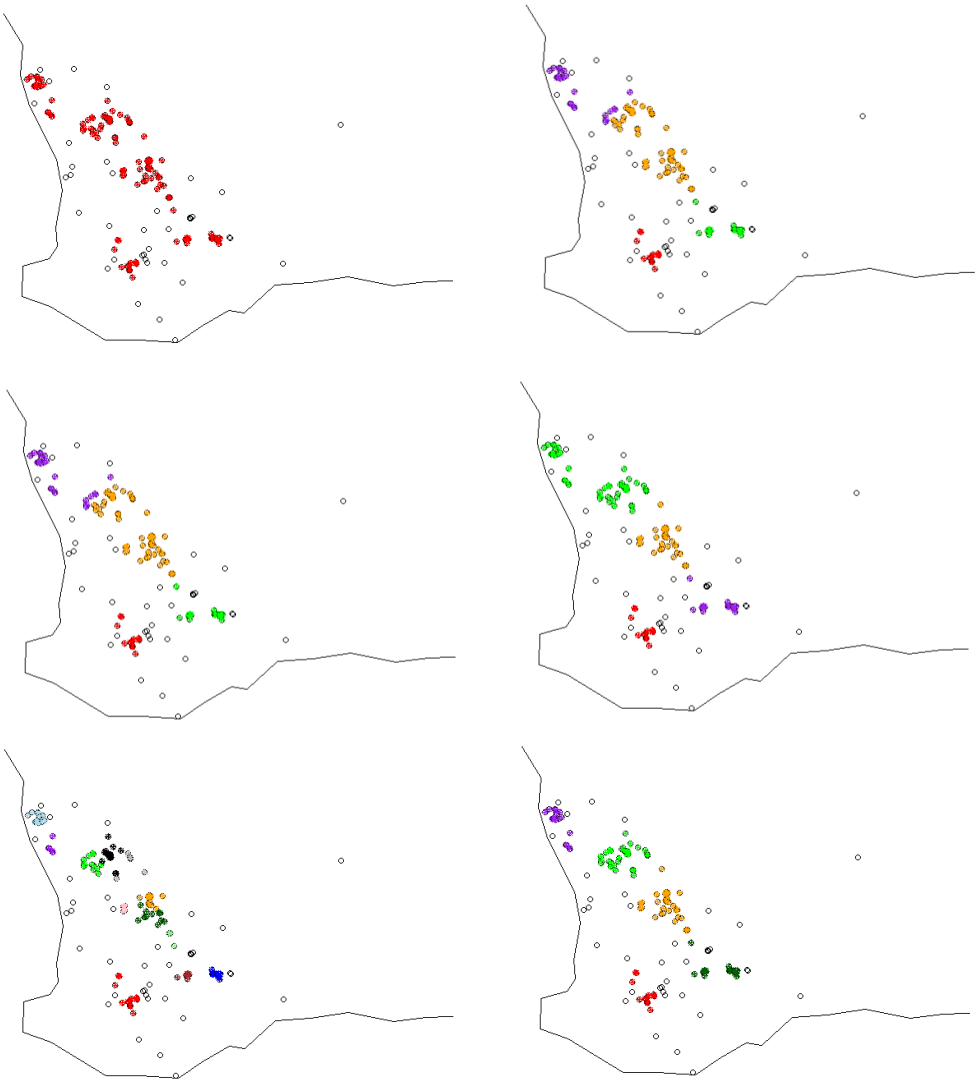

Figura 2.6: Representación de los puntos de $\hat{\mathcal{L}}(\hat{f}_\tau)$ para $\tau = 0.2$ (superior izquierda) y componentes conexas estimadas con el método de CFF con los distintos valores de ε: $\varepsilon_1 = 0.6612$ (superior derecha), $\varepsilon_2 = 0.6152$ (centro izquierda), $\varepsilon_3 = 0.5045$ (centro derecha), $\varepsilon_4 = 0.1915$ (inferior izquierda) y $\varepsilon_5 = 0.3830$ (inferior derecha).

la triangulación y el diagrama asociado, se emplearán las librerías `interp` (Gebhardt et al, 2024) y `tripack` (Gebhardt et al, 2024). En la Figura 2.7 se muestra tanto el diagrama de Voronoi a la izquierda como la triangulación de Delaunay a la derecha de la muestra de `grevillea`. En la parte izquierda vemos que aparece el espacio dividido en regiones y cada uno de los puntos que se muestran son los vértices de los polígonos que forman el diagrama conocidos como celda de Voronoi. Estos no coinciden con las observaciones sino que son puntos del plano que se encuentran equidistantes a tantas observaciones como aristas se junten en ellos. Por otro lado, en la parte derecha, aparece la triangulación de Delaunay; en este caso los puntos en rojo sí coinciden con las observaciones de la muestra y están unidas a otras formando triángulos según lo explicado previamente.

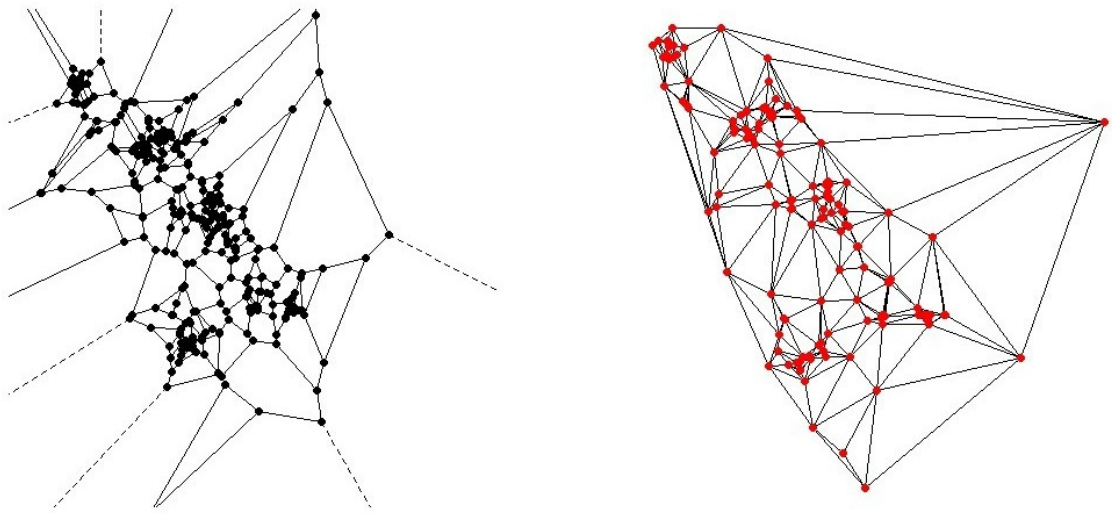

Figura 2.7: Diagrama de Voronoi (izquierda) y triangulación de Delaunay (derecha) obtenidos sobre la muestra de `grevillea`.

Tal como mencionan en Azzalini y Menardi (2014) se puede obtener directamente la triangulación de Delaunay sin necesidad de calcular el diagrama de Voronoi completo que puede llegar a ser complejo computacionalmente. Una vez tenemos la triangulación hecha nos vamos a quedar solo con las observaciones pertenecientes a $\hat{\mathcal{L}}(\tau)$, es decir, se eliminan las observaciones con densidad menor que el umbral $\hat{f}_\tau$ y todos aquellos segmentos pertenecientes a la triangulación de Delaunay que tuvieran como inicio o fin alguna de esas observaciones. Así, nos quedamos solo con los puntos de mayor densidad estimada y segmentos de la triangulación que unen a dos de estos puntos. A partir de esto, dos observaciones de $\hat{\mathcal{L}}(\tau)$ formarán parte de la misma componente conexa si y solo si, hay una arista de la triangulación de Delaunay uniéndolas. De esta manera podemos formar polígonos quedándonos solo con las rectas que unan puntos de $\hat{\mathcal{L}}(\tau)$. Cada una de las figuras que se formen independientes, es decir, que no tengan ninguna arista en común, será una componente conexa. Notemos que en el caso de que haya algún punto aislado no se considerará una nueva componente conexa sino que formará parte de la más cercana. Todo este procedimiento está implementado en la función `pdfcluster`. Notemos que esta tiene por defecto un factor $hmult = 0.75$ por el que se multiplica cada elemento de la matriz H obtenida mediante plug-in a la hora de la estimación no paramétrica de la densidad f_n. Así, a la hora del cálculo de componentes conexas vamos a emplear la matriz H que se nos da por defecto en la función `pdfcluster`. Cabe destacar que, tras diversas pruebas que no se incluyen en el trabajo, cuando se toma el valor $hmult = 1$ se obtienen, en general, menor número de componentes conexas que al tomar los parámetros por defecto, por lo que vemos que, de nuevo, la matriz de suavizado tiene un papel importante.

Veamos esto gráficamente con los datos de `grevillea` y para el conjunto de nivel con $\tau = 0.2$ en la Figura 2.8. En la parte izquierda tenemos los puntos que se han obtenido con densidad estimada mayor que $\hat{f}_\tau$ con $\tau = 0.2$ según la función `pdfcluster` con los valores por defecto. En azul están dichas observaciones y en rojo el resto. A la derecha de ella están esas mismas observaciones de mayor densidad divididas en dos grupos que forman las dos componentes conexas: uno en la parte izquierda superior formado por las observaciones en azul claro y otro de mayor tamaño en verde. De nuevo, en rojo, se muestran las observaciones con menor densidad estimada.

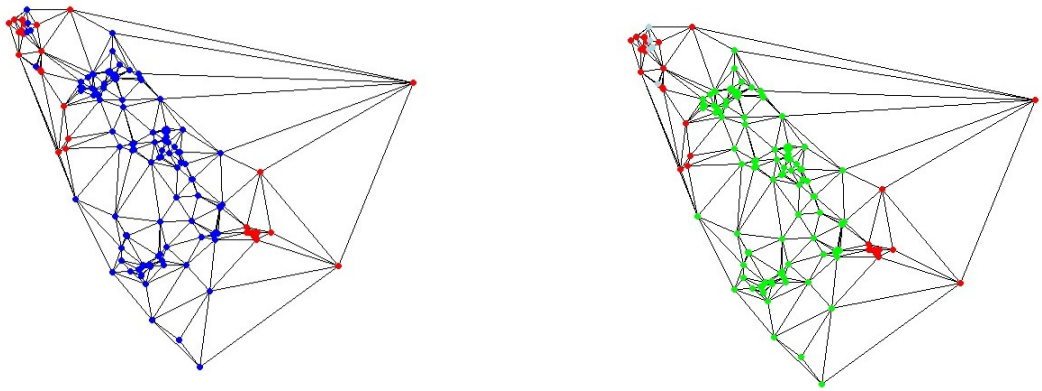

Figura 2.8: Representación de los puntos de `grevillea` con densidad estimada mayor que $\hat{f}_\tau$ con $\tau = 0.2$ obtenidos con `pdfcluster` (izquierda en azul) y clusters formados (derecha en verde y azul claro).

No coincide lo obtenido con el método CFF para los distintos valores de ε con el número de componentes conexas calculadas con este método. Por ello en el capítulo siguiente se compararán los resultados obtenidos tras realizar estimaciones con diferentes metodologías, umbrales y distribución de las observaciones de la muestra.

2.2. Exceso de masa

Otra metodología para estimar conjuntos de nivel se ha estudiado bajo la premisa de que el conjunto de nivel a estimar $G(t)$ cumpla alguna condición geométrica. Un ejemplo podría ser asumir que el conjunto de nivel fuese convexo. En este método, a diferencia de los anteriores y de los métodos híbridos, no hay estimación de la función de densidad. Las referencias empleadas son Hartigan (1987) o Polonik (1995)

Se emplea el hecho de que $G(t)$, el conjunto de nivel t, maximiza el funcional de exceso de masa siguiente

$$H_t(B) = \mathbb{P}(B) - t\mu(B)$$

donde B es un conjunto de Borel, $\mathbb{P}$ es la medida de probabilidad inducida por f y μ es la medida de Lebesgue. La ventaja de H_t es que se puede estimar empíricamente mediante

$$H_{n,t}(B) = \mathbb{P}_n(B) - t\mu(B)$$

donde $\mathbb{P}_n$ denota la probabilidad empírica inducida por la muestra $\mathcal{X}_n$. De esta forma tenemos, asumiendo una condición geométrica sobre el conjunto $G(t)$ podemos obtener un estimador $\hat{G}(t)$ maximizando el exceso de masa empírico $H_{n,t}$ anterior.

En la Figura 2.9 se tiene una representación del funcional exceso de masa unidimensional en gris para un cierto nivel t. En la zona inferior, en naranja, se muestra $t\mu(B)$. En este caso el conjunto de nivel estaría compuesto por una única componente conexa.

Al igual que en la Sección 2.1 podemos cambiar t por un valor f_τ que asegure que el conjunto contenga contenido en probabilidad mínimo $1 - \tau$, $\tau \in (0, 1)$.

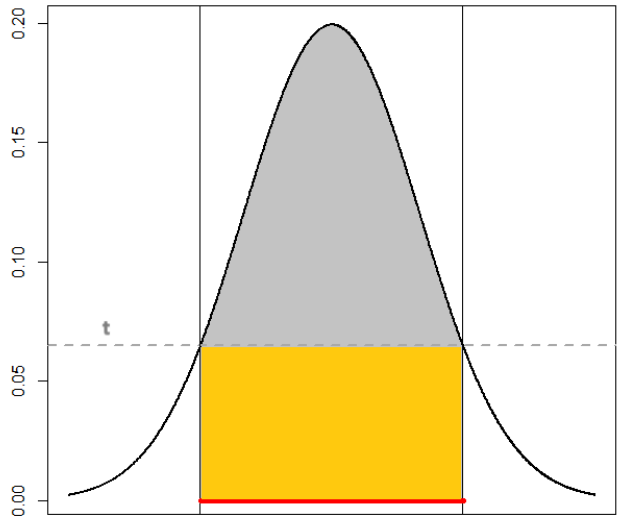

Figura 2.9: Representación de $H_t(B)$ (gris) y $t\mu(B)$(naranja) para un cierto nivel t y una distribución unidimensional normal.

Debido a que las condiciones geométricas en $\mathbb{R}^2$ pueden ser muy diversas, esta metodología únicamente se menciona pero no se va a aplicar a lo largo del documento. En el artículo Grübel (1988) se trata el problema de exceso de masa asumiendo convexidad en $\mathbb{R}$; en Hartigan (1987) se proponen métodos en el caso bidimensional bajo la hipótesis de convexidad. Para ver resultados asintóticos de la estimación por exceso de masa se puede consultar Polonik (1995).

2.3. Métodos híbridos

La última metodología que se presenta es la *híbrida* en la cual se usa tanto la estimación no paramétrica de la función de densidad f como la asunción de condiciones geométricas a priori en el conjunto de nivel $\mathcal{L}(\tau)$.

A pesar de que existen más métodos híbridos en la literatura en este trabajo trataremos tres: el método de la envoltura convexa (véase Walther (1997)), el método de la envoltura $r-$convexa con r dependiente de los datos desarrollado en Rodríguez-Casal y Saavedra (2022) y un método de suavizado granulométrico con un valor de r también seleccionado de la muestra presentado en Bolón et al (2024).

Al igual que en el plug-in, dada una muestra $\mathcal{X}_n$ se realiza una estimación no paramétrica de la densidad f_n a partir de la que se construyen los siguientes conjuntos tras calcular el umbral $\hat{f}_\tau$ que se emplearán para la estimación de los conjuntos de nivel:

$$\mathcal{X}_{n,+}(\hat{f}_\tau) = \{x \in \mathcal{X}_n : f_n(x) \geq \hat{f}_\tau\} \qquad \mathcal{X}_{n,-}(\hat{f}_\tau) = \mathcal{X}_n \setminus \mathcal{X}_{n,+}(\hat{f}_\tau).$$

En el primero de los métodos se asume que el conjunto de nivel es convexo, lo cual es bastante

restrictivo como veremos a continuación; en el segundo suponemos que $\mathcal{L}(\tau)$ es $r-$convexo para algún $r < 0$ y el mayor problema es obtener un valor de r óptimo que nos separe de manera correcta las posibles modas; en el último de los métodos se asume que una bola de radio r rueda libremente dentro de $\mathcal{L}(\tau)$. De nuevo nos encontramos con el problema de obtener un radio r con el que podamos construir un estimador de $\mathcal{L}(\tau)$ aceptable.

2.3.1. Método de la envoltura convexa

Tal como comentábamos previamente en esta metodología asumimos la convexidad del conjunto de nivel $\mathcal{L}(\tau)$. Los pasos a seguir son los siguientes:

1. Se calcula un estimador no paramétrico de la densidad f_n y calculamos el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau) = \{x \in \mathcal{X}_n : f_n(x) \geq \hat{f}_\tau\}$ donde el umbral $\hat{f}_\tau$ se estima se estima como el cuantil de orden τ de $f_n(\mathcal{X}_n) = \{f_{(1)}, \ldots, f_{(n)}\}$[3]

2. Una vez tenemos el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau)$, obtenemos la estimación del conjunto de nivel como la envoltura convexa del conjunto anterior, esto es:

$$\hat{\mathcal{L}}(\tau) = \text{conv}(\mathcal{X}_{n,+}(\hat{f}_\tau))$$

Notemos que debido a la condición de convexidad, el número de agrupaciones de grupos que se van a estimar es siempre una, lo que en muchos casos puede ser muy restrictivo. Aplicado a los datos de `grevillea` tenemos la representación del conjunto de nivel obtenido mediante este método para $\tau = 0.2$ en la Figura 2.10. Vemos que tal como adelantábamos solo hay una componente conexa y además hay muchos puntos que forman parte de $\mathcal{X}_{n,-}$ en el interior del conjunto de nivel. Por lo tanto no será un método que funcione bien en la mayoría de los casos y por ello no se aplicará en los próximos capítulos.

2.3.2. Método de la envoltura $r-$convexa

La condición de convexidad es restrictiva en muchos casos por lo que una alternativa más aplicable y que permite la obtención de más de una componente conexa es usar como restricción de forma la $r-$convexidad. La forma de obtener el conjunto de nivel estimado $\hat{\mathcal{L}}(\tau)$ es mediante los siguientes pasos:

1. Se elige un valor de $r > 0$.

2. Se calcula un estimador no paramétrico de la densidad f_n y calculamos el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau) = \{x \in \mathcal{X}_n : f_n(x) \geq \hat{f}_\tau\}$ donde el umbral $\hat{f}_\tau$ se estima se estima como el cuantil de orden τ de $f_n(\mathcal{X}_n) = \{f_{(1)}, \ldots, f_{(n)}\}$.

3. Una vez tenemos el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau)$, obtenemos la estimación del conjunto de nivel como la envoltura $r-$convexa del conjunto anterior (ver Definición 1.7), esto es:

$$\hat{\mathcal{L}}(\tau) = C_r(\mathcal{X}_{n,+}(\hat{f}_\tau))$$

.

[3]Una vez ordenados de forma creciente todas las evaluaciones $f_n(x), x \in \mathcal{X}_n$ sobre la muestra,$f_{(k)}$ es el valor que ocupa la posición $k-$ésima; esto es, $f_{(1)}$ sería el menor y $f_{(n)}$ el mayor.

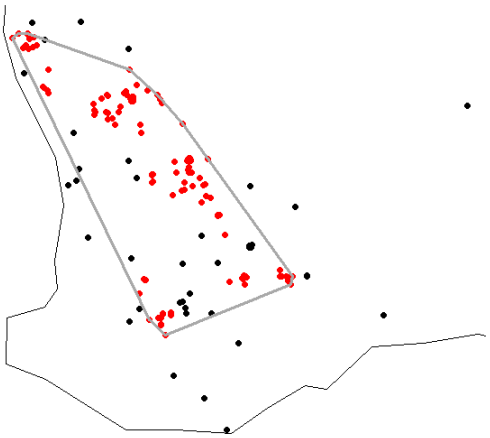

Figura 2.10: Representación de los puntos de grevillea (negro), el conjunto $\mathcal{X}_{n,+}$ (rojo) y el conjunto de nivel estimado por el método de la envoltura convexa (gris).

En este caso, tenemos más flexibilidad pero se añade una nueva dificultad: la obtención del parámetro r. En el caso de que r sea muy pequeño, se formarán numerosos grupos pequeños mientras que si r es muy grande el conjunto de nivel estimado se parecerá cada vez más al obtenido con la envoltura convexa.

En la Figura 2.11 tenemos el conjunto de nivel estimado para $\tau = 0.2$ de los datos de `grevillea` para cuatro valores distintos de r: a la izquierda arriba un valor muy pequeño[4], $r = 0.4$, generando 6 componentes conexas, a la derecha arriba un valor un poco superior, $r = 0.6$, dividiendo el conjunto de nivel estimado en 4 componentes conexas, y en la parte inferior dos valores más grandes: $r = 1$ a la izquierda y $r = 4$ a la derecha, que en ambos casos solo dan lugar a una componente conexa.

Así, vemos que para los valores pequeños de r se nos generan grupos de unas pocas observaciones o incluso un único punto aislado y a medida que aumentamos el valor de r, el número de grupos disminuye hasta que para un valor alto, como puede ser en este caso $r = 4$, el conjunto de nivel estimado con el método $r-$convexo se parece mucho al obtenido con la metodología de la envoltura convexa que se muestra en la Figura 2.10.

Por tanto ya hemos visto la importancia de tener un valor adecuado para r. A continuación se presenta la obtención de un valor óptimo que se desarrolla en Rodríguez-Casal y Saavedra (2022).

Elección óptima del parámetro r

Primero se introduce la idea que hay detrás de la elección del valor óptimo de r y después se presenta el algoritmo empleado en la aplicación y simulaciones, todo extraído de Rodríguez-Casal y Saavedra (2022). Si conociésemos la función de densidad f, podríamos definir el valor óptimo de r como se muestra en la Definición 2.1.

[4]Notemos que un mismo valor de r puede ser grande para un conjunto de datos y pequeño para otros; por ejemplo si tenemos por un lado la ubicación de una especie de flor en una parcela y por otro la ubicación de esta misma especie en toda una provincia, un valor de r que en el caso de la parcela sea grande, puede ser pequeño a nivel provincial.

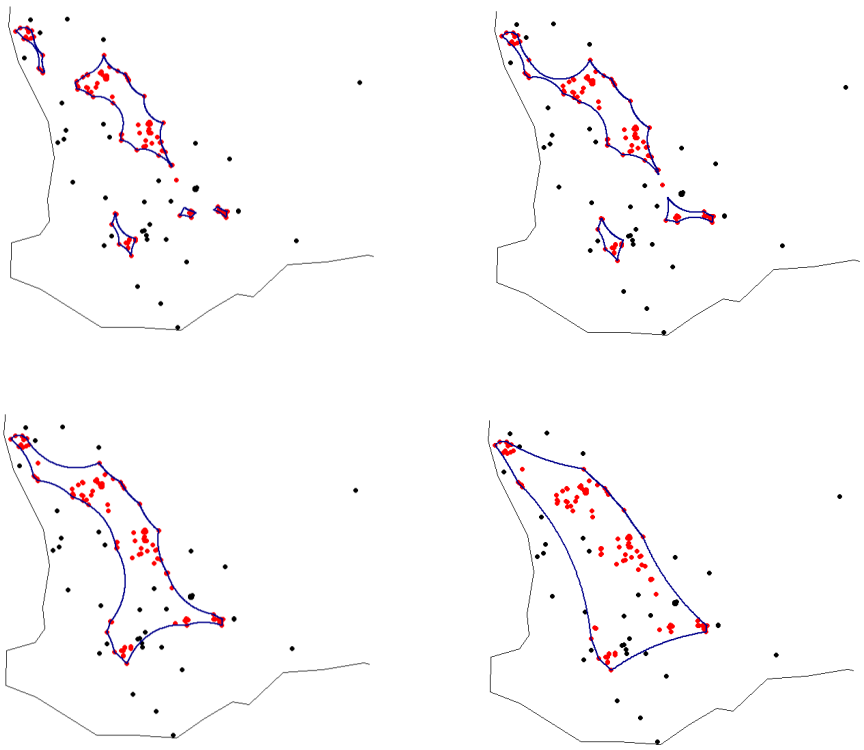

Figura 2.11: Representación de los puntos de grevillea (negro), el conjunto $\mathcal{X}_{n,+}$ (rojo) y los conjuntos de nivel estimados por el método de la envoltura $r-$convexa (azul) para cuatro valores de r: $r = 0.4$ arriba a la izquierda, $r = 0.6$ arriba a la derecha, $r = 1$ abajo a la izquierda y $r = 4$ abajo a la derecha.

Definición 2.1 *Sea $G(t)$ un conjunto de nivel compacto, no vacío, no convexo y $r-$convexo para algún $r > 0$. Se define el parámetro óptimo como:*

$$r_0(t) = \sup\{\gamma > 0 : C_\gamma(G(t)) = G(t)\}$$

Notemos que se pide no convexidad del conjunto de nivel porque en ese caso el valor de r_0 sería infinito y la envoltura $r-$convexa sería igual a la convexa pudiéndose aplicar directamente el método desarrollado en la Sección 2.3.1 con lo que el coste computacional sería mucho menor y se ahorraría tiempo de ejecución.

El supremo de la definición anterior es un máximo bajo la siguiente propiedad geométrica:

(R_r^λ): Una bola cerrada de radio $\lambda > 0$ rueda libremente en $G(t)$ y una bola cerrada de radio $r > 0$ rueda libremente en $G(t)^c$.

Claramente, obtener r_0 no es realista en la práctica ya que no vamos a conocer la función de densidad f y por tanto tampoco $G(t)$. Así que tendremos que obtener un estimador del r_0 anterior. Para la obtención del estimador de r_0 vamos a utilizar un estimador piloto de f al que denotaremos por g_n[5] y asumimos que el conjunto

$$G_n(t) = \{x \in \mathbb{R}^d : g_n \geq t\}$$

[5]Notemos que el estimador piloto g_n es diferente al estimador f_n calculado para reconstruir el conjunto de nivel.

sea lo suficientemente suave para asegurar que

$$r_n(t) = sup\{\gamma > 0 : C_\gamma(G_n(t)) = G_n(t)\}$$

está bien definido. El estimador g_n debe verificar ciertas condiciones para que se cumpla lo anterior. Primeramente, la condición **A**. que se desarrolla a continuación, junto con que sea una función C^2 con matriz hessiana acotada. También, al igual que f, será un estimador núcleo para el que debemos elegir una matriz (o parámetro en el caso unidimensional) de suavizado a partir de la muestra y además, el núcleo Φ debe verificar la condición **K**. para asegurar que g_n es suficientemente suave para que g_n y ∇g_n estiman uniformemente f y ∇f. El kernel gaussiano satisface **K**. por lo que será el empleado para estimar g_n.

A. 1. El umbral t de $G(t)$ pertenece a $[l, u]$ donde $-\infty < l \leq u < \sup(f) < \infty$

2. $f \in C^p(U), p \geq 1$ donde U es un conjunto abierto y acotado que contiene a $\overline{G(l-\zeta)} \setminus \text{Int}(G(u+\zeta))$ para algún $\zeta > 0$, donde $G(u+\zeta)$ es acotado.

3. El gradiente de f, ∇f, satisface $|\nabla f| \geq m > 0$ y la condición Lipschitz en U:

$$D_2(\nabla f(x), \nabla f(y)) \leq k D_2(x - y) \text{ para } x, y \in U$$

K. El núcleo Φ está dado por $\Phi(x_1, \ldots, x_d) = \prod_{k=1}^{d} \phi_k(x_k)$. Si $\mathbf{BC}^r$ denota la colección de funciones con derivadas continuas y acotadas hasta el orden r, se supone que $\phi_k \in \mathbf{BC}^2$, con todas las derivadas hasta el orden 2 anulándose en el infinito. También se asume que ϕ_k es no negativa, integra uno, el primer momento es cero, $\int x\phi_k(x)\,dx = 0$, y el segundo es finito, $\int x^2\phi_k(x)\,dx < \infty$.

La condición (R_r^λ) asegura que $\{\gamma > 0 : C_\gamma(G_n(t)) = G_n(t)\}$ es no vacío para n grande y con probabilidad uno. Además, considerando la condición **A**. sobre la función de densidad f, se garantiza que el conjunto $G(t)$ es compacto, no vacío y verifica (R_r^λ), para $r = \lambda = m/k \quad \forall t \in [l, u]$ tal como se desarrolla en el Teorema 2 de Walther (1997). Por tanto vamos a asumir que f también cumple la condición **A**.

En Rodríguez-Casal y Saavedra (2022) se encuentran diversos resultados que garantizan que $r_n(t)$ está bien definido, acotado y es un buen estimador de r_0.

Por lo tanto tenemos resultados que aseguran que r_n sería un gran candidato de radio para la metodología de la envoltura $r-$convexa pero calcularlo no es tarea fácil en la práctica. Por ello, en vez de r_n, se va a emplear el estimador $\tilde{r}_n$ presentado en (2.3) que, como comentan en Rodríguez-Casal y Saavedra (2022), también cumple estar bien definido, acotado y ser un buen estimador de r_0 bajo las mismas condiciones que r_n.

$$\tilde{r}_n = sup\{\gamma > 0 : C_\gamma(\mathcal{X}_{g_n,+}(t)) \cap \mathcal{X}_{g_n,-}(t) = \emptyset\} \tag{2.3}$$

donde

$$\mathcal{X}_{g_n,+}(t) = \{x \in \mathcal{X}_n : g_n(x) \geq t\} \qquad \mathcal{X}_{g_n,-}(t) = \mathcal{X}_n \setminus \mathcal{X}_{g_n,+}(t).$$

El cálculo de $\tilde{r}_n$ es mucho más sencillo en la práctica, por lo que será este el que se va a calcular en este y los próximos capítulos.

La forma de calcular $\tilde{r}_n$ es mediante un algoritmo iterativo. Iremos aumentando el valor de r hasta encontrar un r' para el que se verifique $C'_r(\mathcal{X}_{g_n,+}(t)) \cap \mathcal{X}_{g_n,-}(t) \neq \emptyset$[6]. Esta forma de proceder es posible

[6]Tal como se va a comentar en el algortimo, el valor de t depende del conjunto de nivel y será $\hat{f}_\tau$ dado $\tau \in (0, 1)$, donde $1 - \tau$ es el contenido en probabilidad mínimo del conjunto de nivel a estimar.

gracias a la monotonía de la envoltura $r-$convexa en r; es decir, a medida que aumentamos el radio, también lo hace el área que ocupa y con ello el contenido en probabilidad. De esta forma nos quedamos con el valor de r anterior a r' que consideraremos como el mayor radio para el que se verifica

$$C_r(\mathcal{X}_{g_n,+}(t)) \cap \mathcal{X}_{g_n,-}(t) = \emptyset.$$

El algoritmo para obtener el conjunto de nivel mediante el método de la envoltura $r-$convexa con r óptimo explicado en Rodríguez-Casal y Saavedra (2022) se presenta en el Algoritmo 2.

Algoritmo 2 Método $r-$convexo

Entrada: Valor $\tau \in (0,1)$ y muestra iid $\mathcal{X}_n = \{X_1, \ldots, X_n\}$

Paso 1: Estimación de la densidad y umbral

1: Obtener el estimador no paramétrico f_n a partir de $\mathcal{X}_n$
2: Calcular $f_j = f_n(X_j)$ para $1 \leq j \leq n$
3: Obtener $\hat{f}_\tau$ como el cuantil de orden τ de $f_n(\mathcal{X}_n) = \{f_{(1)}, \ldots, f_{(n)}\}$
4: Calcular los conjuntos $\mathcal{X}_{n,+}(\hat{f}_\tau)$ y $\mathcal{X}_{n,-}(\hat{f}_\tau)$

Paso 2: Búsqueda del radio óptimo $\tilde{r}_n$

5: Obtener g_n a partir de la muestra
6: Calcular $\mathcal{X}_{g_n,+}(\hat{f}_\tau)$ y $\mathcal{X}_{g_n,-}(\hat{f}_\tau)$
7: Inicializar $r \leftarrow 0.05$, $ps \leftarrow 0.005$
8: **while** $C_r(\mathcal{X}_{g_n,+}(\hat{f}_\tau)) \cap \mathcal{X}_{g_n,-}(\hat{f}_\tau) \neq \emptyset$ **do**
9: $\quad r \leftarrow r + ps$
10: **end while**
11: $\tilde{r}_n \leftarrow r - ps$

Paso 3: Estimación del conjunto de nivel

12: Calcular $C_{\tilde{r}_n}(\mathcal{X}_{n,+}(\hat{f}_\tau))$
13: **if** $C_{\tilde{r}_n}(\mathcal{X}_{n,+}(\hat{f}_\tau)) \cap \mathcal{X}_{n,-}(\hat{f}_\tau) = \emptyset$ **then**
14: $\quad \hat{\mathcal{L}}(\tau) \leftarrow C_{\tilde{r}_n}(\mathcal{X}_{n,+}(\hat{f}_\tau))$
15: **si no**
16: $\quad$ Sea $f_{(i)} = \hat{f}_\tau$
17: $\quad$ **if** $\exists j \in \{i+1, \ldots, n\}$ tal que $\mathbb{P}\Big(C_{\tilde{r}_n(f_{(j)})}(\mathcal{X}_{n,+}(f_{(j)}))\Big) \leq 1-\tau$ **then**
18: $\quad\quad l \leftarrow \text{mín}\left\{j > i : \mathbb{P}\Big(C_{\tilde{r}_n(f_{(j)})}(\mathcal{X}_{n,+}(f_{(j)}))\Big) \leq 1-\tau\right\}$
19: $\quad\quad \hat{f}_\tau \leftarrow f_{(l-1)}$
20: $\quad$ **si no**
21: $\quad\quad \hat{f}_\tau \leftarrow f_{(n)}$
22: $\quad$ **end if**
23: $\quad \hat{\mathcal{L}}(\tau) \leftarrow C_{\tilde{r}_n}(\mathcal{X}_{n,+}(\hat{f}_\tau))$
24: **end if**

Salida: Estimación del conjunto de nivel $\hat{\mathcal{L}}(\tau)$

Notemos que el parámetro ps y el valor con el que inicializamos r en la línea 7 del Algoritmo 2 se pueden modificar en función de la escala de los datos para así disminuir el tiempo de ejecución del algoritmo y obtener un estimador r adecuado.

En el paso 3 estamos comprobando que el r óptimo obtenido en el paso 2 reconstruya el conjunto de nivel de manera que tenga el porcentaje de datos que debe tener y no exista otro radio mayor con

el que se pueda construir otro conjunto de nivel que contenga una proporción de puntos $1-\tau$. Esta comprobación se realiza porque los conjuntos utilizados para estimar $\tilde{r}_n$ $(\mathcal{X}_{g_n,+}, \mathcal{X}_{g_n,-})$ no coinciden con el que se usa para obtener el conjunto de nivel $\hat{\mathcal{L}}(\tau)$ que es $\mathcal{X}_{n,+}$.

En la Figura 2.12 tenemos el conjunto de nivel estimado para $\tau = 0.2$ usando el método de la envoltura $r-$convexa con el parámetro óptimo calculado $\tilde{r}_n = 0.615$. El resultado es un valor de r intermedio entre todos los que se usaron en la Figura 2.11 y un conjunto de nivel estimado formado por tres componentes conexas.

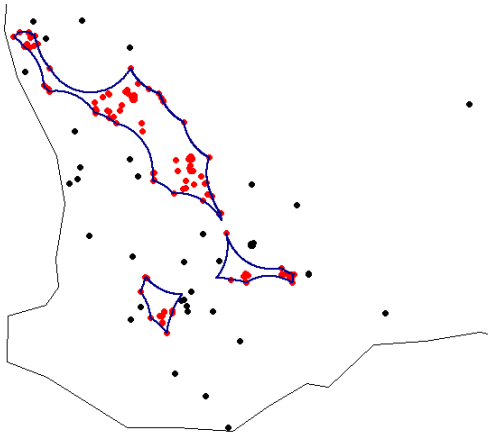

Figura 2.12: Representación de los puntos de la muestra de grevillea del conjunto $\mathcal{X}_n$, (negro), del conjunto $\mathcal{X}_{n,+}$ (rojo) y el conjunto de nivel estimado por el método de la envoltura $r-$convexa con el r estimado siguiendo el Algoritmo 2: $\tilde{r}_n = 0.615$ (azul).

Cálculo del número de componentes conexas

Para el cálculo de grupos en dos dimensiones vamos a hacer uso de la forma que toman las componentes gracias a calcular el conjunto de nivel como una envoltura $r-$convexa. Las componentes conexas que tenemos están formadas por complementarios de bolas por lo que los bordes van a ser arcos de circunferencia. Cabe destacar que en el caso de otras dimensiones de los datos este procedimiento no sería válido.

Vamos a calcular los puntos que forman la frontera del conjunto de nivel y a partir de ellos se detectan los ciclos[7] que se forman y además habrá que comprobar si alguno de los que se han obtenido están en el interior de un ciclo más grande. En caso de que eso ocurriese, no se podría contabilizar ya que no daría lugar una componente conexa nueva, sino que formaría parte de una ya existente. De esta forma obtendríamos el número de componentes conexas como el número de ciclos formados menos el de ciclos internos. Notemos que en esta forma de contabilizar componentes, un único punto aislado puede formar un grupo por si misma.

Si nos fijamos en la Figura 2.11, el número de componentes conexas varía en función del radio considerado. En la imagen de arriba a la izquierda se calculan 6 componentes conexas, una grande, dos medianas, dos pequeñas y una formada por un punto aislado que se ve en rojo. A su derecha se tienen

[7] Cuando hablamos de ciclo nos referimos a una unión de arcos cuyo punto inicial y final es el mismo.

4 componentes conexas, una grande, dos medianas y de nuevo un punto aislado. Para los valores más altos tenemos una componente conexa en cada caso. Posteriormente podría modificarse el método para que no tuviera en cuenta estos puntos aislados y los incorporase a la componente conexa más cercana.

2.3.3. Método del suavizado granulométrico

Este método es el híbrido clásico que ya fue propuesto a finales del milenio pasado en Walther (1997). Para este método la condición geométrica asumida es que una bola de radio r ruede libre en el conjunto de nivel y en la clausura de su complementario. Es decir, en otras palabras, pedimos que sea un conjunto con bordes suaves. La forma de obtener $\hat{\mathcal{L}}(\tau)$ empleando esta metodología es mediante los siguientes pasos:

1. Se elige un valor de $r > 0$.

2. Se calcula un estimador no paramétrico de la densidad f_n y calculamos el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau) = \{x \in \mathcal{X}_n : f_n(x) \geq \hat{f}_\tau\}$ donde el umbral $\hat{f}_\tau$ se estima como el cuantil de orden τ de $f_n(\mathcal{X}_n) = \{f_{(1)}, \ldots, f_{(n)}\}$

3. Una vez tenemos los subconjuntos anteriores obtenemos la estimación del conjunto de nivel como la unión de bolas de radio r cuyos centros son aquellas observaciones de $\mathcal{X}_{n,+}(\hat{f}_\tau)$ que distan más de r de las observaciones de $\mathcal{X}_{n,-}(\hat{f}_\tau)$:

$$\hat{\mathcal{L}}(\tau) = (\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus rB)^c) \oplus rB = \bigcup_{\substack{X_i \in \mathcal{X}_{n,+}(\hat{f}_\tau) \\ B[X_i,r] \cap \mathcal{X}_{n,-}(\hat{f}_\tau) = \emptyset}} B[X_i, r].$$

De nuevo aquí tenemos el problema de elegir el parámetro r óptimo aunque en este caso es un radio contenido en el conjunto de nivel, no como en el método anterior. La idea es conseguir un radio fijo que sea lo suficientemente bueno para tener un buen estimador del conjunto de nivel, para lo que necesitaremos un r lo suficientemente grande pero no demasiado ya que en algunos casos donde r tome un valor alto, el conjunto de nivel podría ser vacío.

En la Figura 2.13 se muestra la unión de bolas que da lugar al conjunto de nivel estimado empleando este método para los datos de `grevillea` y $\tau = 0.2$ para tres valores distintos de r. En la parte izquierda tenemos un valor de r pequeño, $r = 0.1$, donde se tiene que el número de centros es elevado y el conjunto de nivel está formado por numerosos grupos ya que con el valor pequeño de r no se logran formar agrupaciones. En la parte central se ha tomado un valor de r intermedio, $r = 0.25$, donde se ve que el número de centros ha disminuido pero ya podemos distinguir cuatro clusters formados por uniones de bolas que no se cortan. Por último, a la derecha, tenemos un valor alto de r, $r = 0.65$ y vemos que existen muchos menos centros que en los casos anteriores. Resalta que una gran parte de la muestra $\mathcal{X}_{n,+}$ no está contenida en el conjunto de nivel estimado. Además, únicamente tenemos una componente conexa en aquella zona con mayor cantidad de observaciones de $\mathcal{X}_{n,+}$.

Fijémonos además en que, tal como dice la definición, las bolas que forman el conjunto de nivel no contienen a puntos de la muestra $\mathcal{X}_{n,-}$, y no necesariamente tienen que contener a todas las observaciones de $\mathcal{X}_{n,+}$. Si escogiésemos un valor de r aun más grande obtendríamos un conjunto de nivel que sería el conjunto vacío ya que no podríamos construir bolas de un radio tan amplio que no contengan puntos de $\mathcal{X}_{n,-}$.

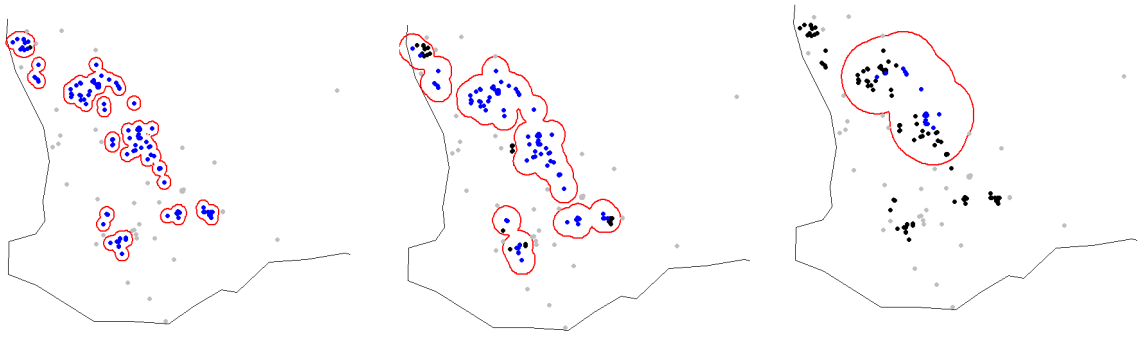

Figura 2.13: Representación de los puntos de `grevillea`, en gris las observaciones de la submuestra $\mathcal{X}_{n,-}$, en azul los centros, en negro el resto de puntos de $\mathcal{X}_{n,+}$ que no son centros y en rojo las bolas que forman los conjuntos de nivel estimados por el método de suavizado granulométrico para tres valores de r: $r = 0.1$ (izquierda), $r = 0.25$ (centro) y $r = 0.65$ (derecha).

Con este ejemplo hemos visto la importancia de r. Por ello introducimos un radio obtenido a partir de la muestra de datos que da lugar a estimaciones adecuadas de los conjuntos de nivel.

Elección del parámetro r a partir de la muestra

La elección del radio r partir de las observaciones de la muestra que vamos a presentar es la introducida en Bolón et al (2024). Lo denotamos por r_{opt} y será un estimador del radio r_0 que obtendríamos si conociésemos el conjunto de nivel real $\mathcal{L}(\tau)$:

$$r_0 = \sup\left\{r > 0 : d_H\left(\mathcal{L}(\tau) \cap (\mathcal{L}(\tau)^c \oplus rB)^c, \mathcal{L}(\tau)\right) \leq r\right\}$$

Tomaríamos como radio el mayor r que cumple que la distancia Hausdorff entre el conjunto de nivel y los puntos de la intersección del conjunto de nivel y el complementario de la unión de bolas de radio r y centro los puntos el complementario del conjunto de nivel sea menor que r.

Empleando las submuestras $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$ como sustitutos de $\mathcal{L}(\tau)$ y $\mathcal{L}(\tau)^c$ respectivamente, la definición del radio óptimo es la siguiente:

$$r_{opt} = \sup\left\{r > 0 : d_H\left(\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus rB)^c, \mathcal{X}_{n,+}(\hat{f}_\tau)\right) \leq r + h_n\right\} \tag{2.4}$$

donde h_n es una secuencia de valores que dependen de la muestra. Su función es minimizar el error que se comete por usar $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$ en vez de el conjunto de nivel desconocido. En Bolón et al (2024) se proponen dos opciones: $h_n = 1/\log(n)$ y $h_n = \sigma_n/\log(n)$ donde $\sigma_n = 2n^{-1}(n-1)^{-1}\sum_{1\leq i<j\leq n} D_2(X_i, X_j)$ en el caso de datos bidimensionales. En la segunda opción se está teniendo en cuenta la escala de los datos y, tras haber probado ambos al realizar las simulaciones, esta segunda proporcionaba resultados más cercanos a los valores reales sin aumentar el coste computacional, por lo que será la que se empleará en lo que resta de trabajo.

La interpretación de r_{opt} es el mayor radio que nos permita reconstruir desde el interior con bolas de radio r_{opt} el conjunto de nivel de manera que todos los puntos de $\mathcal{X}_{n,+}$ estén contenidos en el conjunto de nivel estimado. Así estaríamos evitando quedarnos con un valor pequeño como ocurría a la izquierda en la Figura 2.13. Además, gracias a que se pide que la distancia Hausdorff entre $\mathcal{X}_{n,+}$

y la intersección entre la submuestra anterior y el complementario de la unión de bolas de radio r y centro los puntos de $\mathcal{X}_{n,-}$ (primer conjunto de la distancia Hausdorff de 2.4) sea menor que r sumado a un factor, nos aseguramos que el radio no puede ser tan grande como para que se queden puntos de $\mathcal{X}_{n,+}$ fuera de la estimación del conjunto de nivel. Así estamos evitando que ocurra lo que veíamos en la parte derecha de la Figura 2.13.

El procedimiento a seguir es el siguiente: partimos de $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$ y nos quedamos con las mínimas distancias euclídeas de cada observación del primero a los puntos del segundo subconjunto. Estos mínimos serán los radios candidatos a óptimo. Iremos testeando si cumplen la condición de la definición en 2.4 y entre todos los que la verifiquen nos quedaremos con el mayor. Cabe destacar que el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus rB)^c$ disminuye a medida que r aumenta ya que el conjunto $\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus rB$ es más grande y por tanto el complementario será más pequeño de manera que el conjunto $\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus rB)^c$ estará formado por menos puntos de $\mathcal{X}_{n,+}$ según aumentamos el radio.

En el Algoritmo 3 se encuentra el procedimiento que acabamos de comentar y que se sigue para obtener la estimación del conjunto de nivel siguiendo lo presentado en Bolón et al (2024).

Algoritmo 3 Método del suavizado granulométrico

Entrada: Valor $\tau \in (0,1)$ y muestra iid $\mathcal{X}_n = \{X_1, \ldots, X_n\}$

Paso 1: Estimación de la densidad, umbral y submuestras $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$

Obtener el estimador no paramétrico f_n a partir de $\mathcal{X}_n$
Tomar $\hat{f}_\tau$ como el cuantil de orden τ de $f_n(\mathcal{X}_n) = \{f_{(1)}, \ldots, f_{(n)}\}$
Calcular los conjuntos $\mathcal{X}_{n,+}(\hat{f}_\tau)$ y $\mathcal{X}_{n,-}(\hat{f}_\tau)$

Paso 2: Búsqueda del radio óptimo $\tilde{r}_n$

Calcular la matriz de distancias D entre $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$
Para cada $k \in 1, ..., n_+$, $n_+ = n^{\underline{o}}$ de observaciones $\mathcal{X}_{n,+}$ se elige la mínima distancia r_k, $k \in 1, ..., n_+$ a $\mathcal{X}_{n,-}$.
Se ordenan los r_k de manera ascendiente: $r_{(1)}, ..., r_{n_+}$
Calculamos h_n a partir de $\mathcal{X}_n$.
for $k \in \{1, ..., n_+ - 1\}$ **do**
    Calculamos $\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus r_{(k)}B)^c$ y $d_H\left(\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus r_{(k)}B)^c, \mathcal{X}_{n,+}(\hat{f}_\tau)\right)$
    if $d_H\left(\mathcal{X}_{n,+}(\hat{f}_\tau) \cap (\mathcal{X}_{n,-}(\hat{f}_\tau) \oplus r_{(k)}B)^c, \mathcal{X}_{n,+}(\hat{f}_\tau)\right) \leq r_{(k)} + h_n$ **then**
        Añadimos $r_{(k)}$ conjunto de radios factibles
    end if
end for
$r_{opt} = \text{máx}\{\text{radios factibles}\}$

Paso 3: Estimación del conjunto de nivel

Se calculan los centros de las bolas como: Centros=$\{X_i \in \mathcal{X}_{n,+}(\hat{f}_\tau)$ tal que $B[X_i, r_{opt}] \cap \mathcal{X}_{n,-}(\hat{f}_\tau) = \emptyset\}$
$\hat{\mathcal{L}}(\tau) = \bigcup_{C \in \text{Centros}} B[C, r_{opt}]$

Salida: Estimación del conjunto de nivel $\hat{\mathcal{L}}(\tau)$

Si obtenemos el radio a partir de la nuestra muestra de `grevillea` siguiendo los pasos anteriores, llegamos a $r_{opt} = 0.321$. El resultado del estimador del conjunto de nivel es el que se muestra en la Figura 2.14 con el que obtenemos tres componentes conexas.

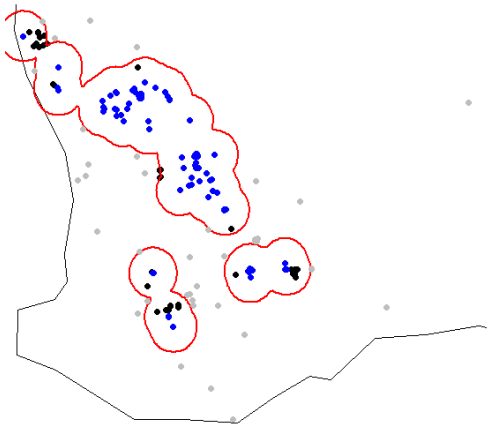

Figura 2.14: Representación de los puntos de `grevillea`, en gris las observaciones de la submuestra $\mathcal{X}_{n,-}$, en azul los centros y en rojo las bolas que forman los conjuntos de nivel estimados por el método de suavizado granulométrico con r óptimo, $r_{opt} = 0.321$.

Cálculo del número de componentes conexas

Una vez hemos realizado la estimación del conjunto de nivel con el método anterior y teniendo un valor de radio r_{opt} podemos obtener el número de componentes conexas en las que está dividido el conjunto de nivel estimado $\hat{\mathcal{L}}(\tau)$. Para ello vamos a usar que la estimación es una unión de bolas de radio fijo r_{opt} y por lo tanto, dos puntos contenidos en el conjunto de nivel pertenecerán a la misma componente conexa si se encuentran a menos de dos veces el valor del radio r_{opt}; en caso contrario, pertenecerán a dos componentes conexas distintas. Gráficamente cada componente conexa estará formada por una unión de bolas que se cortan. En R haremos uso de la librería `igraph`; primero vamos a obtener los bordes del conjunto de nivel empleando los centros de las bolas y el radio fijo r_{opt} y a partir de estos bordes podemos crear un grafo al que se le pueden calcular las componentes conexas gracias a la función `components()`. Esta forma de estimar componentes conexas es igual a la presentada en CFF pero automática con funciones de R, la diferencia es la estimación del conjunto de nivel.

Así, gráficamente podíamos saber el número de componentes conexas de los conjuntos de nivel estimados en la Figura 2.13 tal como comentábamos previamente: en la parte izquierda hay once grupos, en el centro cuatro y en la derecha únicamente uno.

Capítulo 3

Estudio de simulación

Tras presentar diferentes métodos de estimación de componentes conexas, interesa saber cómo de bien funcionan en diferentes casos. Para ello, vamos a elegir cuatro modelos para los que conozcamos el número real de componentes conexas en función del valor de τ y para distintos tamaños de muestra n se hará un estudio del funcionamiento de los métodos. Los procedimientos para calcular componentes conexas elegidos son algunos de los presentados en el Capítulo 2: el primero de ellos será el método CFF con valor de $\varepsilon = \varepsilon_2(\beta = 0.1)$; se ha elegido únicamente este por simplicidad en la presentación de los datos, ya que tras realizar las simulaciones, los resultados para ε_1 y ε_3 son muy similares. En segundo lugar la metodología basada en `pdfcluster` tomando como matriz de suavizado H por defecto. Después tenemos el método de la envoltura $r-$convexa con el valor de r seleccionado con los datos y por último, el método del suavizado granulométrico con el valor de r también dependiente de la muestra. Los modelos sobre los que se van a aplicar estos procedimientos son mixturas de normales con distinto número de modas y diferentes separaciones entre ellas, que representamos de la forma:

$$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2, \rho_i)$$

donde w_i es el peso que tiene la normal $i-$ésima en el modelo, donde $\sum_{i=1}^{k} w_i = 1$; μ_{i1} y μ_{i2} las componentes del vector de medias de la normal $i-$ésima; σ_{i1}^2 y σ_{i2}^2 son las componentes de la diagonal de la matriz de covarianzas de la $i-$esima normal y ρ_i es la correlación de la normal $i-$esima Todos los modelos, junto con la notación empleada están sacados de Wand y Jones (1993) con ligeras modificaciones en el segundo de los modelos para que tenga una distribución similar a los datos de seísmos. Para más información de los modelos se puede consultar Wand y Jones (1993). Definimos los modelos que se van a emplear en la Tabla 3.1.

El primero y segundo de los modelos están formados por la mixtura de dos normales, el segundo de ellos con las modas ubicadas de forma similar a los conjuntos de datos de seísmos que se van a emplear en el Capítulo 4. El tercer modelo es una mixtura de tres normales y el cuarto de cuatro. En la Figura 3.1 podemos ver una representación tridimensional de cada una de las densidades que nos hacen intuir que el número de modas de cada una es igual al número de normales que se han mezclado. También en la Figura 3.2 tenemos una representación bidimensional de los conjuntos de nivel de los distintos modelos para los valores $\tau = 0.8, 0.5, 0.3$ que será para los que se calculará el número de componentes conexas. Cuanto más oscura sea la zona implica que hay mayor concentración de puntos, y es por ello que a mayor valor de τ nos quedamos con menos partes claras y predominan las zonas de colores más oscuros.

Modelo	$\sum_{i=1}^{k} w_i \mathcal{N}(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2, \rho_i)$
Bimodal I	$\frac{1}{2}\mathcal{N}(-1, 0, \frac{4}{9}, \frac{4}{9}, 0) + \frac{1}{2}\mathcal{N}(1, 0, \frac{4}{9}, \frac{4}{9}, 0)$
Bimodal II	$\frac{1}{2}\mathcal{N}(-3, 2, \frac{4}{9}, \frac{4}{9}, -0.3) + \frac{1}{2}\mathcal{N}(3, -2, 1, 2, 0.35)$
Trimodal	$\frac{1}{3}\mathcal{N}(0, 0, 1, 1, 0.8) + \frac{1}{3}\mathcal{N}(6, 0, \frac{1}{2}, \frac{1}{2}, -\frac{1}{2}) + \frac{1}{3}\mathcal{N}(3, 6, \frac{1}{2}, \frac{1}{2}, 0)$
Quadrimodal	$\frac{1}{4}\mathcal{N}(-1.2, 1.5, \frac{4}{9}, \frac{4}{9}, \frac{2}{5}) + \frac{1}{4}\mathcal{N}(-1.2, -1.5, \frac{4}{9}, \frac{4}{9}, \frac{3}{5}) + \frac{1}{4}\mathcal{N}(1.2, -1.5, \frac{4}{9}, \frac{4}{9}, -\frac{7}{10}) + \frac{1}{4}\mathcal{N}(1.2, 1.5, \frac{4}{9}, \frac{4}{9}, -\frac{1}{2})$

Tabla 3.1: Modelos empleados para las simulaciones.

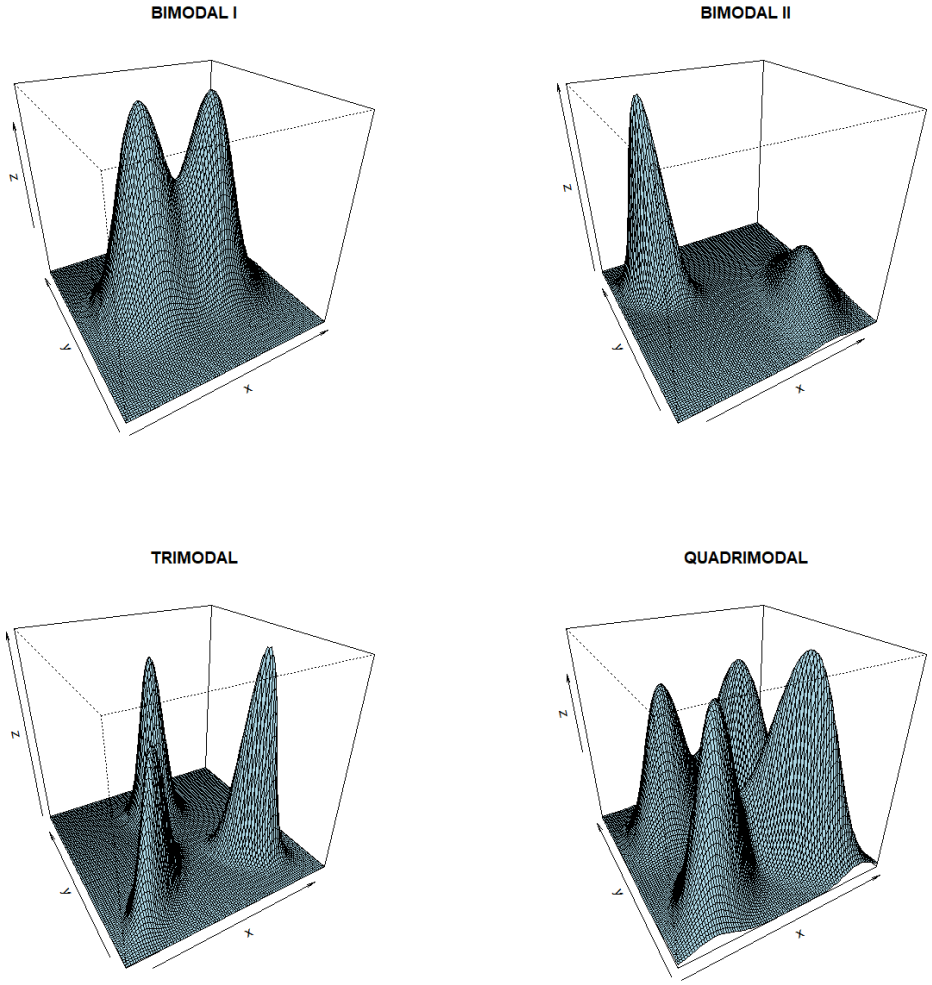

Figura 3.1: Representación tridimensional de los cuatro modelos definidos que se emplean en el estudio de simulación.

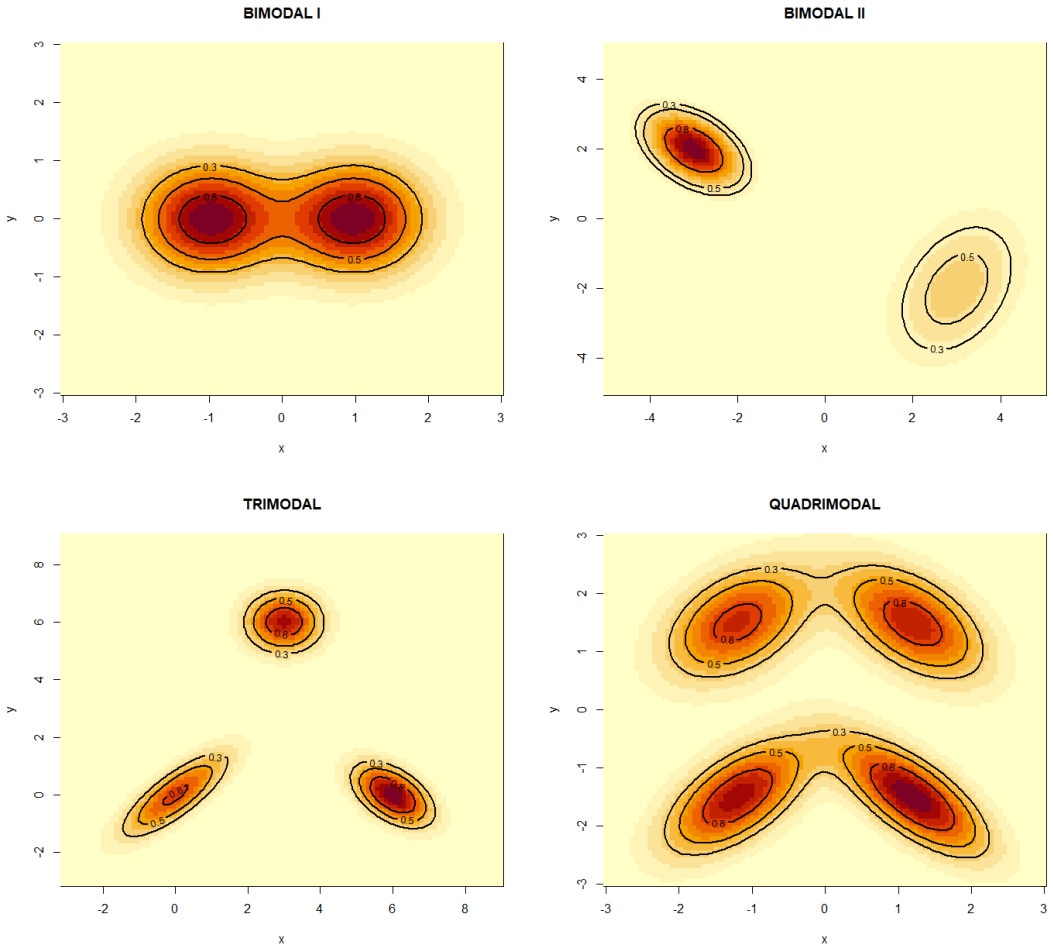

Figura 3.2: Representación bidimensional de los conjuntos de nivel para los cuatro modelos definidos y valores de $\tau = 0.8, 0.5, 0.3$.

Además en la Figura 3.2 se pueden ver el número de componentes conexas reales que hay para los diferentes contenidos en probabilidad y modelos. En la Tabla 3.2 se resumen las componentes reales en cada uno de los casos que se van a estudiar. Estos serán los valores con los que compararemos los resultados de las simulaciones.

En total se realizarán 300 simulaciones de los modelos anteriores con cada uno de los tamaños muestrales siguientes: $n = 100, 200, 500, 1000$. Se estimará el conjunto de nivel para los valores de τ mencionados y usando los métodos presentados se obtendrá un estimador del número de componentes conexas para cada una de las simulaciones. Como resultado, para cada valor de τ y cada modelo, se mostrará una tabla con la media del número de componentes conexas calculadas y, entre paréntesis, la desviación típica del conjunto. Con estas tablas tendremos una idea de qué metodología es la más acertada en cada caso y compararemos su funcionamiento para los distintos modelos y veremos cómo afecta el tamaño muestral y los valores de τ. Así, en un caso como en el del capítulo siguiente en el que no conocemos el número real de componentes conexas podríamos tener una guía del método que mejor nos podría funcionar y de su variabilidad.

		τ		
		0.8	0.5	0.3
Modelos	Bimodal I	2	1	1
	Bimodal II	1	2	2
	Trimodal	3	3	3
	Cuadrimodal	4	4	2

Tabla 3.2: Número de componentes conexas teóricas que hay para cada uno de los modelos según el contenido en probabilidad $1-\tau$ considerado.

Tras esta introducción, se exponen y comentan los resultados obtenidos tras la realización de las simulaciones presentados en múltiples tablas.

Comenzamos con el primer modelo: BIMODAL I. En la Tabla 3.3 tenemos los resultados para $\tau = 0.8$. Lo primero que llama la atención son los 20 grupos detectados por CFF para $n = 100$. Esto es así porque empieza con esos 20 grupos iniciales pero seguramente debido al pequeño tamaño muestral, el método no es capaz de agrupar. Sin embargo para n mayores las estimaciones mejoran y a medida que aumenta el tamaño muestral se va ajustando más la media al valor real 2 y disminuyendo la desviación típica. Esto ocurre de igual modo para el resto de métodos pero vemos que mientras que los tres primeros se quedan en general por debajo del valor real, el suavizado granulométrico sobreestima para muestras pequeñas y medianas y es el que mayor desviación tiene a excepción de para $n = 1000$. En el otro lado, CFF es la de menor variabilidad teniendo en cuenta todos los valores de n.

BIMODAL I $\tau = 0.8$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	20(0)	1.63(0.58)	1.45(0.74)	2.31(1.67)	2
200	1.98(0.67)	1.79(0.61)	1.58(0.66)	2.22(1.43)	
500	1.86(0.40)	1.97(0.49)	1.75(0.53)	2.13(0.91)	
1000	1.93(0.28)	2.03(0.41)	1.86(0.40)	2.04(0.34)	

Tabla 3.3: Número medio y desviación típica de las componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.8$ en el modelo BIMODAL I.

Seguimos con el mismo modelo pero para $\tau = 0.5$, es decir, conjuntos de nivel con mayor contenido en probabilidad que en el caso anterior. El número real de componentes conexas es una a diferencia

de antes y los resultados de las simulaciones se pueden ver en la Tabla 3.4. De nuevo tenemos que la desviación disminuye a medida que se aumenta el tamaño muestral para todos los métodos. En este caso vemos que el método que más se ajusta a la realidad es el de la $r-$convexidad con valores muy cercanos a uno en la media y los más bajos en desviación; en el lado contrario, es el suavizado granulométrico, de nuevo, el que presenta desviaciones más altas. En general, las estimaciones son mejores para $\tau = 0.5$.

BIMODAL I $\tau = \mathbf{0.5}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	1.55(0.54)	1.16(0.38)	1.10(0.36)	1.15(0.82)	1
200	1.32(0.49)	1.14(0.35)	1.08(0.33)	1.21(0.61)	
500	1.17(0.38)	1.04(0.22)	1.03(0.23)	1.14(0.37)	
1000	1.07(0.25)	1.03(0.17)	1.01(0.10)	1.13(0.34)	

Tabla 3.4: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.5$ en el modelo BIMODAL I.

Por último tenemos los resultados para $\tau = 0.3$, donde estimaremos los conjuntos de nivel con mayor contenido de los que vamos a considerar. Los resultados de las simulaciones están en la Tabla 3.5. Se puede ver como estamos ante las desviaciones más bajas de entre todas las simulaciones del Modelo BIMODAL I, llegando a ser cero en los tamaños muestrales más elevados. Esto se debe a que hay más observaciones de la muestra $\mathcal{X}_n, +$ que se pueden emplear y por lo tanto más precisas serán las estimaciones. En este caso, el método de pdfCluster es el que mejor ajusta el número de componentes conexas al tener una media casi perfecta y la desviación más baja. De nuevo, el suavizado granulométrico es la metodología que presenta mayor varianza excepto para $n = 1000$.

BIMODAL I $\tau = \mathbf{0.3}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	1.23(0.43)	1.02(0.14)	1.03(0.26)	1.14(0.60)	1
200	1.09(0.30)	1.01(0.08)	1.01(0.12)	1.06(0.49)	
500	1.02(0.15)	1.00(0.08)	1.00(0)	1.02(0.29)	
1000	1.00(0)	1.01(0.08)	1.00(0)	1.00(0)	

Tabla 3.5: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.3$ en el modelo BIMODAL I.

De esta manera finalizamos la presentación de resultados para BIMODAL I donde hemos visto que a medida que el contenido en probabilidad aumenta, esto es, τ disminuye, las estimaciones son mejores.

Pasamos ahora al segundo modelo: BIMODAL II. Para $\tau = 0.8$ tenemos que el número de componentes conexas real es uno y los resultados de las simulaciones se muestran en la Tabla 3.6. Para $n = 500$ y $n = 1000$ se estima 1 componente en todos los casos para pdfCluster, el método de la $r-$convexidad y el suavizado granulométrico. CFF también proporciona buenos resultados pero peores en todos los casos que el resto de métodos. En este caso es el método híbrido de la $r-$convexidad el que mejor estima el número de componentes conexas.

BIMODAL II $\tau = \mathbf{0.8}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	20(0)	1.06(0.23)	1.04(0.20)	1.13(0.56)	1
200	1.33(0.64)	1.02(0.12)	1.00(0.06)	1.06(0.23)	
500	1.01(0.11)	1.00(0)	1.00(0)	1.00(0)	
1000	1.01(0.11)	1.00(0)	1.00(0)	1.00(0)	

Tabla 3.6: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.8$ en el modelo BIMODAL II.

Tomando $\tau = 0.5$ y $\tau = 0.3$, el número real son dos componentes conexas y los resultados de las simulaciones se pueden ver en las Tabla 3.7 y 3.8 respectivamente. La metodología que menos desviación típica presenta y mejor ajusta el número de componentes conexas es pdfCluster para ambos umbrales, llegando a realizar las estimaciones en las 300 simulaciones correctamente para los mayores

BIMODAL II $\tau = \mathbf{0.5}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	1.88(0.43)	1.79(0.41)	1.71(0.69)	2.37(1.20)	2
200	1,96(0.26)	1.92(0.28)	1.91(0.43)	2.34(0.78)	
500	2.00(0.13)	2.00(0.08)	2.03(0.21)	2.33(0.56)	
1000	2.01(0.08)	2.00(0)	2.01(0.16)	2.30(0.46)	

Tabla 3.7: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.5$ en el modelo BIMODAL II.

tamaños muestrales. Los métodos híbridos son los que dan lugar a mayor varianza: el método del suavizado granulométrico en el primer caso y el segundo para los dos valores mayores de n y $r-$convexo para los n más pequeños . Además, al igual que anteriormente, para $\tau = 0.5$, a medida que aumentamos n se disminuye la varianza pero es llamativo como en $\tau = 0.3$ y la metodología CFF, la varianza oscila entre 0.16 y 0.17 sin importar el tamaño muestral. Es señal de que esa varianza probablemente no disminuiría aunque aumentásemos el valor de n.

BIMODAL II $\tau = \mathbf{0.3}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	2.03(0.16)	1.97(0.16)	1.99(0.86)	2.26(0.64)	2
200	2.03(0.17)	2.00(0.06)	2.04(0.62)	2.18(0.55)	
500	2.03(0.16)	2.00(0)	2.05(0.30)	2.10(0.31)	
1000	2.03(0.17)	2.00(0)	2.02(0.22)	2.08(0.28)	

Tabla 3.8: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.3$ en el modelo BIMODAL II.

Pasemos ya al tercer modelo y comenzamos con el valor de $\tau = 0.8$, cuyos resultados se muestran en la Tabla 3.9. Vemos que para $n = 100$ el número medio de componentes que se estima se queda muy por debajo del número real. Esto se va compensando en todos los métodos a excepción de CFF al aumentar el tamaño muestral, además de disminuir la desviación. El método que más se acerca al valor real 3 es el suavizado granulométrico aunque su desviación es a su vez la más alta excepto para el valor más alto de n. Llama la atención lo comentado sobre CFF que no aumenta el número medio de componentes estimados ni disminuye la desviación por lo que es probable que esta metodología no realizaría buenas estimaciones aunque sigamos aumentando el tamaño muestral.

TRIMODAL $\tau = \mathbf{0.8}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gr	Real
100	20(0)	1.74(0.58)	1.85(0.66)	2.33(0.90)	3
200	2.39(0.75)	1.98(0.58)	2.03(0.69)	2.52(0.82)	
500	2.33 (0.66)	2.16(0.45)	2.13(0.69)	2.59(0.71)	
1000	2.34(0.71)	2.26(0.36)	2.27(0.67)	2.62(0.55)	

Tabla 3.9: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.8$ en el modelo TRIMODAL.

En las Tablas 3.10 y 3.11 tenemos los resultados de las simulaciones para el modelo **TRIMODAL** y los umbrales $\tau = 0.5$ y $\tau = 0.3$ respectivamente. Podemos ver que las estimaciones del número de componentes conexas son mucho más cercanas al valor real tres que para el umbral $\tau = 0.8$. La metodología pdfCluster es capaz de estimar perfectamente en todas las simulaciones con $n = 1000$ para ambos umbrales y con $n = 500$ para el τ más pequeño. Además es el de menor desviación para $\tau = 0.5$, mientras que el suavizado granulométrico lo es para $\tau = 0.3$ los dos tamaños muestrales menores. El método con mayor desviación típica es para ambos umbrales el de la envoltura $r-$convexa. Por último resaltamos que aunque CFF no realizaba buenas estimaciones para $\tau = 0.8$, estas mejoran para los valores de $\tau = 0.5$ y $\tau = 0.3$.

TRIMODAL $\tau = \mathbf{0.5}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gr	Real
100	2.80(0.53)	2.70(0.47)	2.43(0.72)	3.02(0.72)	3
200	2.97(0.37)	2.88(0.33)	2.35(0.69)	3.09(0.43)	
500	3.01(0.17)	2.99(0.08)	2.64(0.67)	3.04(0.19)	
1000	3.04(0.19)	3.00(0.00)	2.76(0.55)	3.01(0.09)	

Tabla 3.10: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.5$ en el modelo TRIMODAL.

TRIMODAL $\tau = \mathbf{0.3}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gr	Real
100	3.01(0.23)	2.52(0.59)	2.48(0.82)	3.02(0.14)	3
200	3.03(0.17)	2.89(0.32)	2.59(0.79)	3.02(0.13)	
500	3.04(0.20)	3.00(0)	2.70(0.65)	3.01(0.10)	
1000	3.05(0.22)	3.00(0)	2.90(0.42)	3.00(0.06)	

Tabla 3.11: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.3$ en el modelo TRIMODAL.

Por último tenemos las simulaciones realizadas con muestras del modelo **QUADRIMODAL** que para el caso de $\tau = 0.8$ y $\tau = 0.5$, son los resultados de las Tablas 3.12 y 3.13. El número de componentes conexas reales es el mayor de todos los modelos: cuatro. Vemos que en ambos casos las estimaciones del número de componentes se encuentran por debajo del valor real, siendo la diferencia mayor para el valor más alto de τ. El método que mayor desviación presenta y el que resultados más cercanos a

cuatro estima para $\tau = 0.8$ es el suavizado granulométrico. Para $\tau = 0.5$ sigue siendo el suavizado granulométrico la metodología que se queda más próxima a cuatro pero es CFF para todos los tamaños muestrales excepto para $n = 100$. En ambos casos la menor desviación ha sido de las estimaciones con pdfCluster aunque el número medio de componentes es el que más se aleja del real para $\tau = 0.8$ y es el segundo mejor para $\tau = 0.5$.

QUADRIMODAL $\tau = \mathbf{0.8}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	20(0)	2.18(0.68)	2.27(0.91)	2.99(1.24)	4
200	3.04(0.85)	2.79(0.75)	2.91(0.93)	3.28(1.21)	
500	3.43(0.68)	3.42(0.58)	3.48(0.68)	3.81(0.90)	
1000	3.70(0.58)	3.77(0.43)	3.83(0.45)	3.95(0.59)	

Tabla 3.12: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.8$ en el modelo QUADRIMODAL.

QUADRIMODAL $\tau = \mathbf{0.5}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	2.98(0.94)	3.08(0.86)	2.86(0.94)	3.49(1.07)	4
200	3.30(0.94)	3.71(0.56)	3.21(0.80)	3.93(0.73)	
500	3.52(0.90)	3.91(0.30)	3.38(0.62)	3.99(0.48)	
1000	3.70(0.74)	3.98(0.14)	3.50(0.56)	4.02(0.29)	

Tabla 3.13: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.5$ en el modelo QUADRIMODAL.

Por último tenemos los resultados para $\tau = 0.3$ en la Tabla 3.14 donde vemos que el número de componentes reales es dos. Lo primero destacable es que usando pdfCluster, a medida que aumenta el tamaño muestral también lo hace el número de componentes conexas estimadas alejándose del valor real aunque la desviación disminuye. En el resto de metodologías, para $n = 100$ el número medio de componentes estimadas es alto y va bajando al aumentar el tamaño muestral, al igual que la desviación. El método de la envoltura $r-$convexa es el de menor desviación en este caso además de que es el que más ajusta el número de componentes conexas estimadas al valor real para $n = 100, 200, 1000$. En el lado contrario, la metodología CFF es la que más desviación presenta.

En general, para este modelo **QUADRIMODAL** se han obtenido los resultados con las desviaciones más elevadas para todos los contenidos en probabilidad.

QUADRIMODAL $\tau = \mathbf{0.3}$					
n	CFFε_2	pdfCluster	$r-$conv	suav-gran	Real
100	2.53(1.13)	2.14(0.82)	2.32(0.72)	2.97(0.96)	2
200	2.32(1.11)	2.63(0.73)	2.31(0.58)	2.92(0.81)	
500	1.9(1.03)	2.79(0.60)	2.22(0.45)	2.86(0.67)	
1000	1.48(0.86)	2.74(0.62)	2.20(0.49)	2.71(0.59)	

Tabla 3.14: Número medio de componentes conexas estimadas con 300 simulaciones para $n \in 100, 200, 500, 1000$ y $\tau = 0.3$ en el modelo QUADRIMODAL.

Tras analizar los resultados, podemos concluir varias cosas. Primeramente, notamos que, en general, las desviaciones disminuyen al aumentar el tamaño muestral y además, en muchos casos son más altas para mayor valor real de número de componentes conexas y, dentro de un mismo modelo e igual valor real, a mayor contenido en probabilidad, menor es la desviación, es decir, más exactas son las estimaciones. Esto lo podemos ver en los resultados de los dos primeros modelos donde vemos que se alcanza la desviación nula en varias ocasiones. Comparando modelos, pdfCluster es el que mejores resultados proporciona en general pero para n grande, el suavizado granulométrico lo supera en numerosas ocasiones. Sin embargo este último, es el que tiene en la mayoría de los casos las desviaciones más elevadas. El método con la desviación más estable independientemente del modelo es el de la envoltura $r-$convexa. La metodología CFF no tiene un correcto funcionamiento para tamaños muestrales pequeños ($n = 100$) y valores de τ altos pero sí proporciona resultados correctos en el resto de casos, en especial, en los dos primeros modelos con menor número de componentes conexas reales para tamaños muestrales distintos de $n = 100$.

Capítulo 4

Aplicación a datos reales

Una vez hecho el estudio de simulaciones presentado en el capítulo anterior y obtenidos los resultados sobre las estimaciones del número de componentes conexas, es interesante realizar una aplicación de los procedimientos testeados sobre muestras de datos reales como pueden ser los seísmos expuestos en el primer capítulo. Se aplicarán los cuatro métodos de reconstrucción de conjuntos de nivel y obtención de componentes conexas que se pusieron a prueba en las simulaciones: plug-in usando el contador de componentes conexas de CFF y pdfCluster, el método de la envoltura $r-$convexa y el del suavizado granulométrico. Los valores de τ para los que se van a obtener los conjuntos de nivel son los mismos que en el Capítulo 3: $\tau = 0.8, 0.5, 0.3$ y las dos muestras sobre los que se calculan serán, por un lado, los terremotos que han tenido lugar en Grecia entre 1972 y 1992 (antiguos, S.XX) y por otro, los ocurridos entre 2003 y 2023 (actuales, S.XXI). La primera de las muestras tiene un total de 1284 observaciones mientras que la segunda 1430 y de todas las variables que se presentaron en el Capítulo 1 solo se van a utilizar las coordenadas de los terremotos. Notemos que el número de observaciones es similar al mayor tamaño muestral que usamos en las simulaciones y con el que teníamos unos resultados muy cercanos al valor real en media y con una desviación típica pequeña. Por lo tanto se podría esperar que los métodos también proporcionen buenos resultados sobre ambas muestras.

Obtenemos la función f_n usando un kernel normal bidimensional, y para cada valor de τ mencionados anteriormente y cada muestra calculamos el estimador $\hat{f}_\tau$ correspondiente obteniendo:

Datos S.XX:	$\hat{f}_{0.8} = 0.104$	$\hat{f}_{0.5} = 0.0529$	$\hat{f}_{0.3} = 0.0343$
Datos S.XXI:	$\hat{f}_{0.8} = 0.0898$	$\hat{f}_{0.5} = 0.0489$	$\hat{f}_{0.3} = 0.0294$

Podemos ver como en todos los casos, el umbral estimado es ligeramente menor para la muestra de datos actuales pero los valores son bastante similares para ambas muestras. Así, con estos valores podemos obtener el conjunto de X_i tales que $f_n(X_i)$ se encuentra por encima del umbral para cada una de las muestras y τ, es decir, los puntos de la muestra pertenecientes a $\hat{\mathcal{L}}(\tau)$ o, siguiendo la notación de los métodos híbridos, el conjunto $\mathcal{X}_{n,+}$. La representación de estas observaciones se encuentran en la Figura 4.1: En la parte superior se muestran en rojo los puntos por encima del estimador del umbral para la muestra de los datos del siglo pasado para, de izquierda a derecha, $\tau = 0.8, 0.5, 0.3$. En la parte inferior tenemos lo mismo en color azul para los datos actuales. En primer lugar, a medida que disminuye τ, mayor número de observaciones estarán contenidas en el conjunto de nivel; por eso de izquierda a derecha vemos que las zonas coloreadas aumentan. En segundo lugar se puede comparar

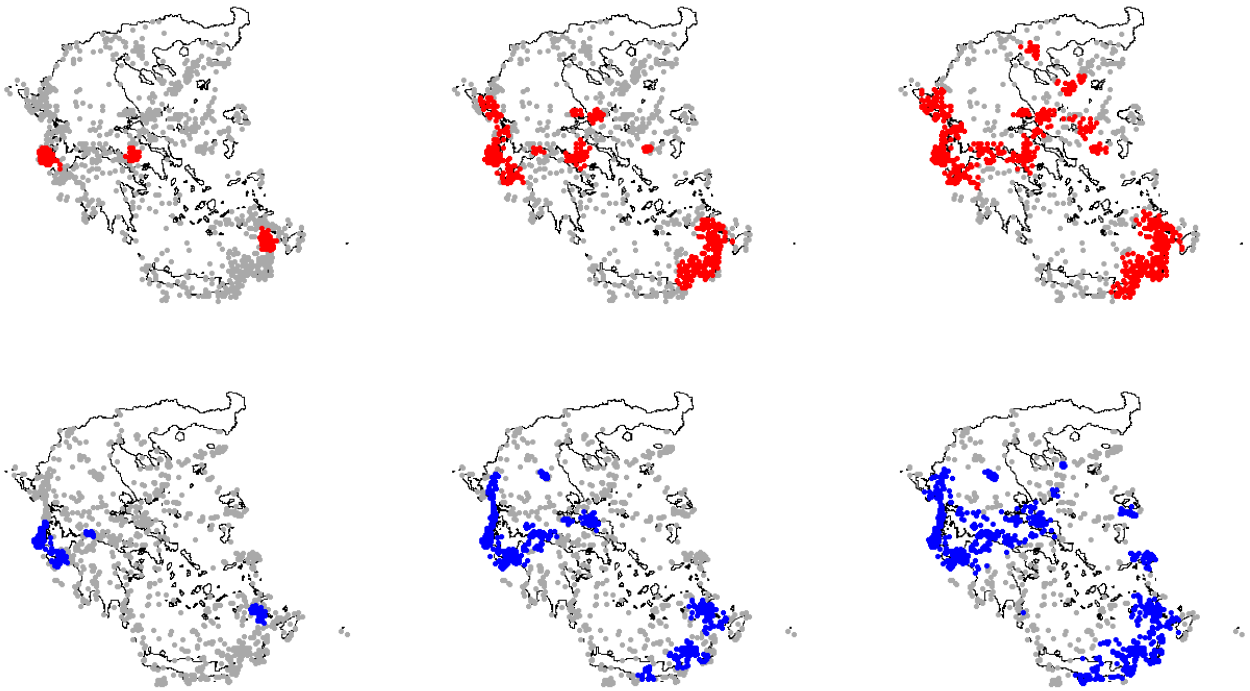

Figura 4.1: Observaciones por encima de $\hat{f}_\tau$ con $\tau = 0.8, 0.5, 0.3$ de izquierda a derecha para los datos de seísmos en Grecia entre 1972 y 1992 (parte superior, en rojo) y entre 2003 y 2023 (parte inferior, en azul).

lo obtenido para los datos del siglo XX. con los actuales. Se puede observar que para $\tau = 0.8$ hay tres zonas coloreadas de las cuales dos coinciden en ubicación y la que está en el medio es la que difiere: en los datos antiguos se encuentra más a al este que en los actuales. Para $\tau = 0.5$ podemos ver para las dos muestras que hay una zona coloreada al sureste del país y otra ubicada en el centro y noroeste. Por tanto vemos que son bastante similares los resultados para ambas. Al tomar $\tau = 0.3$, laz zonas de alta densidad para los datos del siglo XX. se encuentran en el sur, y hacia el norte, mientras que en los del siglo XXI no alcanzan zonas tan al norte pero en compensación las islas situadas al este están afectadas por mayor número de seísmos.

Una vez tenemos esta idea inicial y hemos obtenido las observaciones pertenecientes a $\hat{\mathcal{L}}(\tau)$ (según la notación de plug-in) y a $\mathcal{X}_{n,+}$ (siguiendo la notación de los métodos híbridos) para cada umbral, vamos a estimar el número de componentes conexas usando las metodologías mencionadas

4.1. CFF

Al igual que en las simulaciones únicamente se va a emplear el parámetro ϵ_2 para estimar componentes. Así, para los diferentes umbrales τ y las dos muestras de seísmos, se obtienen los valores de ε_2 que se muestran en la Tabla 4.1. Como era de esperar el valor de ε aumenta a medida que τ disminuye. Esto es, estaríamos estimando el número de componentes conexas generadas por una unión de bolas

de un radio más grande a medida que hay mayor contenido en probabilidad en el conjunto de nivel, es decir, menor τ.

Valor de ε_2

τ	0.8	0.5	0.3
Muestra S.XX	0.2499	1.0892	2.8709
Muestra S.XXI	0.3022	1.2518	2.4850

Tabla 4.1: Valores del parámetro ε_2 para las dos muestras de terremotos y los tres umbrales: $\tau = 0.8, 0.5, 0.3$.

Una vez tenemos el valor de ε para cada caso se pueden estimar el número de componentes conexas siguiendo los pasos del Algoritmo 1. El resultado de los grupos formados se muestra gráficamente en la Figura 4.2 y el número de componentes estimadas se puede consultar en la Tabla 4.2.

Nº de componentes conexas estimadas

τ	0.8	0.5	0.3
Muestra S.XX	3	3	2
Muestra S.XXI	4	2	2

Tabla 4.2: Número de componentes conexas estimadas para las dos muestras de terremotos y los tres umbrales: $\tau = 0.8, 0.5, 0.3$ empleando el método CFF con el parámetro ε_2.

Vemos que únicamente coinciden para $\tau = 0.3$, donde el número de componentes conexas estimadas es dos. Fijándonos en la parte inferior de la Figura 4.2 se tiene que hay un grupo en el sureste del país y otro más hacia el norte en ambos casos. La mayor diferencia es que las islas del este están más afectadas en la actualidad formando parte del conjunto de nivel estimado mientras que a partir de los registros del siglo XX, estas zonas no formarían parte del conjunto de nivel. Para $\tau = 0.5$ tenemos tres componentes conexas para la muestra del siglo XX y dos para la actual. La diferencia está, tal como vemos en las dos imágenes centrales de la Figura 4.2, en que las zonas de alta densidad situadas en la zona norte, en el caso de la muestra del siglo pasado forman parte de dos componentes conexas distintas mientras que en la actual todas las observaciones están en una única. Por último, las componentes conexas estimadas para $\tau = 0.8$ son tres para los datos históricos y cuatro para los actuales. Fijándonos en la Figura 4.2, en la muestra del siglo pasado, las componentes conexas coinciden en localización con las obtenidas para $\tau = 0.5$. Esto no ocurre para los datos actuales donde se tienen cuatro agrupaciones de observaciones, una en el sur y otras tres bastante cercanas en la parte noroeste; en la zona donde, para mayor contenido en probabilidad, teníamos una única componente conexa, tomando para $\tau = 0.8$ se han estimado tres de menor tamaño.

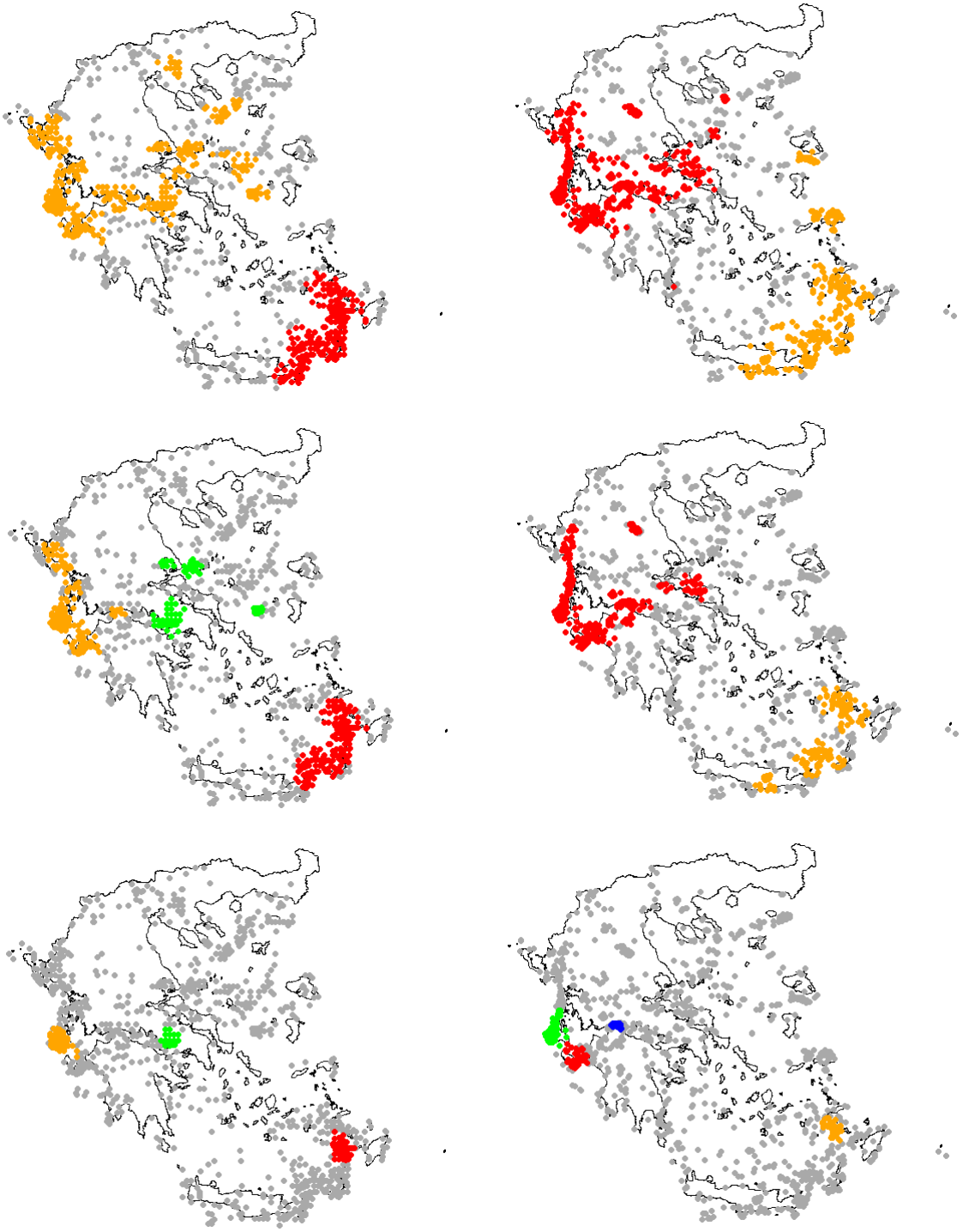

Figura 4.2: Grupos formados usando el método de estimación del número de componentes conexas CFF para los datos de seísmos en Grecia entre 1972 y 1992 (izquierda) y entre 2003 y 2023 (derecha). En la fila superior están los resultados para $\tau = 0.3$, en el medio, $\tau = 0.5$ y en la inferior $\tau = 0.8$.

4.2. pdfCluster

Para este segundo método vamos a necesitar la triangulación de Delaunay calculada para ambas muestras. Estas se pueden ver en la Figura 4.3, a la izquierda la obtenida para la muestra de seísmos del siglo pasado y a la derecha la del actual.

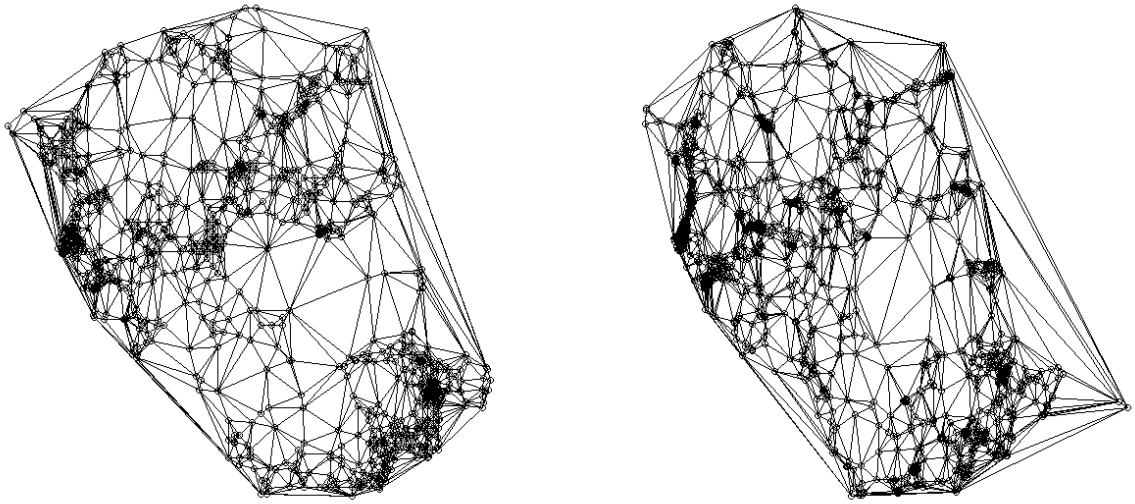

Figura 4.3: Triangulación de Delaunay obtenida a partir de la muestra de seísmos registrados entre 1972 y 1992 (izquierda) y entre 2003 y 2023 (derecha).

Tras aplicar el método mediante la función `pdfCluster` de R sobre ambas muestras con los valores de τ anteriormente mencionados, se han estimado el número de componentes conexas que se muestra en la Tabla 4.3 en cada caso. Además, los grupos generados se pueden observar gráficamente en la Figura 4.4 donde tenemos a la izquierda los grupos para la muestra de seísmos en el siglo XX para $\tau = 0.3$ arriba, $\tau = 0.5$ en el centro y $\tau = 0.3$ en la parte inferior; en la columna derecha se encuentra lo mismo pero para la muestra de datos del siglo XXI.

Nº de componentes conexas estimadas

τ	0.8	0.5	0.3
Muestra S.XX	3	2	2
Muestra S.XXI	2	2	2

Tabla 4.3: Número de componentes conexas estimadas para las dos muestras de terremotos y los tres umbrales: $\tau = 0.8, 0.5, 0.3$ mediante el método denotado por pdfCluster.

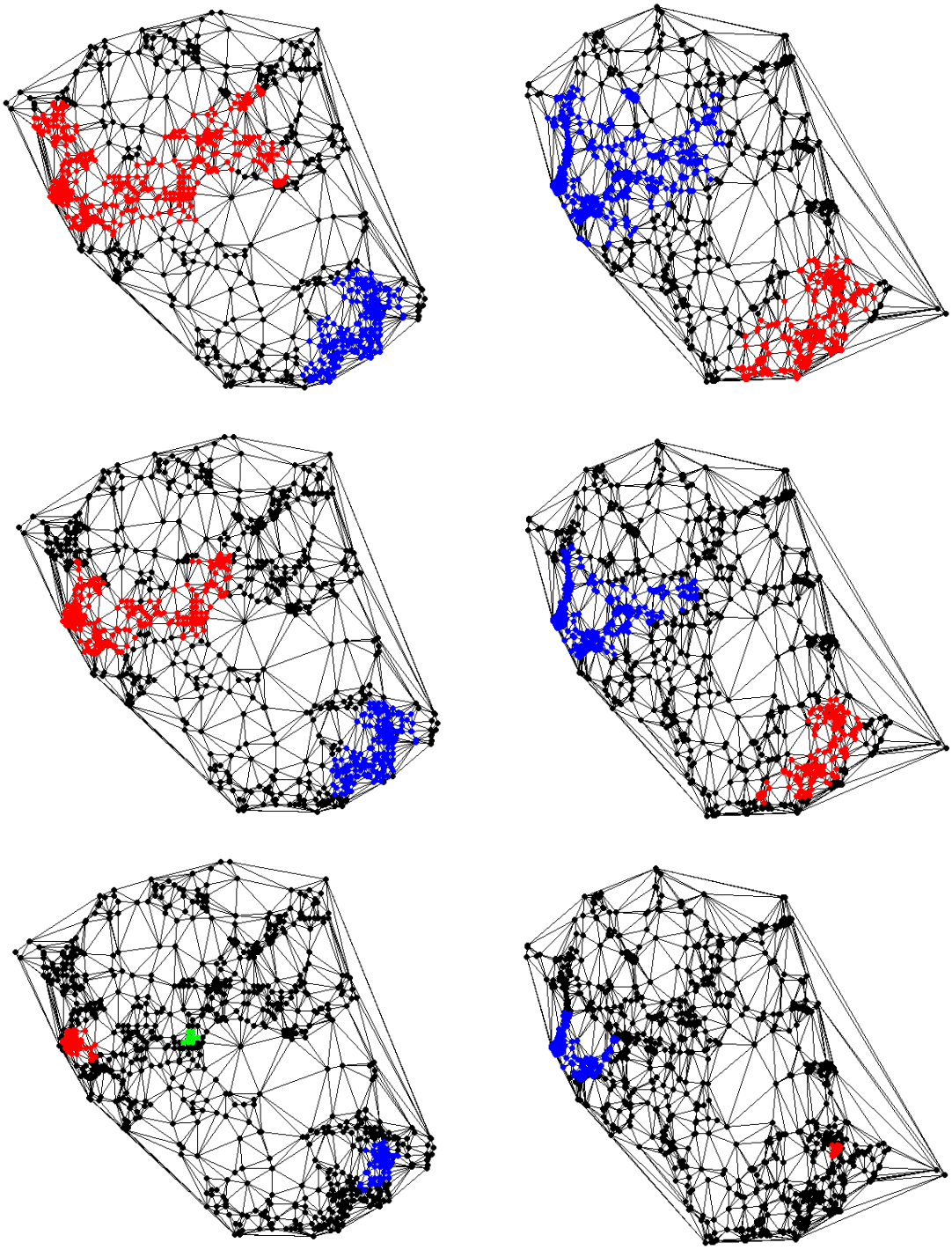

Figura 4.4: Grupos formados usando el método de estimación del número de componentes conexas pdfCluster para los datos de seísmos en Grecia entre 1972 y 1992 (izquierda) y entre 2003 y 2023 (derecha). En la fila superior están los resultados para $\tau = 0.3$, en el medio, $\tau = 0.5$ y en la inferior $\tau = 0.8$.

Vemos como en este caso, excepto para la muestra antigua y $\tau = 0.8$, los grupos generados son dos: uno en la zona noroeste y otra en la sureste. Las principale diferencia con el método anterior es que para $\tau = 0.5$ y la muestra antigua, el grupo más situado al norte estaba dividido en dos y para $\tau = 0.8$ y la muestra actual el grupo que se encuentra en el noroeste estaba dividido en tres.

4.3. Método de la envoltura r-convexa

Ahora pasamos ya a los métodos híbridos, comenzando con el método de la envoltura $r-$convexa. Para cada muestra y cada valor de τ debemos calcular el valor del radio óptimo obtenido mediante el Algoritmo 2. Los resultados para cada muestra y umbral se pueden ver en la Tabla 4.4. Vemos que el valor de r varía según el umbral; en este caso, el mayor valor se alcanza para ambas muestras para $\tau = 0.8$ y el menor para $\tau = 0.5$. Así, con estos valores de los radios se puede estimar el conjunto de nivel como la envoltura $r-$ convexa de la muestra $\mathcal{X}_{n,+}$ y a partir de ella calculamos el número de componentes conexas por las que está formado el conjunto estimado. Los resultados se pueden ver en la Tabla 4.5 y gráficamente en la Figura 4.5.

Radio óptimo

τ	0.8	0.5	0.3
Muestra S.XX	1.645	0.17	0.255
Muestra S.XXI	0.44	0.195	0.375

Tabla 4.4: Radios óptimos obtenidos a partir de las muestras de seísmos en Grecia entre 1972 y 1992 (arriba) y entre 2003 y 2023 (abajo) para el método de la envoltura $r-$convexa con los diferentes valores de τ.

Nº de componentes conexas estimadas

τ	0.8	0.5	0.3
Muestra S.XX	3	9	8
Muestra S.XXI	3	8	8

Tabla 4.5: Número de componentes conexas estimadas para las dos muestras de terremotos y los tres umbrales: $\tau = 0.8, 0.5, 0.3$ mediante el método de la envoltura $r-$convexa con r óptimo.

Como vemos los resultados para los conjuntos de nivel con mayor contenido en probabilidad no se parecen a los obtenidos para los métodos de estimación de componentes conexas que no tenían en cuenta condiciones de forma. En este caso los conjuntos de nivel estimados están formados por un mayor número de componentes conexas y eso se debe a la restricción sobre el conjunto de nivel de que

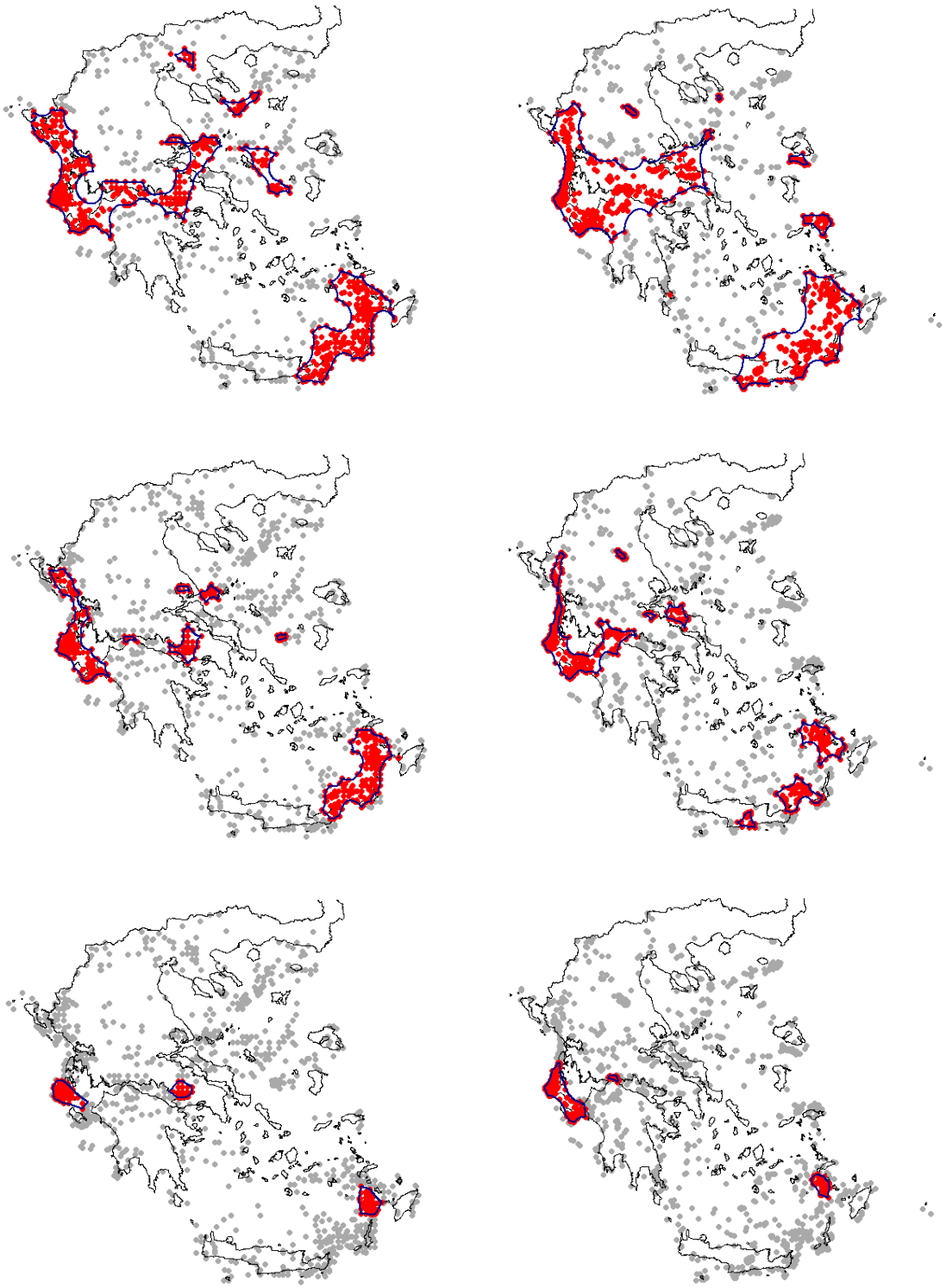

Figura 4.5: Grupos formados usando el método de estimación de conjuntos de nivel y componentes conexas de la envoltura $r-$convexa con r óptimo para los datos de seísmos en Grecia entre 1972 y 1992 (izquierda) y entre 2003 y 2023 (derecha). En la fila superior están los resultados para $\tau = 0.3$, en el medio, $\tau = 0.5$ y en la inferior $\tau = 0.8$.En rojo los puntos de $\mathcal{X}_{n,+}$ y en azul los conjuntos de nivel.

que contenga todas las observaciones de la submuestra $\mathcal{X}_{n,+}$ pero que ningún punto de $\mathcal{X}_{n,-}$ se encuentre en su interior. Así, para los valores $\tau = 0.8$ tenemos para ambas muestras tres componentes conexas; la situada en el sureste y en el oeste coinciden pero la restante se ubica en lugares distintos. Para $\tau = 0.5$ tenemos 9 componentes conexas para la muestra de terremotos del siglo pasado y 8 para las del actual. En el caso de la primera, tal como vemos en la parte central izquierda de la Figura 4.5, solo dos de las componentes conexas se encuentran en la parte sur y una de ellas es un punto aislado; sin embargo en la muestra del siglo XXI la parte del sureste está dividida en tres zonas en vez de estar todos los seísmos concentrados en la misma. Esto deja de ocurrir para $\tau = 0.3$ que las diferencias se encuentran más en la zona norte aunque el número de componentes conexas estimadas coinciden. Se puede observar en la parte superior de la Figura 4.5 como las zonas de alta densidad de seísmos se han desplazado hacia el este en la zona norte en la muestra del siglo actual frente al pasado.

4.4. Método del suavizado granulométrico

El último método también forma parte de las metodologías que asumen condiciones de forma y en este caso asumimos que el conjunto de nivel está formado por bolas de radio fijo. Para cada muestra y cada valor de τ tenemos un radio distinto que se muestran en la Tabla 4.6. Al igual que ocurría antes, los valores del radio no siguen una distribución monótona sino que, como se ve en la Tabla 4.6, pueden tomar un valor más alto para $\tau = 0.8$, diminuir para $\tau = 0.5$ y volver a aumentar cuando $\tau = 0.3$. Los resultados del número de componentes conexas se pueden ver en la Tabla 4.5 y gráficamente en 4.6.

Radio obtenido a partir de los datos

τ	0.8	0.5	0.3
Muestra S.XX	0.20	0.12	0.20
Muestra S.XXI	0.08	0.06	0.095

Tabla 4.6: Radios óptimos obtenidos a partir de las muestras de seísmos en Grecia entre 1972 y 1992 (arriba) y entre 2003 y 2023 (abajo) para el método del suavizado granulométrico con los diferentes valores de τ.

Nº de componentes conexas estimadas

τ	0.8	0.5	0.3
Muestra S.XX	3	7	5
Muestra S.XXI	3	8	11

Tabla 4.7: Número de componentes conexas estimadas para las dos muestras de terremotos y los tres umbrales: $\tau = 0.8, 0.5, 0.3$ mediante el método del suavizado granulométrico con r óptimo.

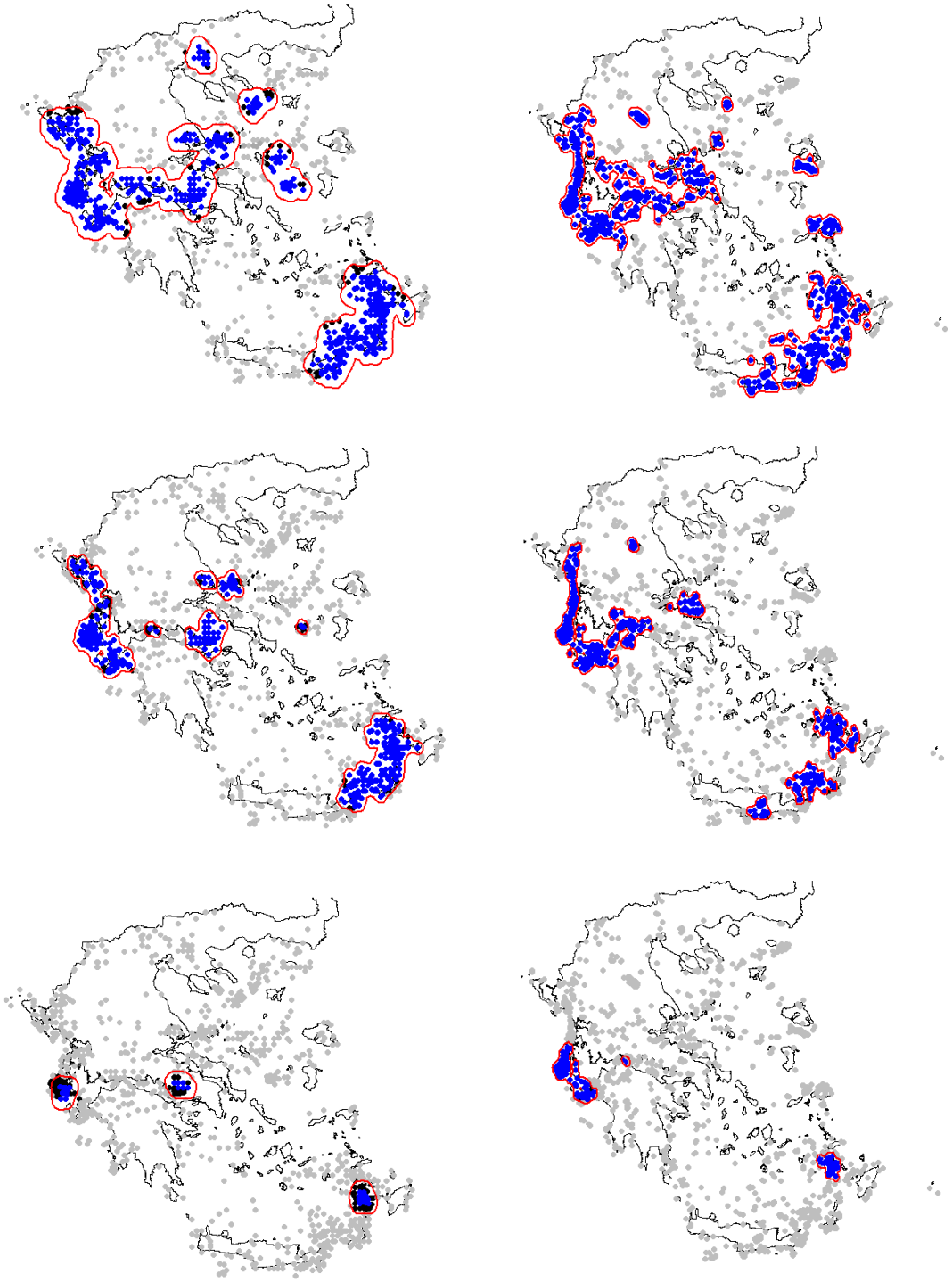

Figura 4.6: Grupos formados usando el método de estimación de conjuntos de nivel y componentes conexas de la envoltura $r-$convexa con r obtenido a partir de la muestra para los datos de seísmos en Grecia entre 1972 y 1992 (izquierda) y entre 2003 y 2023 (derecha). En la fila superior están los resultados para $\tau = 0.3$, en el medio, $\tau = 0.5$ y en la inferior $\tau = 0.8$. En azul los centros de las bolas, en negro las observaciones de $\mathcal{X}_{n,+}$, en gris las de $\mathcal{X}_{n,-}$y en rojo los conjuntos de nivel.

Este método, al igual que el anterior, estima un número de componentes conexas alto y, de nuevo, esto puede ser consecuencia de la cercanía de los puntos de la muestra $\mathcal{X}_{n,-}$ a las observaciones de $\mathcal{X}_{n,+}$. Para $\tau = 0.8$ se estiman tres componentes conexas para cada muestra y coinciden en espacio con las estimadas mediante el método de la envoltura $r - convexa$. En el caso de $\tau = 0.5$ tenemos siete componentes conexas en el caso de los datos de seísmos del siglo XX y ocho en las del siglo actual. Para la primera muestra tenemos dos componentes menos que para el método de la envoltura $r-$convexa y esta diferencia viene dada por dos puntos aislado que siguiendo la metodología de esta sección forman parte de una componente conexa de mayor tamaño cercana. Para la muestra de datos del siglo XXI se estiman el mismo número de componentes conexas y cada una está formada por el mismo conjunto de observaciones que usando la envoltura $r-$convexa. Por último, para $\tau = 0.3$ hay una gran diferencia en las componentes estimadas: para la muestra del siglo XX se tienen cinco y para el del actual once. Para la primera muestra tal como vemos en la parte superior izquierda de la Figura 4.6 tenemos una componente conexa en el sureste (al igual que en todos los métodos anteriores) y una de las diferencias se encuentra en la zona norte: usando la envoltura $r-$convexa teníamos tres puntos aislados que empleando esta metodología forman parte de una componente conexa cercana de mayor tamaño, igual que para $\tau = 0.5$. La mayor diferencia viene en la muestra actual. Probablemente debido a la cercanía de las observaciones de la muestra $\mathcal{X}_{n,-}$, el radio óptimo es pequeño y eso lleva a una estimación de un mayor número de componentes conexas.

4.5. Conclusiones

Tras aplicar los diferentes métodos a dos bases de datos reales y de gran tamaño, podemos concluir varias cosas.

Primero presentamos las conclusiones acerca de las metodologías. Vemos como los métodos híbridos para valores de $\tau = 0.5, 0.3$ estiman un mayor número de componentes conexas que los contadores que no tienen en cuenta condiciones geométricas. Sin embargo para $\tau = 0.8$ los tres métodos tienen unos resultados prácticamente iguales, de tres componentes para la muestra de seísmos del siglo XX para todos los métodos y entre dos y cuatro para la otra muestra. Por tanto hemos visto la importancia de la distancia entre los puntos de $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$ en los métodos híbridos ya que de esta depende el radio óptimo estimado en ambos. Podríamos pensar que debido a la pequeña distancia entre las dos submuestras mencionadas se está sobreestimando el número de componentes conexas con los métodos híbridos pero esto no se detectaba en el estudio de simulación. Otra opción sería que los dos primeros métodos estuvieran estimando un número de componentes por debajo del real y no son capaces de diferenciar todos los grupos correctamente. Podría ser interesante para futuros trabajos ver si la distancia entre las submuestras influye a la hora de sobreestimar el número de componentes conexas.

Por último, tenemos la conclusión sobre la evolución de las zonas de alta densidad de seísmos. Fijándonos en los lugares más afectados ($\tau = 0.8$), tenemos que tanto las islas del oeste del país como las del sureste eran zonas muy afectadas según la muestra del siglo XX y lo siguen siendo actualmente. La diferencia está en el centro del país, al oeste de Atenas. En los registros del período entre 1972 y 1992 se tenía esa zona como un lugar con alta densidad de terremotos y sin embargo en la muestra de seísmos entre 2003 y 2023 esa zona de alta densidad se ha desplazado hacia el oeste y ha disminuido considerablemente su tamaño. Si tomamos un umbral más bajo como $\tau = 0.5$ tenemos varias zonas de alta densidad; de nuevo las islas del oeste y del sureste son las zonas principales pero se une a estas la zona central del país. Además, según la muestra de datos del siglo XX tenemos alguna isla pequeña del este también se encuentra en peligro pero esto no ocurre en la actualidad sino que es la

costa al este de Atenas la zona afectada. Por último, fijándonos en lo obtenido para $\tau = 0.3$, notamos algunas pequeñas diferencias. Por un lado que las islas situadas al este del país están más afectadas por los terremotos en la actualidad de lo que lo estaban según la muestra del siglo XX, por lo que deberían tomarse nuevas medidas de seguridad en ellas. Por otro lado, parece que ahora los seísmos se concentran además en la franja central, en especial las islas del oeste, y en el sureste mientras que para la muestra obtenida en el período entre 1972 y 1992 se detectaba una zona con gran número de seísmos en el norte, cercano a la zona centro de Macedonia. En la muestra del siglo XXI esto no ocurre, siendo una región sin excesivo peligro.

En general no se detectan grandes variaciones en los seísmos en Grecia entre las dos muestras de datos salvo las comentadas previamente y donde destacan la menor afectación del oeste de la capital del país ($\tau = 0.8$) y el aumento de los seísmos registrados sobre las islas del este más cercanas a Turquía ($\tau = 0.3$).

Apéndice A

Funciones Kernel unidimensionales y multidimensionales

1. Kernels unidimensionales

Kernels unidimensionales	
Kernel	**K(u)**
Uniforme	$\frac{1}{2}\mathbb{1}(\lvert u\rvert \leq 1)$
Triangular	$(1-\lvert u\rvert)\mathbb{1}(\lvert u\rvert \leq 1)$
Epanechnikov	$\frac{3}{4}(1-u^2)\mathbb{1}(\lvert u\rvert \leq 1)$
Biweight	$\frac{15}{16}(1-u^2)^2\mathbb{1}(\lvert u\rvert \leq 1)$
Gaussiano	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$
Coseno	$\frac{\pi}{4}\cos(\frac{\pi}{2})\mathbb{1}(\lvert u\rvert \leq 1)$

Tabla A.1: Funciones kernel unidimensionales más empleadas para la estimación tipo núcleo de la densidad.

2. Kernels multidimensionales

Kernel normal $d-$dimensional: $K(\mathbf{u}) = (2\pi)^{-d/2} exp(-\frac{1}{2}\mathbf{u^T u})$

Kernel multiplicativo: $K(\mathbf{u}) = \prod_{i=1}^{d} \kappa(u_i)$, donde $\kappa(\cdot)$ es un kernel unidimensional de los que se han definido previamente.

Bibliografía

Azzalini, A., Menardi, G. (2014). Clustering via nonparametric density estimation: the R package pdfCluster. Journal of Statistical Software, 57(11), 1-26, URL: http://www.jstatsoft.org/v57/i11/.

Azzalini A., Torelli N. (2007). Clustering via nonparametric density estimation. Statistics and Computing. 17, 71-80.

Baíllo, A., Cuevas, A. (2006). Parametric versus nonparametric tolerance regions in detection problems. Computational Statistics, 21, 523-536

Becker, RA., Minka, TP., Deckmyn, A. (2023). _maps: Draw Geographical Maps_. R package version 3.4.2, https://CRAN.R-project.org/package=maps. Accedido 15 de octubre de 2024.

Biau, G., Cadre, B., Pelletier, B. (2007). A graph-based estimator of the number of clusters. ESAIM: Probability and Statistics, 11, 272-280.

Bird, P. (2003). An updated digital model of plate boundaries. Geochemistry, Geophysics, Geosystems 4(3).

Bolón, D., Crujeiras, R. M., Rodríguez-Casal, A. (2024). Granulometric Smoothing on Manifolds. arXiv preprint arXiv:2407.07559.

Chacón, J. E., Duong, T. (2018). Multivariate kernel smoothing and its applications. Chapman and Hall/CRC.

Cuevas, A., Febrero, M.,Fraiman, R. (2000). Estimating the number of clus ters. Canadian Journal of Statistics, 28, 367-382.

Devroye, L., Wise, G. L. (1980). Detection of abnormal behavior via non parametric estimation of the support. SIAM Journal on Applied Mathematics, 38, 480-488.

Duong, T. (2024). _ks: Kernel Smoothing_. R package version 1.14.2, https://CRAN.R-project.org/package=ks. Accedido 15 de octubre de 2024.

Duong, T., Hazelton, M.L. (2003). Plug-in bandwith matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics, 15, 17-30

Gebhardt, A., Bivand, R., Sinclair, D. (2024). interp: Interpolation Methods. R package version 1.1-6, <https://CRAN.R-project.org/package=interp>.

Gebhardt, A., Renka, R., Eglen, S., Zuyev, S., White, D. (2024). tripack: Triangulation of Irregularly Spaced Data. R package version 1.3-9.2, <https://CRAN.R-project.org/package=tripack>.

Grübel, R. (1988). The length of the short. Annals of Statistics, 16, 619-628.

Hall, P., Marron, J. S. (1987). On the amount of noise inherent in bandwidth selection for a kernel density estimator. The Annals of Statistics, 163-181.

Hartigan, J.A. (1975). Clustering Algorithms. John Wiley & Sons, New York.

Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. Journal of the American Statistical Association, 82, 267-270.

Hyndman, R.J. (1996). Computing and graphing highest density regions. The American Statistician, 50(2), 120-126.

Kanamori, H., Brodsky, E.E. (2004). The physics of earthquakes. Reports on progress in physics, 67(8), 1429-1496.

Menardi, G., Azzalini, A. (2014). An advancement in clustering via nonparametric density estimation. Statistics and Computing. DOI: 10.1007/s11222-013-9400-x.

Pateiro-López, B., Rodríguez-Casal, A. (2022). alphahull: Generalization of the Convex Hull of a Sample of Points in the Plane. R package version 2.5, <https://CRAN.R-project.org/package=alphahull>.

Polonik, W. (1995). Measuring mass concentration and estimating density contour clusters-an excess mass approach. Annals of Statistics, 23, 855-881.

Qiao, W. (2020). Asymptotics and optimal bandwidth for nonparametric estimation of density level sets.

Rényi, A., Sulanke, R. (1963). Über die konvexe Hülle von n zufällig gewählten Punkten. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 2, 75-84.

Rényi, A., Sulanke, R. (1964). Über die konvexe Hülle von n zufällig gewählten Punkten. Z. Wahrscheinlichkeitstheorie und Verw. Gebiete, 3, 138-147.

Rodríguez-Casal, A., Saavedra-Nieves, P. (2022). A data-adaptive method for estimating density level sets under shape conditions. The Annals of Statistics, 50(3), 1653-1668.

Saavedra-Nieves, P. (2014). Nonparametric data-driven methods for set estimation. Tesis, Universidade de Santiago de Compostela.

Saavedra-Nieves, P., Crujeiras, R.M. (2022). HDiR: An R Package for Computation and Nonparametric Plug-in Estimation of Directional Highest Density Regions and General Level Sets.

Samworth, R.J., Wand, M.P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. The Annals of Statistics, 38, 1767-1792.

Silverman, B.W. (1986). Density estimation for statistics and data analysis. Monographs on Statistics and Applied Probability.

Walther, G. (1997). Granulometric Smoothing. The Annals of Statistics, 25, 2273-2299.

Wand, M. P., Jones, M. C. (1993). Comparison of smoothing parameterizations in bivariate kernel density estimation. Journal of the American Statistical Association, 88(422), 520-528.

Wanli, Q. (2020). Asymptotics and optimal bandwidth for nonparametric estimation of density level sets.