Trabajo Fin de Máster

Estimación no paramétrica de la extensión de ocurrencia

Raquel García Jaen

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación non paramétrica da extensión de ocurrencia

Título en español: Estimación no paramétrica de la extensión de ocurrencia

English title: Nonparametric estimation of the extent of occurrence

Modalidad: Modalidad A

Autor/a: Raquel García Jaen, Universidad de Santiago de Compostela

Director/a: Alberto Rodríguez Casal, Universidad de Santiago de Compostela

Tutor/a: Paula Saavedra Nieves, Universidad de Santiago de Compostela

Breve resumen del trabajo:
La Unión Internacional de Conservación de la Naturaleza define la extensión de ocurrencia (*extent of occurrence*, EOO) como el área contenida dentro del límite imaginario continuo más corto que puede dibujarse para abarcar todos los lugares conocidos, inferidos o proyectados de la ocurrencia actual de un taxón, excluyendo los casos de vagancia.

En general, los biólogos y ecólogos estiman la EOO de un taxón como la envoltura convexa de las coordenadas geográficas correspondientes a sus posiciones de ocurrencia. Sin embargo, la convexidad puede ser una condición de forma restrictiva en la práctica. En este trabajo, se revisará un estimador no paramétrico de la EOO más flexible, la envoltura r–convexa. Como aplicación práctica, su comportamiento será ilustrado analizando la reconstrucción de la EOO para la avispa velutina en Galicia.

Recomendaciones: Cursar previamente Estadística No Paramétrica

Otras observaciones: El trabajo forma parte de una línea de investigación conjunta entre los dos directores.

Agradecimientos

Me gustaría agradecer a los profesores Alberto Rodríguez Casal y Paula Saavedra Nieves su labor como directores de este Trabajo Fin de Máster. Muchas gracias por el apoyo recibido y las horas de dedicación empleadas en este trabajo.

Quiero dar también las gracias a mi familia por el apoyo y ánimos recibidos. En especial a mis padres, Miguel y Ana, que desde Canarias siempre han sabido transmitirme los valores que me hacen estar hoy aquí.

Por último, gracias también a todos mis compañeros de máster por los dos años compartidos, a mis amigos por su apoyo incondicional y a mi pareja Pablo por no dejar de creer en mí.

Índice general

Resumen

Resumen en español

La Vespa asiática, también conocida como Vespa velutina es una especie de avispa originaria del sudeste asiático. En solo siete años, la velutina ha colonizado toda Galicia y ha pasado de ser un insecto desconocido a generar alarma entre la población por su presencia cada vez más intensa. En este trabajo, abordaremos la reconstrucción de la extensión de ocurrencia (EOO) de la velutina como un problema de estimación del soporte. Dada una muestra aleatoria simple generada por una distribución desconocida, revisaremos los métodos propuestos en la literatura para la estimación del soporte. En particular, bajo la hipótesis de que el soporte S es r–convexo, el estimador natural de S resulta ser la envoltura r–convexa de la muestra de puntos. Dicho estimador presenta el inconveniente de que el parámetro de forma r es desconocido en la práctica. Para solventar este problema, emplearemos el método implementado por Rodríguez-Casal y Saavedra-Nieves (2019) para estimar dicho parámetro r y poder aplicar dicho estimador sobre nuestro conjunto de datos.

English abstract

The Asian Vespa, also known as Vespa velutina is a species of native wasp of the Asian Southeast. In just seven years, velutina has colonized all Galicia and has gone from being an unknown insect to generate alarm among the population due to its increasingly intense presence. In this work, we will approach the reconstruction of the extent of occurrence (EOO) of velutina as a support estimation problem. Given a random sample from some unknown distribution, we will review the methods proposed in literature for estimating support. In particular, under the hypothesis that the support S is r–convex, the natural estimator of S turns out to be the r–convex hull of the sample of points. This estimator has the drawback that the shape parameter r is usually unknown in practice. To solve this problem, we will use the method proposed in Rodríguez-Casal y Saavedra-Nieves (2019) to estimate the parameter r and be able to apply this estimator on our dataset.

Capítulo 1

Introducción

La variedad *Vespa velutina nigrithorax*, popularmente conocida como avispa asiática, es una especie de avispa originaria del sudeste asiático. Esta avispa es considerada una especie invasora en Europa desde su llegada al suroeste francés en 2004. Se cree que podría haber sido importada accidentalmente desde China a través del comercio hortícola, ver Monceau *et al.* (2014). Desde entonces, ha invadido unos $120000\,km^2$ y atacado colmenas, causando importantes daños y alarma social en los lugares afectados. Las abejas melíferas europeas no cuentan con una estrategia de defensa eficaz, y por ello un grupo de estas avispas es capaz de acabar con gran parte de una colmena en poco tiempo y mermar su productividad.

En 2010, se confirmó en Navarra su llegada a la península ibérica a través de los Pirineos y desde entonces, además de en el País Vasco y Navarra, la avispa asiática se ha asentado en Cataluña, Cantabria, Galicia, Asturias, La Rioja, la provincia de Burgos en Castilla y León e incluso ha llegado a las islas Baleares, donde parece haber sido controlada. Debido a su potencial colonizador y a que constituye una grave amenaza para las especies autóctonas, los hábitats o los ecosistemas, esta especie ha sido incluida en el Catálogo Español de Especies Exóticas Invasoras, aprobado por Real Decreto 630/2013, del 2 de agosto. Para España, se han definido un conjunto de estrategias de gestión, control y posible erradicación.

En Galicia se detectó su existencia al encontrarse ejemplares y nidos aislados en 2012, en la zona de Burela del norte de Lugo, y de Baiona en el sur de la provincia de Pontevedra. Esto hace pensar que la entrada en Galicia tuvo lugar simultáneamente por dos puntos distintos, probablemente a través del puerto de Burela por el norte, y procedente de territorios colonizados en Portugal por la zona sur. En sólo siete años, la velutina ha colonizado toda Galicia y ha pasado de ser un insecto desconocido a generar alarma entre la población por su presencia cada vez más intensa y por el miedo a su picadura que, aunque los expertos de la Sociedad Española de Alergología e Inmunología Clínica (SEAIC) aseguran que no es más peligrosa que la de otras especies de avispa, esta ya ha producido al menos tres muertes en la comunidad según recoge el periódico La Voz de Galicia. Una de las vías de investigación sobre la especie se centra en la localización de los nidos, tanto para su destrucción como para evitar accidentes entre los profesionales que manipulan árboles.

En este trabajo, abordaremos la reconstrucción de la extensión de ocurrencia (*extent of occurrence*, EOO) de la velutina. Actualmente, el concepto de extensión de ocurrencia es uno de lo más utilizados en el estudio de la presencia de especies para los diseños de redes de reservas naturales. De hecho, la Unión Internacional para la Conservación de la Naturaleza (UICN) establece la EOO como una medida clave del riesgo de extinción. En términos generales, la UICN define la EOO como el área contenida dentro de los bordes imaginarios continuos más cortos que pueden dibujarse para incluir todos los sitios conocidos, inferidos o proyectados en los que un taxón se halle presente, ver Bland *et*

al. (2016). Con el objetivo de estimar la EOO de la velutina en Galicia, emplearemos un conjunto de datos reales proporcionado por la Consellería de Medio Rural, donde se recogen las localizaciones de nidos de velutinas registrados en Galicia entre enero de 2014 y diciembre de 2019. En concreto, dicho conjunto se compone de 87635 coordenadas en UTM. El sistema de coordenadas geográficas UTM (Universal Transverse Mercator) es un sistema de coordenadas basado en la proyección cartográfica transversa de Mercator. A diferencia del sistema de coordenadas geográficas, expresadas en longitud y latitud, las magnitudes en el sistema UTM se expresan en metros al nivel del mar. A la hora de emplear este conjunto de datos a lo largo del trabajo, no hemos necesitado realizar ninguna transformación sobre ellos, sino que hemos utilizado directamente las coordenadas UTM en las que se recogieron. En la Figura 1.1 se muestra la localización geográfica de este conjunto de datos.

Figura 1.1: Localización geográfica de Galicia.

La evolución por años de la velutina en Galicia recogida en nuestros datos se muestra en la Figura 1.2. En la primera de las representaciones, correspondiente al año 2014, podemos distinguir los dos focos de aparición de velutinas comentados anteriormente, uno en el norte de la provincia de Lugo y el otro en el sur de la provincia de Pontevedra, limítrofe con Portugal.

Matemáticamente, la reconstrucción de la EOO puede verse como un problema de estimación del soporte de una distribución. Dicho problema pertenece a la rama de la estadística dedicada a la estimación de conjuntos. Esta trata de reconstruir, a partir de una muestra aleatoria de puntos, un conjunto desconocido del espacio euclidiano $\mathbb{R}^d$ que guarda una estrecha relación con la distribución de los datos observados. Los estimadores resultantes suelen depender de ciertos parámetros de suavizado y los resultados teóricos están enfocados principalmente en obtener propiedades asintóticas, en especial, consistencia y tasas de convergencia. El objetivo es conseguir garantizar resultados teóricos lo más generales posibles bajo condiciones no muy restrictivas sobre la forma del conjunto o la distribución de los datos. Además de la estimación del soporte, se ha abordado la estimación de conjuntos de nivel, para representar gráficamente las zonas de alta concentración de datos, o la estimación de la frontera, para visualizar bien las propiedades geométricas de la región a estimar.

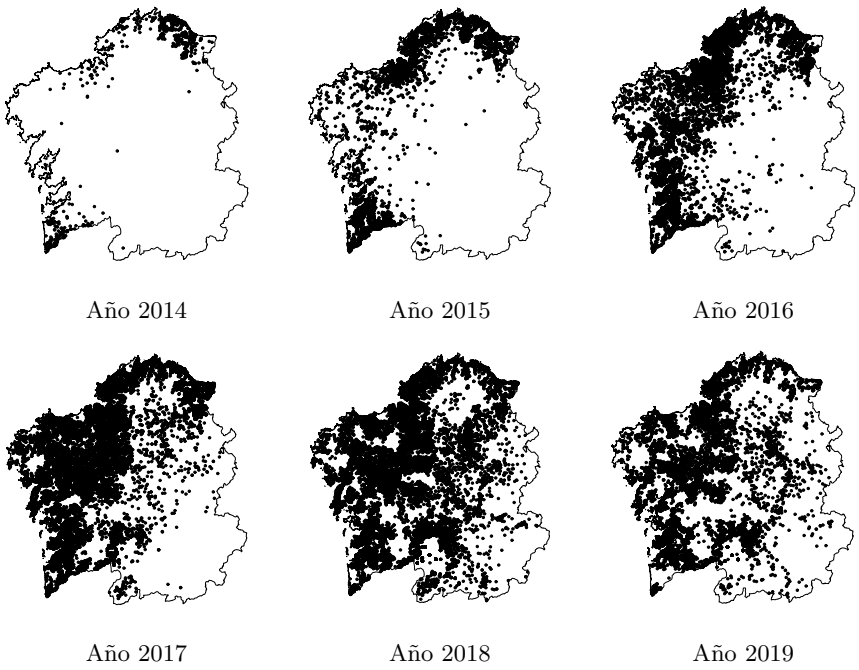


Figura 1.2: Coordenadas UTM de los nidos de avispa velutina en Galicia desde 2014 hasta 2019.

El soporte de una variable absolutamente continua X con distribución de probabilidad $\mathbb{P}_X$ se define como el conjunto cerrado más pequeño con $\mathbb{P}_X$–medida uno. Dada una muestra independiente e idénticamente distribuida, $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, de una variable absolutamente continua X con distribución de probabilidad $\mathbb{P}_X$, el objetivo es estimar el soporte, $S \subset \mathbb{R}^d$, de dicha variable que supondremos compacto y no vacío. Este problema puede abordarse desde dos enfoques distintos dependiendo de si se asumen o no condiciones de forma sobre S, tal como veremos en detalle en el siguiente capítulo.

El contenido del presente Trabajo Fin de Máster se divide en cinco capítulos estructurados de la siguiente forma. En el Capítulo 1, hemos introducido la motivación del trabajo, el problema de la expansión de la vespa velutina en Galicia y la EOO como concepto útil para analizar dicha situación. A continuación, hemos presentado el conjunto de datos con el que vamos a trabajar y la forma de modelar matemáticamente el problema de estimación del soporte.

El Capítulo 2, lo dedicamos a la revisión de los distintos estimadores disponibles en la literatura para este problema. Comenzaremos con el caso más general, en el cual no se asumen hipótesis sobre la forma de S para luego, continuar con los estimadores resultantes de suponer la convexidad y r-convexidad de S. A la vez que vamos presentando los estimadores, iremos introduciendo las herramientas matemáticas necesarias para la reconstrucción del soporte, comenzando con la distancia entre conjuntos y continuando con las condiciones de forma que se pueden establecer sobre dicho conjunto. Para ilustrar cada uno de los estimadores, haremos uso de los datos para su representación y discutiremos las limitaciones de cada uno de ellos.

A lo largo del Capítulo 3, presentaremos el método propuesto por Rodríguez-Casal y Saavedra-Nieves (2019) para estimar el soporte de la velutina bajo la hipótesis de r–convexidad de S. El estimador resultante de aplicar dicho método depende de la estimación del spacing maximal, cuya distribución asintótica es conocida bajo determinadas condiciones sobre la distribución en el muestreo. A continuación, describiremos con detalle el algoritmo implementado en R para su cálculo. Para

terminar, reconstruiremos la EOO de la velutina mediante el algoritmo descrito en distintos periodos temporales. Los estimadores obtenidos, en este caso particular, son extremadamente fragmentados. La diagnosis de las posibles causas se realizará a través de la realización de un estudio de simulación en el Capítulo 4. Dicho estudio nos permitirá identificar cuál es la explicación del funcionamiento anómalo del método presentado en el Capítulo 3. En concreto, emplearemos varios modelos de mixturas de normales bidimensionales para comprobar el calibrado del spacing maximal del que depende el estimador. Para finalizar, realizaremos un análisis del conjunto de datos en el Capítulo 5, estimaremos el soporte centrándonos en distintos periodos temporales y regiones geográficas. También realizaremos distintos niveles de limpieza de los datos a través de la muestra efectiva. Esto nos permitirá concluir si el muestreo realizado es homogéneo o no. A partir del estudio de simulación y del análisis de los datos extraeremos conclusiones, las cuales se expondrán en el Capítulo 6, sobre qué puede estar causando que las estimaciones obtenidas no sean las esperadas.

Capítulo 2

Estimación del soporte

La estimación del soporte es uno de los problemas que aborda la estimación de conjuntos. Dada una muestra $\mathcal{X}_n = \{X_1, \ldots, X_n\}$, de una variable absolutamente continua X con distribución de probabilidad $\mathbb{P}_X$, el objetivo es estimar el soporte de dicha variable, $S \subset \mathbb{R}^d$, que supondremos compacto y no vacío. En este capítulo, presentaremos los estimadores clásicos del soporte disponibles en la literatura. Ilustraremos cada uno de los estimadores a partir de nuestro conjunto de datos presentado en el Capítulo 1 y analizaremos sus ventajas e inconvenientes.

En primer lugar, en la Sección 2.1, presentaremos el estimador para el caso general donde no se realizan restricciones de forma sobre el conjunto S. A continuación, introduciremos las estimaciones resultantes de suponer restricciones de forma sobre S. En la Sección 2.2, revisamos las reconstrucciones resultantes de suponer la convexidad del soporte y en la Sección 2.3 cuando el soporte es r-convexo. Conforme vayamos presentando dichos estimadores, introduciremos las herramientas matemáticas necesarias para la reconstrucción del soporte comenzando con la distancia entre conjuntos y continuando con las condiciones de forma que se pueden asumir en cada uno de los casos sobre S.

2.1. Caso general en la estimación del soporte

En el caso más general no se realiza ninguna hipótesis sobre la forma de S. Por tanto, la única información que disponemos para estimar el soporte S es aquella que nos proporciona la muestra $\mathcal{X}_n$.

Antes de comenzar a presentar los estimadores existentes en este caso, necesitamos definir como evaluar su calidad como estimador del conjunto. Para ello, es necesario medir la distancia entre dicho estimador y el conjunto teórico. La distancia entre puntos de $\mathbb{R}^d$ más común es la distancia euclidiana, pero la distancia entre conjuntos es un concepto distinto. Por ejemplo, podemos pensar que la distancia entre los conjuntos A y C de la Figura 2.1 es cero al poseer un borde común, pero dichos conjuntos son un poco distintos.

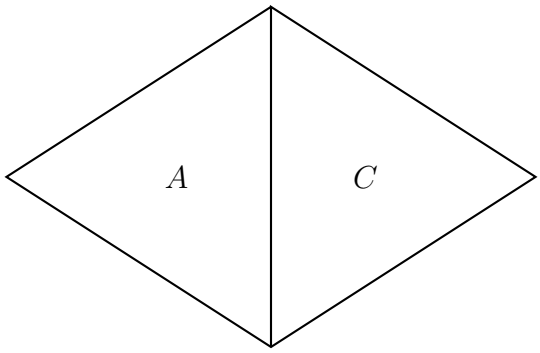


Figura 2.1: A y C poseen un borde común.

Por tanto, para dar una definición adecuada de distancia entre los conjuntos A y C todo parece indicar que debemos de tener en cuenta la distancia entre los puntos de A y el borde de C y viceversa. Una distancia que nos permite abordar este planteamiento es la distancia Hausdorff.

Definición 2.1. Sean A, $C \subset \mathbb{R}^d$ conjuntos compactos no vacíos y $\| \, \|$ la norma euclidiana. Se define la distancia Hausdorff entre A y C como

$$d_H(A, C) = \text{máx}\left\{\sup_{a \in A} d(a, C), \sup_{c \in C} d(c, A)\right\}$$

donde

$$d(a, C) = \text{ínf}\{\|a - c\| \, : \, c \in C\}.$$

La distancia Hausdorff nos da una idea de la proximidad entre dos conjuntos, por tanto, nos puede resultar útil a la hora de evaluar la calidad de un estimador, tal como ocurre en la teoría paramétrica clásica. Dado un estimador S_n de S buscaríamos que $d_H(S_n, S) \to 0$. Sin embargo, esto no resulta suficiente para asegurar que S_n estima satisfactoriamente a S, ya que S_n puede ser d_H–consistente sin que ambos conjuntos sean necesariamente similares. Esto ocurre, por ejemplo, en el caso de la propia muestra $\mathcal{X}_n$ como estimador del soporte S pues se cumple que $d_H(S, \mathcal{X}_n) \to 0$ cuando $n \to \infty$ con probabilidad uno. Por tanto, necesitamos otra distancia que nos permita medir la similitud en contenido, y no sólo la proximidad física, entre dos conjuntos. Dicha distancia es la distancia en medida.

Definición 2.2. Sean A y C dos conjuntos de Borel acotados. La distancia en medida entre A y C se define como

$$d_\mu(A, C) = \mu(A \triangle C)$$

donde μ denota la distancia de Lebesgue y $\triangle$, la diferencia simétrica, esto es,

$$A \triangle C = A \cup C \setminus A \cap C.$$

En la Figura 2.2, se ilustra el concepto de diferencia simétrica entre dos conjuntos A y C. Con esta medida, dado que $\mathcal{X}_n$ es un conjunto finito, se cumple que $d_\mu(S, \mathcal{X}_n) = \mu(S \triangle \mathcal{X}_n) = \mu(S \setminus \mathcal{X}_n) = \mu(S) > 0$. Dado que $d_\mu(S, \mathcal{X}_n) = \mu(S) > 0$, la distancia en medida, a diferencia de la de Hausdorff, nos está indicando que la propia muestra no es un estimador adecuado para S. La Figura 2.3 ilustra la diferencia entre la distancia euclidiana y la distancia Hausdorff.

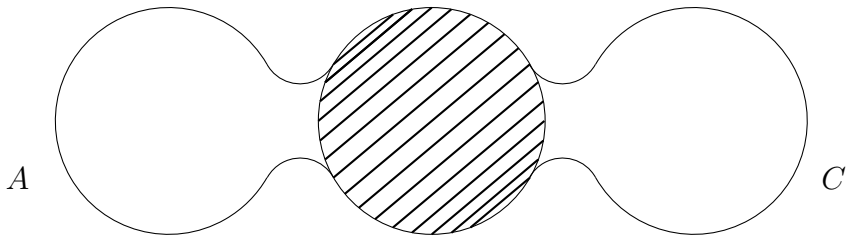


Figura 2.2: Diferencia simétrica entre A y C en $\mathbb{R}^2$.

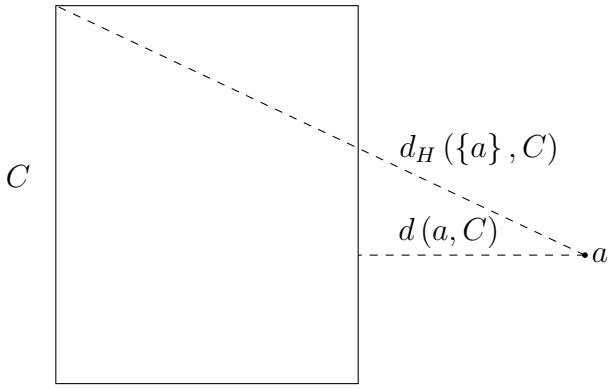


Figura 2.3: Distancia de Hausdorff entre $\{a\}$ y C.

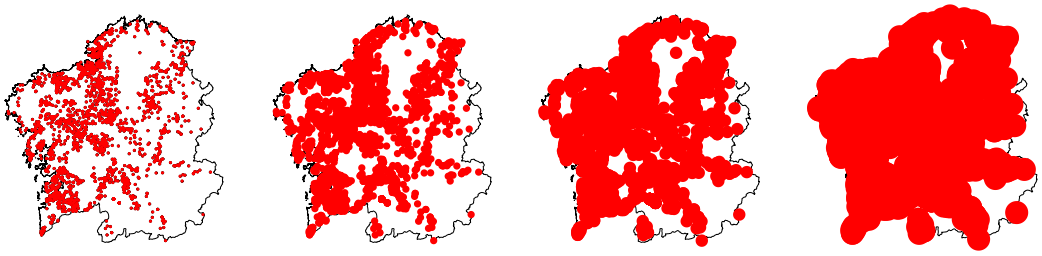
Volviendo a la estimación del soporte, cabe recordar que en el caso general no asumimos ninguna condición de forma sobre S. En consecuencia, la única información que disponemos es la que nos proporciona la muestra de puntos $\mathcal{X}_n$. Sin embargo, esta en sí no es un buen estimador de S como acabamos de ver pues, aunque es d_H–consistente, se tiene que $d_\mu(S, \mathcal{X}_n) > 0$. Por tanto, necesitamos un estimador más flexible que aproxime mejor S. Chevalier (1976) y Devroye and Wise (1980) propusieron como estimador de S la siguiente versión suavizada de la muestra $\mathcal{X}_n$:

$$S_n = \bigcup_{i=1}^{n} B_{\epsilon_n}[X_i], \tag{2.1}$$

donde $B_{\epsilon_n}[X_i]$ denota la bola cerrada de centro X_i y radio $\epsilon_n > 0$.

En la Figura 2.4 se muestra la influencia del parámetro ϵ_n en el estimador para los datos de la velutina de noviembre de 2016. Si dicho parámetro se escoge próximo a cero esto se traducirá en un estimador con muchas componentes conexas, ver Figura 2.4 (izquierda). Sin embargo, para valores muy grandes de ϵ_n se tendrá que $S \subset S_n$ y, por tanto, se sobreestima el soporte considerablemente.

Devroye and Wise (1980) probaron que, cuando $\mathcal{X}_n$ sigue una distribución absolutamente continua en S, el estimador (2.1) es d_μ–consistente en probabilidad y de forma casi segura. Si $\epsilon_n \to 0$ y $n\epsilon_n^d \to \infty$ entonces, $d_\mu(S_n, S) \to 0$ en probabilidad. Nótese que las hipótesis establecidas sobre ϵ_n son idénticas a las que se imponen en el parámetro de ventana para asegurar la consistencia del estimador tipo núcleo en la estimación no paramétrica de la densidad.



$\epsilon_n = 1000$ $\epsilon_n = 3500$ $\epsilon_n = 6000$ $\epsilon_n = 11000$

Figura 2.4: Estimador de Devroye y Wise para distintos valores de ϵ_n en la velutina noviembre de 2018.

Por otro lado, Devroye and Wise (1980) estudiaron el problema de detectar el comportamiento anormal de un sistema. Para ello, consideran una muestra de observaciones independientes $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ según una densidad f cuando la máquina opera con normalidad, estableciendo S^c como zona de peligro. A continuación, dada una nueva observación X_{n+1} con densidad g (que puede ser distinta de f), el objetivo es determinar si el sistema presenta anomalías o no en función de si la distribución de X_{n+1} es distinta de f. El procedimiento que proponen para este contraste consiste en rechazar la hipótesis nula cuando X_{n+1} no pertenece a S_n.

El estimador (2.1) también se ha estudiado para la estimación de la frontera o del número de componentes conexas de S. Cuevas y Rodríguez-Casal (2004) abordaron la estimación de la frontera de S, ∂S, con respecto a la métrica de Hausdorff cuando la distribución es absolutamente continua, y tiene una densidad acotada por abajo. A diferencia de la d_H–consistencia de S_n, los resultados de consistencia para ∂S_n como estimador de ∂S no son tan inmediatos. Para garantizar que ∂S_n es un estimador d_H–consistente de ∂S de forma casi segura, se necesita que $\epsilon_n \to 0$ de forma casi segura y que $S \subset S_n$. Otra manera de asegurar dicha consistencia es asumir que S pertenece al modelo regular de Serra, que definiremos en la Sección 2.3, y seleccionar ϵ_n para que S_n también satisfaga dicha condición.

Cuevas *et al.* (2000) abordaron el problema de estimar el número de componentes conexas de S, $T(S)$, a partir de las de S_n, que denotaremos por T_n. Demostraron que, si los conjuntos de nivel pertenece al modelo regular de Serra y ε tiende a cero, se puede garantizar la consistencia de T_n como estimador de $T(S)$. Además, probaron que tomando ε_n como la mitad del mínimo de las distancias entre las componentes conexas, T_n es un estimador consistente de $T(S)$. Dado que ambos son números enteros, esto implica que $T_n = T(S)$ de forma casi segura. Sin embargo, una elección de ε_n próxima a cero puede conllevar a sobreestimar el número de componentes conexas de S. Por tanto, proponen un algoritmo de selección de ε_n empleando procedimientos de clustering de los datos y evaluando la distancia entre los puntos de un mismo clúster. Con respecto a la estimación de las componentes conexas, a lo largo del trabajo, emplearemos el número de componentes conexas de nuestro estimador para establecer un criterio de parada sobre el algoritmo que presentaremos en el Capítulo 3. Así, evitaremos que nuestro estimador se fragmente demasiado pues, nuestra intuición es que la forma de expandirse de la velutina es de forma continua, no por focos aislados.

2.2. Estimación de un soporte convexo

En esta sección, introduciremos los estimadores resultantes de asumir que nuestro soporte S es un conjunto convexo. En la Definición 2.3 definimos el concepto de conjunto convexo.

Definición 2.3. Un conjunto $A \subset \mathbb{R}^d$ es convexo si para cada par de puntos $x, y \in A$ y para todo $\lambda \in [0, 1]$, se cumple que $\lambda x + (1 - \lambda) y \in A$.

De acuerdo con la definición anterior, en el caso unidimensional, los únicos conjuntos convexos y compactos son los intervalos $[a, b]$ con $a \leq b$. En la Figura 2.5, podemos ver un ejemplo de conjunto convexo y otro no convexo en $\mathbb{R}^2$.

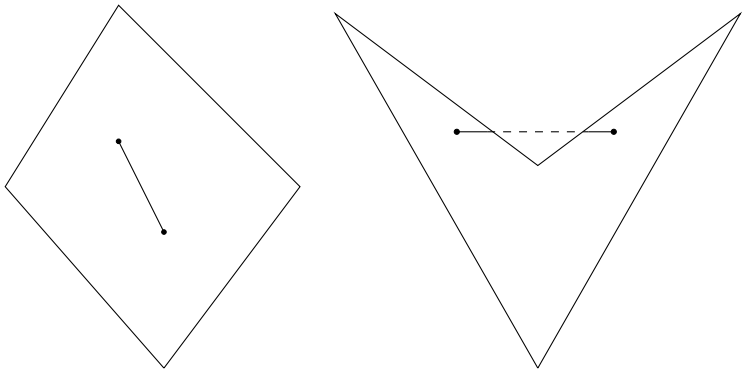


Figura 2.5: Representación conjunto convexo y no convexo.

Dado un conjunto A, el conjunto cerrado y convexo más pequeño que lo contiene es la envoltura convexa, la cual se define de la siguiente forma.

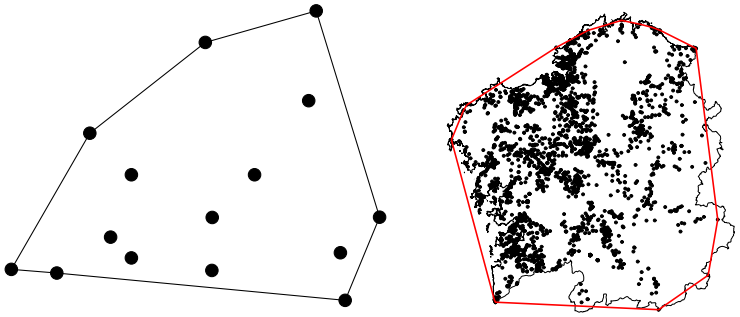
Definición 2.4. Sea $A \subset \mathbb{R}^d$, la envoltura convexa de A, $H(A)$, se define como la intersección de todos los conjuntos cerrados convexos de $\mathbb{R}^d$ que contienen a A.

Bajo la hipótesis de convexidad de S la envoltura convexa de la muestra, $H(\mathcal{X}_n)$, resulta ser un estimador natural del soporte. En la Figura 2.6 (izquierda) representamos la envoltura convexa de un conjunto de puntos. Sin embargo, a pesar de ser un estimador simple, no proporciona resultados satisfactorios cuando S no es convexo, por ejemplo, si posee más de una componente conexa o ciclos internos. En la Figura 2.6 (derecha) podemos ver como la envoltura convexa no sería un buen estimador del soporte para los datos de nidos de velutina en noviembre de 2016 ya que este incluye áreas de mar u océano donde no tiene sentido la existencia de dichos nidos. También podemos observar regiones de la zona suroriental de la provincia de Ourense sin ningún punto muestral, por lo que probablemente no estén contenidas en el soporte teórico y, por tanto, no deberían de estar incluidas en nuestro estimador del soporte.

Dümbgen and Walther (1996) estudiaron como aproxima la envoltura convexa de una muestra uniforme de puntos, $H(\mathcal{X}_n)$, al conjunto S. La proximidad entre dichos conjuntos se estudió en base a la distancia Hausdorff para una dimensión arbitraria d, obteniendo que $d_H(S, H(\mathcal{X}_n)) = O\left((\log n/n)^{1/d}\right)$ de forma casi segura. Además, si ∂S pertenece al modelo regular de Serra, que presentaremos más adelante, se cumple que $d_H(S, H(\mathcal{X}_n))$ es del orden de $(\log n/n)^{2/(d+1)}$.

2.3. Estimación de un soporte r-convexo

En la práctica, la suposición de convexidad sobre S puede resultar muy restrictiva y, por tanto, la envoltura convexa de la muestra puede no ser un estimador adecuado. Una restricción de forma sobre S más flexible que la convexidad es la r–convexidad para algún $r > 0$. Dicha condición la introducimos en la Definición 2.5.



$H(\mathcal{X}_{15})$.

Envoltura convexa de $\mathcal{X}_n$ de noviembre de 2016.

Figura 2.6: Envolturas convexas.

Definición 2.5. Un conjunto cerrado $A \subset \mathbb{R}^d$ es r–convexo si para algún $r > 0$, $A = C_r(A)$ donde

$$C_r(A) = \bigcap_{\{B_r(x):B_r(x)\cap A=\emptyset\}} (B_r(x))^{\mathsf{c}}$$

denota la envoltura r–convexa de A, siendo $B_r(x)$ la bola abierta de centro x y radio r.

De acuerdo con la Figura 2.7, el valor del parámetro r está relacionado con la forma del conjunto A. La envoltura r–convexa de A generaliza de forma natural a la envoltura convexa. La primera de ellas se calcula como la intersección del complementario de las bolas abiertas de radio r que no intersecan a A. Por otro lado, la segunda coincide con la intersección de los semiespacios cerrrados que contienen a A.

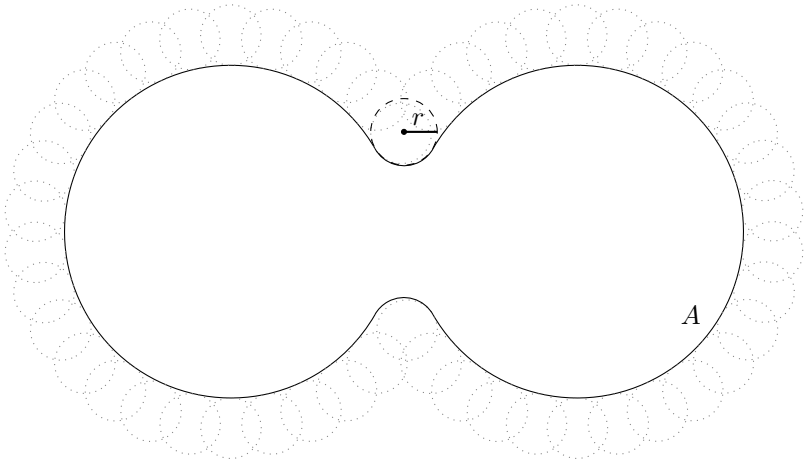


Figura 2.7: El conjunto A es igual a $C_r(A)$. Por tanto, A es r–convexo.

La envoltura r–convexa satisface muchas propiedades interesantes. Por ejemplo, $C_r(A) \subset C_{r^*}(A)$ para todo $r \leq r^*$ como podemos observar si comparamos las Figuras 2.7 y 2.8. Además, resulta fácil probar que si un conjunto A es cerrado y convexo, entonces es r–convexo para todo $r > 0$. En la Figura 2.9 se muestran estas relaciones mediante tres ejemplos.

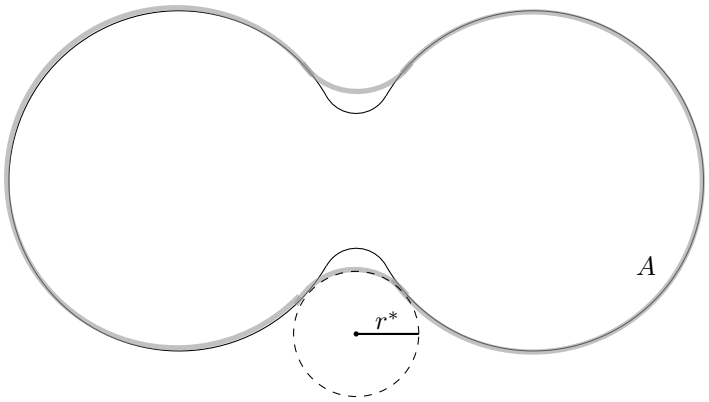


Figura 2.8: El conjunto A en negro no es igual a $C_r^*(A)$ en gris. Por tanto, A no es r^*–convexo.

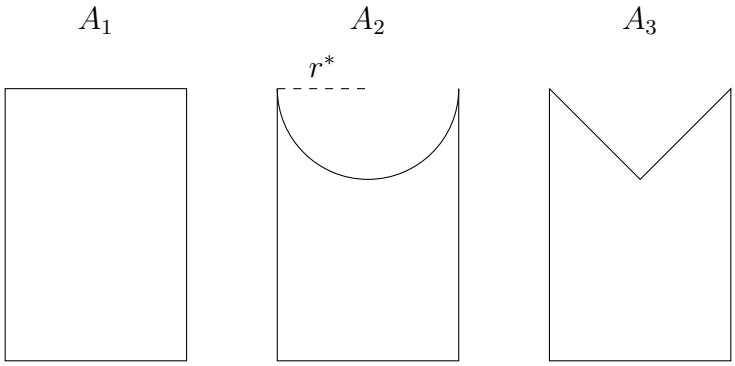


Figura 2.9: A_1 es convexo y, por tanto, r–convexo para todo $r > 0$. A_2 no es convexo pero sí r^*–convexo. A_3 no es convexo ni r–convexo para todo $r > 0$.

Walther (1997) estudió la relación entre la r–convexidad, el modelo regular de Serra y la condición de rodamiento libre que introduciremos a continuación.

Definición 2.6. El modelo regular de Serra se define como la clase de conjuntos compactos A que son morfológicamente abiertos y cerrados con respecto a la bola cerrada $B_r[0]$ de radio r para algún $r > 0$, esto es,

$$A = (A \ominus B_r[0]) \oplus B_r[0] = (A \oplus B_r[0]) \ominus B_r[0]$$

donde $\ominus$ denota la sustracción de Minkowski y $\oplus$ la adición de Minkowski. Ambas se definen como:

$$A \ominus C = \{x : x + C \subset A\},$$
$$A \oplus C = \{a + c : a \in A, c \in C\},$$

donde $x + C$ denota $\{x\} \oplus C$. Para $\delta \in \mathbb{R}$,

$$\delta A = \{\delta a : a \in A\}.$$

Una revisión completa de este modelo la podemos ver en Serra (1984). La envoltura r–convexa de la muestra $\mathcal{X}_n$ también se puede escribir en función de los operadores de adición y sustracción:

$$C_r(\mathcal{X}_n) = (\mathcal{X}_n \oplus B_r(0)) \ominus B_r(0).$$

A continuación, definiremos la condición de rodamiento libre. En la Figura 2.10 se representan dos conjuntos que satisfacen la propiedad de rodamiento libre.

Definición 2.7. Sea $A \subset \mathbb{R}^d$ un conjunto cerrado y $r > 0$. Se dice que una bola de radio r rueda libremente en A si cada punto frontera $a \in \partial A$ está contenido en una bola de radio r cuyo interior no interseca con A.

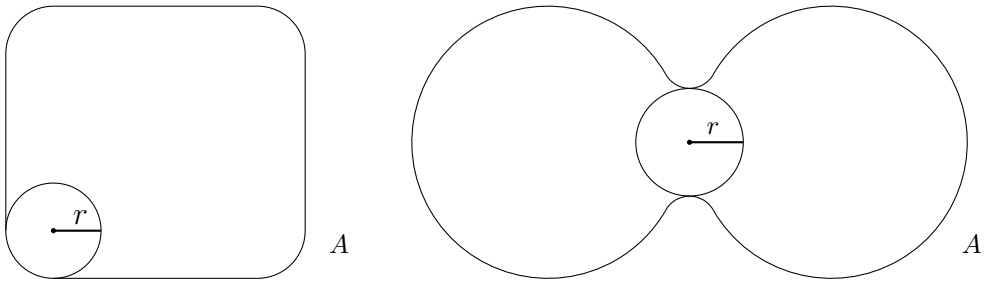


Figura 2.10: Una bola de radio r rueda libremente en $A \subset \mathbb{R}^2$.

En el Teorema 2.8 se enuncia la relación entre estas tres familias de conjuntos. Además, Walther (1997) establece una caracterización geométrica del modelo regular de Serra en términos de la r–convexidad, la condición de rodamiento libre y restricciones de suavidad sobre la frontera.

Teorema 2.8 (Walther, 1997)**.** *Sea $A \neq \emptyset$ un subconjunto compacto de $\mathbb{R}^d$ y $\alpha_0 > 0$. Entonces los siguientes enunciados son equivalentes:*

1. $(A \oplus lB_1[0]) \ominus A = A$ *con* $l \in [0, \alpha_0)$ *y* $(A \ominus lB_1[0]) \oplus A = A$ *con* $l \in [0, \alpha_0)$.
2. *A y $\bar{A}^{\mathsf{c}}$ son α_0–convexos e* $int(A_i) \neq \emptyset$ *para cada componente conexa por caminos* $A_i \subset A$.
3. *Una bola de radio l rueda libremente dentro de cada componente conexa por caminos de A y $\bar{A}^{\mathsf{c}}$ para todo* $0 \leq l \leq \alpha_0$.
4. *∂A es una subvariedad $(d-1)$–dimensional en $\mathbb{R}^d$ con el vector normal exterior $\eta(s)$ en $s \in \partial A$ satisfaciendo la condición de Lipschitz*

$$\|\eta(s) - \eta(t)\| \leq \frac{1}{\alpha_0}\|s - t\| \quad \text{para todo } s,\, t \in \partial A.$$

Además, para algún $\alpha_0 > 0$ lo anterior es equivalente a:

5. *A pertenece al modelo regular de Serra.*

Volviendo a la estimación del soporte, bajo la suposición de r–convexidad, el estimador natural de S será la envoltura r-convexa:

$$S_n = C_r(\mathcal{X}_n).$$

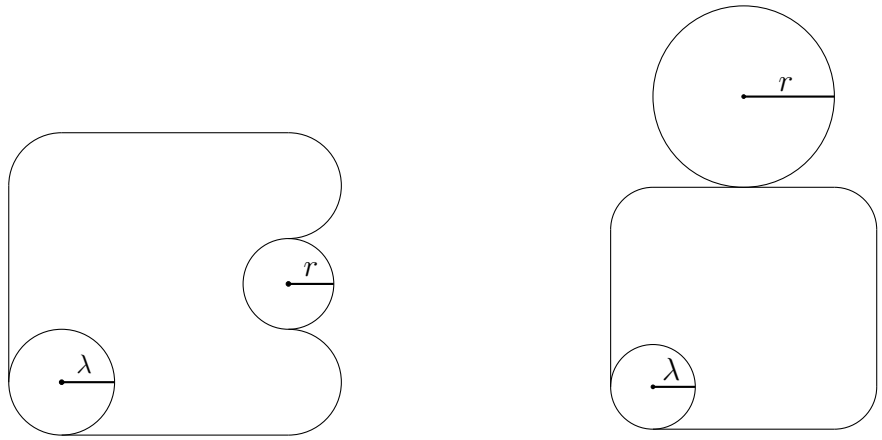
Dicho estimador fue estudiado por Rodríguez-Casal (2007) bajo la hipótesis de que S pertenece al modelo regular de Serra. Si S es r–convexo, se satisface que $d_H(S, C_r(\mathcal{X}_n)) = O\left((\log n/n)^{1/d}\right)$ de forma casi segura. Cabe destacar que, aunque la familia de los conjuntos que satisfacen la propiedad de r–convexidad es mucho más amplia que la de los conjuntos convexos, la tasa de convergencia de $d_H(S, C_r(\mathcal{X}_n))$ es la misma. Sin embargo, si $\mathcal{X}_n$ sigue una distribución absolutamente continua en S, su densidad asociada está acotada por debajo por una constante positiva y S pertenece al modelo regular de Serra se cumple que $d_H(S, C_r(\mathcal{X}_n)) = O\left((\log n/n)^{2/(d+1)}\right)$ de forma casi segura, consiguiendo así la misma que teníamos para $H(\mathcal{X}_n)$. Las mismas tasas de convergencia se obtienen para

$d_H\left(\partial S, \partial C_r\left(\mathcal{X}_n\right)\right)$ y $d_\mu\left(S, C_r\left(\mathcal{X}_n\right)\right)$.

En la literatura es común establecer que S satisfaga las condiciones del Teorema 2.8. Sin embargo, para nuestro objetivo solo es necesario que se cumpla la siguiente condición.

(R) S y $\bar{S^c}$ satisfacen la propiedad de rodamiento para constantes positivas r y λ, respectivamte. (2.2)

En la Figura 2.11 (izquierda) se ilustra dicha propiedad. La condición (R) incluye el caso en el que $r = \infty$ y λ próximo a cero, Figura 2.11 (derecha). Rodríguez-Casal y Saavedra-Nieves (2016) demuestran que, bajo (R) para todo $\lambda > 0$, se cumple que S es r–convexo. Por tanto, (R) es una condición suficiente para garantizar la r–convexidad del soporte S pero, no es una condición necesaria. La Figura 2.12 muestra tres conjuntos r–convexos que no satisfacen la condición (R) para ningún $\lambda > 0$.



(R) es una condición más general. $r > 0$ puede ser muy grande y $\lambda > 0$ cercano a cero.

Figura 2.11: Conjuntos que satisfacen la condición (R).

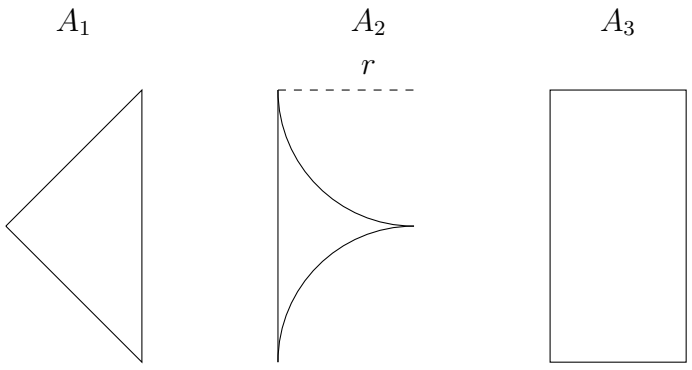
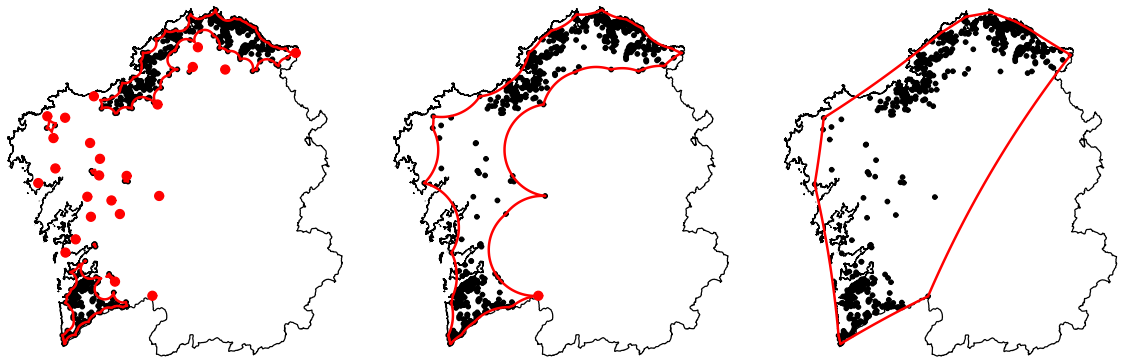


Figura 2.12: A_1, A_2 y A_3 son conjuntos r-convexo que no satisfacen la propiedad (R) para ningún $\lambda > 0$.

En la práctica, S es desconocido y por tanto, también lo es el valor de r. En la Figura 2.13, dada la muestra de 711 nidos de velutinas registrados en julio de 2015, se muestra la influencia que tiene el parámetro r en el estimador $C_r\left(\mathcal{X}_n\right)$.



$C_{6000}(\mathcal{X}_{711})$ $C_{29000}(\mathcal{X}_{711})$ $C_{700000}(\mathcal{X}_{711})$.

Figura 2.13: Influencia del parámetro r en el estimador $C_r(\mathcal{X}_n)$.

Si r es próximo a cero, se pueden dar dos casos: o $C_r(\mathcal{X}_n)$ se ajusta a la muestra de puntos o resulta ser un estimador muy fragmentando. Sin embargo, para valores grandes de r, $C_r(\mathcal{X}_n)$ coincide prácticamente con la envoltura convexa de los puntos como se puede ver en la Figura 2.13 (derecha). En este caso, podemos encontrar una zona o hueco suficientemente grande dentro del estimador que no contiene ningún punto de la muestra o regiones de mar, las cuales tampoco tendrían sentido que formasen parte de nuestro estimador. Por tanto, debemos determinar con precisión el valor óptimo de r para estimarlo a partir de la muestra. En el capítulo siguiente revisaremos los métodos existentes para estimar dicho parámetro y presentaremos detalladamente la propuesta de Rodríguez-Casal y Saavedra-Nieves (2019).

Capítulo 3

Estimación de la EOO

Hasta ahora, hemos visto que el problema de la estimación del soporte trata de reconstruir el soporte compacto no vacío $S \subset \mathbb{R}^d$ de un vector aleatorio absolutamente continuo X a partir de una muestra aleatoria de puntos $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. Bajo la suposición de r–convexidad de S, la envoltura r–convexa de los puntos de la muestra $C_r(\mathcal{X}_n)$ resulta ser un estimador natural de S. En la práctica, dicho estimador se puede calcular como la intersección de los complementarios de todas las bolas de radio mayor o igual que r que no intersecan con $\mathcal{X}_n$.

Sin embargo, $C_r(\mathcal{X}_n)$ depende de un parámetro r desconocido. En el capítulo anterior, en la Figura 2.13, ilustramos la importancia de seleccionar un valor adecuado para este parámetro de suavizado. Si r es próximo a cero, $C_r(\mathcal{X}_n)$ coincide prácticamente con $\mathcal{X}_n$, mientras que para valores grandes de r, se aproxima a la envoltura convexa de los puntos. El objetivo del método que presentaremos a lo largo de este capítulo es seleccionar de forma automática el valor del parámetro de suavizado de la envoltura r–convexa dada la muestra aleatoria de puntos. Comenzaremos introduciendo el problema de estimación de dicho parámetro y comentaremos brevemente los métodos existentes en la literatura. A continuación, presentaremos con detalle la propuesta de Rodríguez-Casal y Saavedra-Nieves (2019) para luego desarrollar el procedimiento de su cálculo. Para ello, introduciremos la noción de spacing maximal y un contraste de hipótesis de r–convexidad. Por último, detallaremos el algoritmo que, bajo la hipótesis de r–convexidad, nos permitirá calcular el estimador del soporte.

3.1. Selección del parámetro óptimo

Dada la influencia del parámetro r en el estimador $C_r(\mathcal{X}_n)$, es necesario conseguir una buena estimación de r a la hora de estimar el soporte r–convexo. En la literatura existen varios métodos dedicados a la estimación de dicho parámetro. Mandal and Murthy (1997) se centran en el caso particular de $\mathbb{R}^2$ y proponen un método de selección basado en la teoría de grafos, en concreto, en el árbol de expansión minimal. Bajo la suposición de uniformidad de la muestra, Rodríguez-Casal y Saavedra-Nieves (2016) presentan un método de selección basado en un contraste de hipótesis de uniformidad de X en S y en la noción de spacing maximal. Este método ha sido generalizado recientemente en Rodríguez-Casal y Saavedra-Nieves (2019) cuando la distribución de la muestra no es necesariamente uniforme, permitiendo así tratar el caso de observaciones sesgadas. Dicho método será el que expliquemos en detalle a continuación.

El primer paso para construir el estimador de r es determinar con precisión el valor óptimo de dicho parámetro. Para ello, se tendrá en cuenta la siguiente propiedad: si S es r–convexo para $r > 0$ entonces es r^*–convexo para todo $0 < r^* \leq r$. Además, se cumple que $C_{r^*}(\mathcal{X}_n) \subset C_r(\mathcal{X}_n)$. Por este

motivo, parece razonable estimar el mayor valor de r para el cual se verifica que S es r–convexo. Sin embargo, si S es convexo se tiene que S es r–convexo para todo $r > 0$. Por tanto, y por simplicidad en la definición, asumiremos que S no es convexo. En consecuencia, si S es r–convexo y no convexo se cumple que $\{\gamma > 0 : C_\gamma(S) = S\}$ es un conjunto no vacío acotado superiormente, pudiendo así presentar la siguiente Definición 3.1.

Definición 3.1. Sea $S \subset \mathbb{R}^d$ un conjunto no vacío, compacto, no convexo y r–convexo para algún $r > 0$. Bajo estas hipótesis definimos

$$r_0 = \sup\{\gamma > 0 : C_\gamma(S) = S\}. \tag{3.1}$$

Resulta inmediato que si S es convexo, r_0 será infinito. La Proposición 2.4 de Rodríguez-Casal y Saavedra-Nieves (2016) garantiza que, bajo (R), el supremo establecido en la Definición 3.1 es un máximo y, en consecuencia, S es r–convexo para todo $r \leq r_0$. Para la demostración, consideran una sucesión $\{r_n\}$ que converge a r_0 tal que $C_{r_n}(S) = S$, dicha sucesión existe por la Definición 3.1. Se puede probar que S satisface la propiedad de r_n–rodamiento, entonces, por el Lema 3.2.9 de Rodríguez-Casal y Saavedra-Nieves (2016), esta propiedad se conserva en el límite y, en consecuencia, S también cumple el r_0–rodamiento. Finalmente, bajo (R), el r_0–rodamiento implica que S es r_0–convexo.

En resumen, bajo (R) queda justificada la optimalidad de (3.1) pues, si S es r–convexo para $r < r_0$, $C_r(\mathcal{X}_n)$ no es un estimador admisible, ya que $C_{r_0}(\mathcal{X}_n)$ siempre nos proporcionará mejores resultados. Esto se debe a que, con probabilidad uno, $C_r(\mathcal{X}_n) \subset C_{r_0}(\mathcal{X}_n) \subset S$. Además, para $r > r_0$, incluso para valores de r muy cercanos a r_0, $C_r(\mathcal{X}_n)$ puede sobrestimar considerablemente a S. Por ejemplo, si S es una corona circular como la de la Figura 3.1 y $r > r_0$, $C_r(S)$ coincide con la bola cerrada cuyo borde se corresponde con la circunferencia exterior.

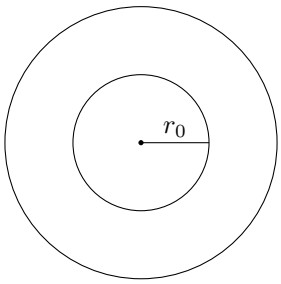


Figura 3.1: Corona circular cuya circunferencia interior es de radio r_0.

A continuación, vamos a introducir algunas nociones de la teoría de spacings maximales para así poder establecer una estimación consistente de r_0. Dicho spacing nos permitirá cuantificar el tamaño de los huecos o zonas sin puntos de la muestra dentro del estimador. De esta forma, si existe un hueco suficientemente grande dentro de $C_r(\mathcal{X}_n)$, esto se traducirá en un valor grande del spacing maximal, indicando así la necesidad de seleccionar un r menor. Esta idea se puede ver en la Figura 3.2 donde representamos en rojo el estimador resultante de tomar $r = 60000$ para la muestra de velutinas de julio de 2017. En ella está representada en azul una bola de radio 1600 contenida dentro de $C_{60000}(\mathcal{X}_{711})$ que no interseca con ningún punto de $\mathcal{X}_n$. Como comentábamos anteriormente, la existencia de dicha bola nos está indicando que estamos sobreestimando S al tomar $r = 60000$ y que debemos seleccionar un r menor.

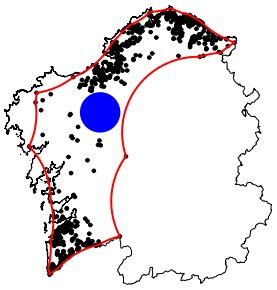


Figura 3.2: $C_{60000}\left(\mathcal{X}_{711}\right)$ en julio de 2017.

La noción de spacing maximal en varias dimensiones fue introducida por Deheuvels (1983) para el caso donde los datos están distribuidos uniformemente en el cubo unidad. A continuación, Janson (1987) extendió estos resultados para datos con distribución uniforme en un conjunto acotado y calculó la distribución asintótica del spacing maximal sin asumir restricciones de forma en el soporte S. Aaron *et al.* (2017) generalizaron los resultados de Janson (1987) al caso no uniforme.

La forma de los spacings considerados vendrá dada por un conjunto $A \subset \mathbb{R}^d$ que supondremos compacto y convexo. Desde el punto de vista práctico, consideraremos el cubo unidad $A = [0,1]^d$ o la bola cerrada de centro 0 y radio 1, $A = B_1[0]$. Para el caso d–dimensional y muestreo uniforme, la definición de spacing maximal para S introducida por Jason (1987) es:

$$\Delta_n^*\left(\mathcal{X}_n\right) = \left\{\gamma \,:\, \exists x \text{ tal que } \{x\} \oplus \gamma A \subset S \setminus \mathcal{X}_n\right\}.$$

Aaron *et al.* (2017) extendieron la definición del spacing maximal bajo las hipótesis de que $\mathcal{X}_n$ se distribuye de acuerdo a una función de densidad f con soporte acotado S, la medida de Lebesgue del conjunto A es uno y su baricentro es el origen de $\mathbb{R}^d$. Además, la función f debe de satisfacer la siguiente condición:

$\left(f_{0,1}^L\right)$ La restricción de la densidad f a S es Lipschitz continua (existe k_f tal que $\forall x, y \in S$, $|f(x) - f(y)| \leq k_f \|x - y\|$) y existe $f_0 > 0$ tal que $f(x) \geq f_0$ para todo $x \in S$. Además, denotaremos $f_1 = \text{máx}_{x \in S} f(x)$.

De aquí en adelante, asumiremos que la muestra aleatoria de puntos, $\mathcal{X}_n$, ha sido generada a partir de una densidad f que satisface la condición $\left(f_{0,1}^L\right)$. En este caso general, el spacing maximal se define como

$$\Delta_n\left(\mathcal{X}_n\right) = \left\{\gamma \,:\, \exists x \text{ tal que } \{x\} \oplus \frac{\gamma}{f(x)^{1/d}} A \subset S \setminus \mathcal{X}_n\right\}$$

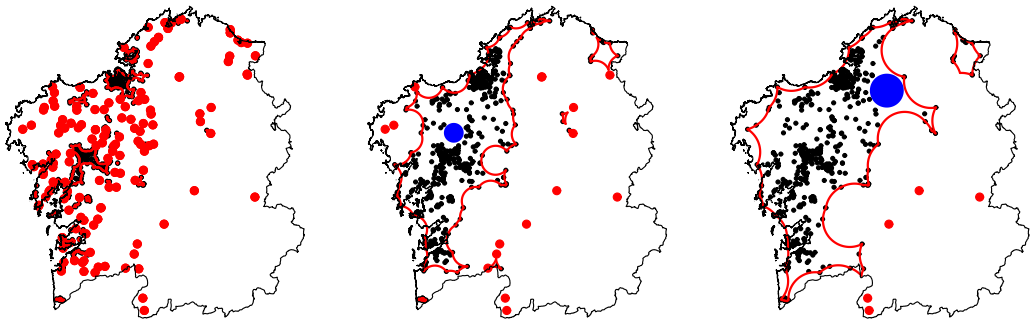
y

$$V_n\left(\mathcal{X}_n\right) = \Delta_n\left(\mathcal{X}_n\right)^d.$$

Obsérvese que la definición anterior depende de la densidad f, por tanto, $\Delta_n\left(\mathcal{X}_n\right)$ distinguirá las regiones de poca, de las de mucha densidad. Por otro lado, dado que $\Delta_n\left(\mathcal{X}_n\right)$ depende de S y de f, ambos desconocidos, deberemos estimarlos.

A continuación, ilustramos a través de nuestros datos la idea del spacing maximal. De acuerdo con la Figura 3.3 (centro y derecha), tanto $C_{10000}\left(\mathcal{X}_{592}\right)$ como $C_{19000}\left(\mathcal{X}_{592}\right)$ poseen zonas sin puntos de la muestra. En ambos casos el estimador contiene una bola, representada en azul, que no interseca con ningún punto de $\mathcal{X}_n$. La existencia de estas zonas nos está indicando que el valor del spacing es

considerablemente mayor que cero. De hecho, podemos esperar que el spacing maximal del estimador $C_{19000}(\mathcal{X}_{592})$ sea mayor que el de $C_{10000}(\mathcal{X}_{592})$. Lo contrario ocurre en Figura 3.3 (izquierda) donde $C_{3000}(\mathcal{X}_{592})$ prácticamente se ajusta a la muestra y aparentemente no hay zonas sin datos, por lo que el spacing maximal será cercano a cero.



$C_{3000}(\mathcal{X}_{592})$ $C_{10000}(\mathcal{X}_{592})$ $C_{19000}(\mathcal{X}_{592})$

Figura 3.3: Ilustración del spacing maximal empleando los datos de velutina de abril de 2017.

Rodríguez-Casal y Saavedra-Nieves (2019) modificaron a su vez las hipótesis sobre f y S que establecieron Aaron *et al.* (2017) para obtener la distribución asintótica de V_n. Obtuvieron así, que dicho resultado sigue verificándose si se asume que S cumple la condición (R) definida en (2.2). Conocer la distribución asintótica de V_n nos permite poder aproximar la magnitud de las zonas sin muestra dentro de nuestro estimador. Por ejemplo, empleando dicha distribución asintótica podemos saber si las bolas azules de la Figura 3.3 (centro y derecha) son demasiado grandes y debemos considerar valores de r menores para ajustar mejor el soporte. El resultado obtenido por Rodríguez-Casal y Saavedra-Nieves (2019) se enuncia a continuación.

Teorema 3.2. *Sea $\mathcal{X}_n$ una muestra aleatoria i.i.d de acuerdo a una densidad f que cumple $\left(f_{0,1}^L\right)$ con soporte compacto no vacío S bajo (R). Sea U una variable aleatoria con distribución*

$$\mathbb{P}(U \leq u) = \exp(-\exp(-u)) \text{ para } u \in \mathbb{R}$$

y sea

$$\beta = \frac{1}{d!}\left(\frac{\sqrt{\pi}\,\Gamma\left(\frac{d}{2}+1\right)}{\Gamma\left(\frac{d+1}{2}\right)}\right)^{d-1}$$

Entonces, se cumple que

$$U(\mathcal{X}_n) \xrightarrow{d} U \text{ cuando } n \to \infty,$$

$$\liminf_{n\to\infty} \frac{nV_n(\mathcal{X}_n) - \log(n)}{\log(\log(n))} \geq d-1 \text{ c.s., } \limsup_{n\to\infty} \frac{nV_n(\mathcal{X}_n) - \log(n)}{\log(\log(n))} \leq d+1 \text{ c.s.,}$$

donde

$$U(\mathcal{X}_n) = nV_n(\mathcal{X}_n) - \log(n) - (d-1)\log(\log(n)) - \log(\beta).$$

Entonces, como vimos anteriormente, para estimar $\Delta_n(\mathcal{X}_n)$ tendremos que estimar a su vez S y f, pues ambos son desconocidos en la práctica. Bajo la suposición de r–convexidad, S se estimará a partir de la envoltura r–convexa de la muestra, $C_r(\mathcal{X}_n)$. Para estimar la correspondiente función de densidad f introduciremos el siguiente estimador.

Definición 3.3. Sean $r > 0$ y $Vor(X_i)$ la región de Voronoi correspondiente al punto X_i (es decir, $\text{Vor}(X_i) = \{x : \|x - X_i\| = \text{mín}_{y \in \mathcal{X}_n} \|x - y\|\}$). Si K es una función kernel (es decir, $K \geq 0$, $\int K = 1$ y $\int uK(u)\,du = 0$) y $f_n(x) = \frac{1}{n h_n^d} \sum K((x - X_i)/h_n)$ denota al estimador kernel usual de la densidad, definimos

$$\widehat{f}_n(x) = \underset{i: x \in Vor(X_i)}{\text{máx}} f_n(X_i).$$

Este estimador, el cual difiere ligeramente del propuesto por Aaron *et al.* (2017), solo toma n valores distintos correspondientes al valor del estimador kernel f_n en los puntos de la muestra. De hecho, para cada punto $x \in S$, $\widehat{f}_n(x)$ es igual a $f_n(X_i)$ donde X_i es el punto de la muestra más próximo a x. Se puede probar que, con probabilidad 1 y para n suficientemente grande, existe un punto de $\mathcal{X}_n$ tan cerca de x como se quiera siempre que $x \in S$. Por tanto, este estimador de la densidad propuesto resulta ser una simplificación del habitual. Esto se traduce en ventajas a nivel computacional a la hora de estimar Δ_n y V_n.

Por otro lado, se deben establecer algunas hipótesis técnicas sobre la función kernel.

$\left(\mathcal{K}_\phi^p\right)$ La función Kernel K pertenece al conjunto de los kernels, $\mathcal{K}$, tales que $K(u) = \phi(p(u))$ donde p es un polinomio y ϕ es una función real acotada con varianza acotada, satisfaciendo que $c_K = \int \|u\| K(u)\,du < \infty$, $K \geq 0$ y existen r_k y $c'_K > 0$ tal que $K(x) \geq c'_K$ para todo $x \in B_{r_K}[0]$.

Por ejemplo, el kernel gaussiano cumple la condición anterior. Establecidas las condiciones anteriores, Rodríguez-Casal y Saavedra-Nieves (2019) proponen el siguiente estimador plug-in de $\Delta_n(\mathcal{X}_n)$:

$$\widehat{\delta}_n(C_r(\mathcal{X}_n) \setminus \mathcal{X}_n) = \sup\left\{\gamma : \exists x \text{ tal que } \{x\} \oplus \frac{\gamma}{\widehat{f}_n(x)^{1/d}} A \subset C_r(\mathcal{X}_n) \setminus \mathcal{X}_n\right\} \tag{3.2}$$

y

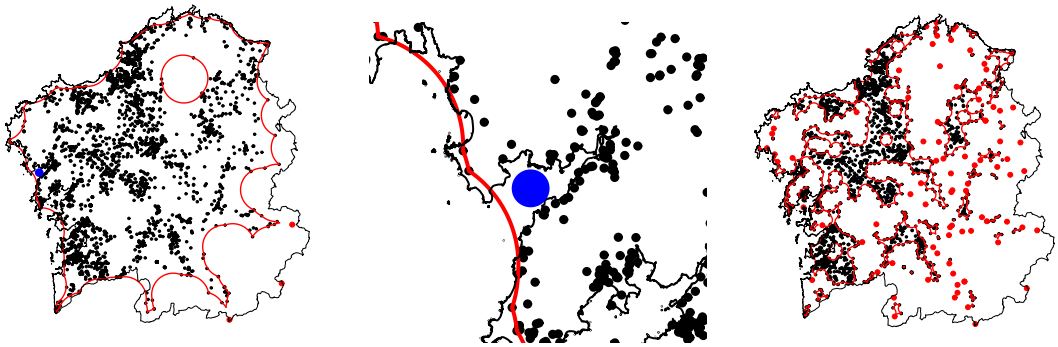
$$\widehat{V}_{n,r} = \widehat{\delta}_n(C_r(\mathcal{X}_n) \setminus \mathcal{X}_n)^d.$$

Dada la definición de $\widehat{f}_n$ y la hipótesis $\left(f_{0,1}^L\right)$, por el Lema 5 de Aaron *et al.* (2017) se espera que $\widehat{f}_n$ no tienda a cero en S. Esto resulta de vital importancia en la fórmula (3.2). Por otro lado, si S es r–convexo, $\widehat{\delta}_n(C_r(\mathcal{X}_n) \setminus \mathcal{X}_n)$ debe converger a cero conforme aumenta el tamaño muestral, pues cada vez habrá menos espacio para que pueda existir un punto en el que podamos centrar una bola de radio $\gamma/\widehat{f}_n(x)^{1/d}$ que no interseque con ningún punto de la muestra. Sin embargo, si $S \subsetneq C_r(S)$, se espera que el estimador plug-in de $\Delta_n(\mathcal{X}_n)$ converja a una constante positiva.

Introducido el concepto de spacing maximal, presentaremos un contraste de hipótesis sobre las r–convexidad de S para estimar r_0. Dado $r > 0$, se contrastará la hipótesis nula de que S es r–convexo empleando $\widehat{V}_{n,r}$ como estadístico. La idea que respalda este procedimiento es que, bajo $\left(f_{0,1}^L\right)$ y (R), si el estadístico toma valores suficientemente grandes significará que el r seleccionado no es apropiado y deberemos considerar un valor más pequeño para r.

Ilustraremos el comportamiento del test con nuestro conjunto de datos de la velutina de noviembre de 2018. Dada la muestra $\mathcal{X}_{2635}$, se representa su envoltura r–convexa para $r = 15000$ en rojo en la Figura 3.4 (izquierda). En ella se puede observar que parte de la Ría de Muros y Noia está incluida en el estimador, de hecho hemos representado una bola azul en esta zona de mar dentro de $C_{15000}(\mathcal{X}_{2635})$. Por tanto, el soporte de nuestros datos, es decir la EOO, está siendo sobrestimado y, en consecuencia, $\widehat{V}_{2635,15000}$ será grande. De hecho, su volumen tenderá a una constante por mucho que aumente el tamaño muestral ya que siempre podremos encontrar dicha bola azul dentro de nuestro estimador. En

conclusión, la hipótesis nula de 15000–convexidad será rechazada. Si contrastáramos la hipótesis de 4000–convexidad, Figura 3.4 (derecha), ocurre lo contrario. En este caso, $\widehat{V}_{2635,4000}$ será claramente más pequeño y, además, su volumen tenderá a cero según aumente el tamaño muestral. Por tanto, no rechazaremos la hipótesis de 4000–convexidad y podemos considerar un r mayor. Formalmente, el comportamiento asintótico del test se enuncia en el Teorema 3.4.



$C_{15000}\left(\mathcal{X}_{2635}\right)$ $C_{4000}\left(\mathcal{X}_{2635}\right)$

Figura 3.4: Ilustración del test mediante la velutina en noviembre de 2018.

Teorema 3.4 (Rodríguez-Casal y Saavedra-Nieves, 2019)**.** *Sean $r > 0$ y $\mathcal{X}_n$ una muestra aleatoria i.i.d según una densidad f que cumple $\left(f_{0,1}^{L}\right)$ con soporte compacto no vacío S bajo (R). Sea f_n el estimador de la densidad introducido en la Definición 3.3 cuya función kernel satisface la condición $\left(\mathcal{K}_{\phi}^{p}\right)$ y la sucesión de parámetros de suavizado h_n cumple que $h_n = O\left(n^{-\varsigma}\right)$ para algún $0 < \varsigma < 1/d$. Sea $\widehat{\delta}_n\left(C_r\left(\mathcal{X}_n\right) \setminus \mathcal{X}_n\right)$ el estimador del spacing maximal introducido en la ecuación (3.2). Dado el siguiente contraste de hipótesis:*

$$H_0 : S \text{ es } r\text{–convexo versus } H_1 : S \text{ no es } r\text{–convexo.}$$

(a) El test basado en el estadístico $\widehat{V}_{n,r} = \widehat{\delta}_n\left(C_r\left(\mathcal{X}_n\right) \setminus \mathcal{X}_n\right)^d$ con región crítica $RC = \left\{\widehat{V}_{n,r} > c_{n,\alpha}\right\}$, donde

$$c_{n,\alpha} = \frac{1}{n}\left(-\log\left(-\log\left(1-\alpha\right)\right) + \log\left(n\right) + \left(d-1\right)\log\left(\log\left(n\right)\right) + \log\left(\beta\right)\right)$$

tiene un nivel asintónico menor o igual que α.

(b) Además, si S no es r–convexo, se cumple que $\mathbb{P}\left(\widehat{V}_{n,r} > c_{n,\alpha}, c.s.\right) = 1$.

Observación 3.5. El Teorema 3.4 es una adaptación para conjuntos r–convexos del Teorema 4 de Aaron *et al.* (2017) para el caso de soporte desconocido y convexo. En el Teorema 4 de Aaron *et al.* (2017) se plantea el contraste de convexidad de S. En él se emplea un estimador de la densidad ligeramente distinto y se toma como estimador del spacing maximal $\widehat{\delta}_n\left(H\left(\mathcal{X}_n\right) \setminus \mathcal{X}_n\right)$, resultante de sustituir $C_r\left(\mathcal{X}_n\right)$ por $H\left(\mathcal{X}_n\right)$ en la ecuación (3.2). Las condiciones que se imponen sobre la densidad f y la función kernel coinciden en ambos teoremas. Por último, en el Teorema 4 también se prueba que el test tiene nivel asintótico α, y que bajo la hipótesis alternativa, el test se rechaza, de forma casi segura, para n suficientemente grande.

Observación 3.6. Nótese que la sucesión habitual de parámetros de suavizados para estimar f, $h_n = h_0 n^{-1/(d+4)}$, satisface las hipótesis del Teorema 3.4. Por tanto, cualquier selector de ventana razonable será adecuado para testar la r–convexidad.

Dado que el parámetro de forma r_0 bajo (R) es el máximo del conjunto $\{\gamma > 0 : C_\gamma(S) = S\}$, las hipótesis del test en el Teorema 3.4 se pueden reescribir como sigue:

$$H_0 : r \leq r_0 \text{ versus } H_1 : r > r_0.$$

Obsérvese que, bajo H_1, $S = C_{r_0}(S) \subsetneq C_r(S)$. Para n suficientemente grande, $C_r(X_n)$ y S serán lo suficientemente distintos para garantizar que se rechaza la hipótesis nula, tal y como establece el apartado (b) del Teorema 3.4. Así, en base al test que acabamos de presentar y a la Definición 3.1, r_0 se estimará como:

$$\widehat{r}_0 = \{\gamma > 0 : \text{Se acepta } H_0 \text{ en } C_\gamma(\mathcal{X}_n)\}. \tag{3.3}$$

Es decir, se propone como estimador el mayor valor de r compatible con la hipótesis de r–convexidad. Nótese que dicha elección depende del nivel de significación elegido para el test. En el siguiente teorema se recogen las propiedades teóricas del estimador de r_0.

Teorema 3.7 (Rodríguez-Casal y Saavedra-Nieves, 2019)**.** *Sea f una función de densidad que satisface $\left(f_{0,1}^L\right)$ con soporte S no vacío, no convexo y compacto bajo (R). Sea $\widehat{f}_n$ el estimador de la densidad introducido en la Definición 3.3 cuya función kernel cumple $\left(\mathcal{K}_\phi^p\right)$ y la sucesión de parámetros de suavizado h_n tal que $h_n = O\left(n^{-\varsigma}\right)$ para algún $0 < \varsigma < 1/d$. Sea r_0 el parámetro definido en (3.1) y $\widehat{r}_0$ definido en (3.3). Sea $\{\alpha_n\} \subset (0,1)$ una sucesión de niveles de significación convergente a 0 tal que $\log(\alpha_n)/n \to 0$. Entonces, $\widehat{r}_0$ converge a r_0 en probabilidad.*

Una vez calculado $\widehat{r}_0$, el estimador propuesto para el soporte será $C_{\widehat{r}_0}(\mathcal{X}_n)$. La consistencia de dicho estimador se puede demostrar a partir del Teorema 3.7 si se cumple que $\lim_{r \to r_0^+} d_H(S, C_r(S)) = 0$. Sin embargo, dicha consistencia no se puede garantizar si no se satisface esta condición. Esto se puede ver claramente en el caso de la corona circular de la Figura 3.1, pues si consideramos un r ligeramente mayor a r_0, $C_r(\mathcal{X}_n)$ pasará a incluir la bola interna de radio r_0, sobreestimando considerablemente la corona circular. Para solventarlo a nivel teórico, consideraremos el estimador $C_{r_n}(\mathcal{X}_n)$ donde $r_n = \nu\widehat{r}_0$ y $\nu \in (0,1)$ fijo. Esto nos asegura que, para n suficientemente grande, $C_{r_n}(\mathcal{X}_n) \subset S$ con una probabilidad muy alta. En la práctica, la elección de ν no supone ningún problema ya que aproximaremos $\widehat{r}_0$ numéricamente y el estimador resultante siempre cumplirá esta condición sin necesidad de multiplicar por ν. En el siguiente teorema enunciamos una desigualdad para el estimador del soporte $C_{\widehat{r}_0}(\mathcal{X}_n)$.

Teorema 3.8 (Rodríguez-Casal y Saavedra-Nieves, 2019)**.** *Sean $r > 0$ y $\mathcal{X}_n$ una muestra aleatoria i.i.d según una densidad f que cumple $\left(f_{0,1}^L\right)$ con soporte compacto, no convexo y no vacío S bajo (R). Sea r_0 el parámetro definido en 3.1 y $\widehat{r}_0$ definido en 3.3. Sea $\{\alpha_n\}$ una sucesión convergente a cero tal que $\log(\alpha_n)/n \to 0$. Sea $\nu \in (0,\, 1)$ y $r_n = \nu\widehat{r}_0$. Entonces, de forma casi segura,*

$$d_H(S, C_{r_n}(\mathcal{X}_n)) \leq D\left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}$$

para alguna constante positiva D. El mismo radio de convergencia se mantiene para $d_H(\partial S, \partial C_{r_n}(\mathcal{X}_n))$ y $d_\mu(S \triangle C_{r_n}(\mathcal{X}_n))$.

3.2. Algoritmo

Los aspectos numéricos del algoritmo de estimación de r_0 se detallan a continuación. Aunque desde el punto de vista teórico el método depende únicamente de la muestra $\mathcal{X}_n$, su implementación va a depender también del nivel de significación α y del número máximo de componentes conexas $\mathcal{C}$ del estimador resultante.

Aunque bajo las hipótesis del Teorema 3.7 la existencia del estimador $\widehat{r}_0$ está garantizada para n suficientemente grande con probabilidad uno, en la práctica puede darse el caso de que no exista dicho estimador para una muestra específica $\mathcal{X}_n$ y un nivel de significación α. Por tanto, se debe tener en cuenta la influencia de α pues, se rechazará la hipótesis nula de r–convexidad para $0 < r \leq r_0$ con probabilidad α, aproximadamente. Este hecho carece de importancia desde el punto de vista teórico ya que asumimos que $\alpha = \alpha_n$ tiende a cero conforme aumenta el tamaño muestral. Sin embargo, en la práctica, para fijar un valor mínimo aceptable para r supondremos que S, y por tanto su estimador, no tiene más de $\mathcal{C}$ componentes conexas. Entonces, aunque rechacemos H_0 para todos los valores de r, no consideraremos estimadores muy fragmentados y nos quedaremos con el valor mínimo de r para el cual $C_r(\mathcal{X}_n)$ no posea más de $\mathcal{C}$ componentes conexas.

Para calcular $\widehat{r}_0$ emplearemos el algoritmo de dicotomía, para el cual deberemos seleccionar el número máximo de iteraciones I y dos puntos iniciales r_m y r_M con $r_m < r_M$ tales que la hipótesis nula de r_M–convexidad se rechace y la hipótesis nula de r_m–convexidad se acepte. Además, como comentamos anteriormente, supondremos que el número de componentes conexas de $C_{r_m}(\mathcal{X}_n)$ no puede ser mayor que $\mathcal{C}$. Para que se acepte la hipótesis nula de r_m–convexidad bastaría considerar, en principio, r_m próximo a cero. Sin embargo, si no es posible porque para valores muy pequeños de r se rechaza la hipótesis de r–convexidad, estimaremos r_0 como el valor positivo más próximo a cero, r, tal que el número de componentes conexas de $C_r(\mathcal{X}_n)$ es menor o igual que $\mathcal{C}$. Por otro lado, si no se puede encontrar un spacing maximal en $H(\mathcal{X}_n)$ lo suficientemente grande para que el test resulte significativo, ya no será necesario entrar en el algoritmo y propondremos $H(\mathcal{X}_n)$ como estimador del soporte.

Por lo tanto, deberemos introducir los siguientes parámetros en el algoritmo: el nivel de significación $\alpha \in (0,1)$, un máximo de iteraciones I, un máximo de componentes conexas $\mathcal{C}$ y dos valores iniciales r_m y r_M. Dados estos parámetros, el algoritmo para calcular $\widehat{r}_0$ sería el siguiente:

1. En cada iteración y mientras el número de iteraciones sea menor que I y el número de componentes conexas de $C_{r_m}(\mathcal{X}_n)$ menor que $\mathcal{C}$:
   (I) Calcular $r = (r_m + r_M)/2$.
   (II) Si no se rechaza la hipótesis nula de r–convexidad entonces $r_m = r$.
   (III) En otro caso $r_M = r$.
2. Completado el paso anterior tenemos que $\widehat{r}_0 = r$.

A continuación, detallamos la implementación del contraste de hipótesis para la r–convexidad. En el algoritmo descrito anteriormente, necesitamos contrastar la hipótesis nula I veces y, dado que cada contraste implica calcular el spacing maximal, debemos tener en cuenta el coste computacional que ello conlleva. Sin embargo, como se recoge en Rodríguez-Casal y Saavedra-Nieves (2016), no se necesita calcular dicho spacing maximal explícitamente, basta con comprobar si existe un punto x tal que

$$\{x\} \oplus \frac{c_{n,\alpha}^{1/d}}{\widehat{f}_n^{1/d}(x)} A \subset C_r(\mathcal{X}_n) \setminus \mathcal{X}_n.$$

De existir dicho punto, tenemos que $\widehat{V}_{n,\alpha} \geq c_{n,\alpha}$ y, por tanto, se rechaza la hipótesis nula de r–convexidad. Además, nótese que si dicha bola existe entonces $x \notin B_{c_{n,\alpha}^{x,w}}(X_k)$ donde X_k es el punto de la muestra tal que $x \in Vor(X_k)$ y $c_{n,\alpha}^{x,w} = c_{n,\alpha}^{1/d} w_d^{-1/d} \widehat{f}_n^{-1/d}(x)$, siendo w_d la medida de Lebesgue de la bola unitaria d–dimensional. Por tanto, $\widehat{f}_n(x) = f_n(X_k)$.

Así, los centros de las posibles bolas maximales que pertenecen a la celda de Voronoi con núcleo X_i, $i = 1, \ldots, n$, se encuentran necesariamente en el conjunto $B_{c_{n,\alpha}^{X_i,w}}(X_i)^{\mathsf{c}} \cap Vor(X_i)$. Para calcularlo seguiremos los siguientes pasos:

1. Determinar el conjunto de puntos candidatos para los centros de las bolas

$$D(r) = \bigcup_{X_i \in E(m)} \left(\partial B_{c_{n,\alpha}^{X_i,w}}(X_i) \cap Vor(X_i) \right)$$

donde $E(m) \subset \mathcal{X}_n$ denota los extremos de la m–*shape* de $\mathcal{X}_n$ con $m = \text{mín}\left\{ c_{n,\alpha}^{X_j,w} : X_j \in \mathcal{X}_n \right\}$. La elección del conjunto $E(m)$ se debe a que este se compone por los puntos de la muestra que son m–vecinos. Dos puntos X_i y X_j se dicen que son m–vecinos si existe una bola abierta de radio m con ambos puntos en su frontera y que no contiene ningún punto de $\mathcal{X}_n$. Por tanto, si $X_i \notin E(m)$ se cumple que para todo $x \in B_{c_{n,\alpha}^{X_i,w}}(X_i)^{\mathsf{c}} \cap Vor(X_i)$, $B_{c_{n,\alpha}^{X_i,w}}(x)^{\mathsf{c}} \cap \mathcal{X}_i \neq \emptyset$. Permitiendo así, descartar estos puntos a la hora de calcular $D(r)$. El cálculo de $E(m)$ y $C_r(\mathcal{X}_n)$ se realiza mediante las funciones `ashape` y `ahull`, respectivamente, de la librería `alphahull`, Pateiro-López y Rodríguez-Casal (2010).

Para obtener el conjunto $\partial B_{c_{n,\alpha}^{X_i,w}}(X_i) \cap Vor(X_i)$ primero calcularemos un conjunto de puntos del borde de la bola $B_{c_{n,\alpha}^{X_i,w}}(X_i)$ y, a continuación, nos quedaremos solo con aquellos que pertenecen a $Vor(X_i)$. Para descartar aquellos puntos de $\partial B_{c_{n,\alpha}^{X_i,w}}(X_i)$ que no pertenecen al $Vor(X_i)$ seguiremos los siguientes pasos:

(I) Quedarnos solo con aquellos puntos que pertenecen a la envoltura convexa, $H(\mathcal{X}_n)$, recurriendo a la función `in.convex.hull` de la librería `tripack`, ver Eglen *et al.* (2016) para más información sobre la librería.

(II) Con los puntos restantes, calcular para cada uno de ellos cual es el punto de la muestra $\mathcal{X}_n$ más cercano a este considerando la distancia euclidiana. Si se trata del punto X_i lo conservamos como posible candidato, en otro caso, lo descartamos.

Para calcular el conjunto $D(r)$ deberemos repetir el procedimiento anterior para cada punto de la muestra y quedarnos con todos los candidatos de cada caso. Cabe destacar que este procedimiento solo es necesario realizarlo una vez para cada muestra $\mathcal{X}_n$, ya que este no depende de la envoltura r–convexa que estemos considerando.

2. Calculamos $M(r) = \text{máx}\{d(x, \partial C_r(\mathcal{X}_n)) : x \in C_r(\mathcal{X}_n) \cap D(r)\}$. Nótese que si $x \in C_r(\mathcal{X}_n) \cap D(r)$ entonces podemos garantizar que $B_{c_{n,\alpha}^{X_i,w}}(X_i) \cap X_i = \emptyset$, o equivalentemente,

$$\{x\} \oplus \frac{c_{n,\alpha}^{1/d}}{\widehat{f}_n^{1/d}(x)} A \subset C_r(\mathcal{X}_n) \setminus \mathcal{X}_n.$$

En primer lugar, para obtener $d(x, \partial C_r(\mathcal{X}_n))$ basta con calcular el mínimo de la distancia euclidiana de x a los centros de las bolas cuyos arcos definen $\partial C_r(\mathcal{X}_n)$, restarles los radios correspondientes a cada bola y quedarnos con el mínimo. Tanto el centro como el radio de dichas bolas se encuentran en la componente `arcs` del objeto que devuelve la función `ahull`. A continuación, podemos eliminar como posibles candidatos aquellos puntos cuyo resultado sea negativo, ya que se encuentran fuera del envoltura r–convexa. Finalmente, empleamos la función `inahull` de la librería `alphahull` con los puntos restantes para descartar aquellos que no pertenezcan a $C_r(\mathcal{X}_n)$ y calculamos $M(r)$, es decir, el máximo de las distancias de los puntos restantes al borde de la envoltura r–convexa calculadas anteriormente.

3. Si $M(r) \leq \widehat{c}_{n,\alpha}$ entonces no se rechaza la hipótesis nula de r–convexidad.

En lo que respecta al criterio de parada sobre las componentes conexas hemos tenido que implementar una función para calcular el número de componentes conexas de $C_r(\mathcal{X}_n)$. En ella, contabilizamos el número de ciclos de $C_r(\mathcal{X}_n)$ mediante la componente `arcs` del objeto `ahull`, distinguiendo los ciclos internos de los ciclos externos. Así, calculamos el número de componentes conexas de nuestro estimador como el número total de ciclos que posee menos el número de ciclos internos.

3.3. Implementación en el conjunto de datos de la velutina

Para reconstruir el EOO de la velutina en Galicia, emplearemos el algoritmo que acabamos de describir en la sección anterior. Para comenzar, consideraremos el conjunto de datos correspondientes a noviembre de 2018, constituido por 2635 coordenadas UTM de velutinas, las cuales se ilustran en la Figura 3.5. A la hora de ejecutar el algoritmo necesitamos definir los valores de los 5 parámetros de entrada: I, $\mathcal{C}$, α, r_m y r_M. Tomaremos $I = 7$, $\alpha = 0.01$, $\mathcal{C} = 7$, $r_m = 1000$ y $r_M = 4000$. Decidimos que el número máximo de componentes conexas sea 7, porque es un valor suficientemente grande y flexible para la muestra que estamos considerando. Dicha elección se justifica al observar la evolución anual de la velutina, Figura 1.2, ya que esta aparece concentrada principalmente en 2 focos en 2014 y, por tanto, cabría esperar que nuestro estimador no estuviera demasiado fragmentado.

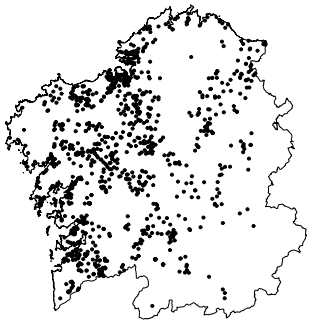
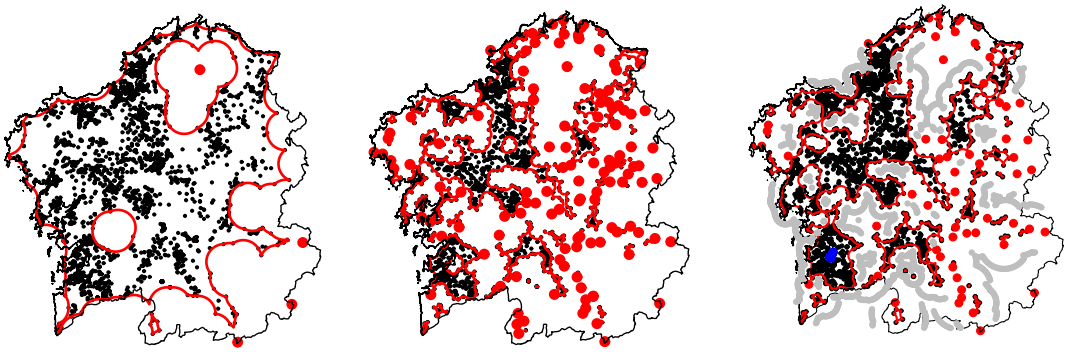


Figura 3.5: Muestra $\mathcal{X}_{2635}$ de noviembre de 2018.

La estimación obtenida para estos parámetros es $\widehat{r_0} = 11863.23$, siendo esta mayor que r_M. Por tanto, el algoritmo no ha entrado en el proceso de dicotomía ya que, para el valor de r donde $C_r(\mathcal{X}_{2635})$ tiene menos de 7 componentes conexas, se rechaza la hipótesis nula de r–convexidad. En la Figura 3.6 (izquierda) se representa $C_{11863.23}(\mathcal{X}_{2635})$. Como podemos observar, el estimador resultante está formado por 6 componentes conexas, pero posee muchas zonas sin datos.

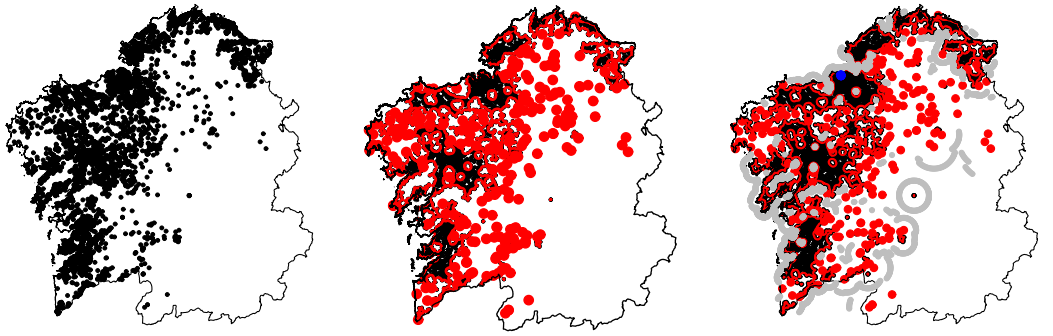
Para ver como sería un estimador sin la restricción de componentes conexas, volvemos a ejecutar el algoritmo pero, esta vez, estableciendo como criterio de parada $\mathcal{C} = 400$. En este caso, obtenemos que $\widehat{r_0} = 3473.374$ y el estimador resultante está formado por 175 componentes conexas. Como podemos ver en la Figura 3.6 (centro), $C_{3473.374}(\mathcal{X}_{2635})$ es un estimador muy fragmentando, posee bastantes puntos aislados y el número de componentes conexas es considerablemente mayor a lo esperado. Para ver qué zonas pueden estar provocando el rechazo de la hipótesis nula de r–convexidad, impidiendo que $\widehat{r_0}$ sea mayor, realizamos el contraste de hipótesis para $r = 4700$. En la Figura 3.6 (derecha) se representa $C_{4700}(\mathcal{X}_{2635})$ en rojo, los puntos grises se corresponden con los candidatos a centro de la bola maximal, que denotábamos en la sección anterior como $D(r)$. Si observamos los puntos que generan el rechazo de H_0, representados en azul, podemos apreciar que estos se ubican en áreas sin muestra rodeada de zonas con alta densidad de puntos.



$C_{11863.23}\,(\mathcal{X}_{2635})$ $C_{3473.374}\,(\mathcal{X}_{2635})$ $C_{3700}\,(\mathcal{X}_{2635})$

Figura 3.6: $C_r\,(\mathcal{X}_{2635})$ para los valores de r comentados para noviembre de 2018.

Para comprobar si el caso de noviembre de 2018 es puntual, volvemos a ejecutar el algoritmo para los datos de julio de 2017, Figura 3.7 (izquierda). Emplearemos los mismos parámetros que antes: $I = 7$, $\alpha = 0.01$, $\mathcal{C} = 400$, $r_m = 1000$ y $r_M = 4000$. En este caso obtenemos un $\widehat{r_0} = 2140.949$ y el estimador del soporte correspondiente, $C_{2140.949}\,(\mathcal{X}_{5033})$, posee 296 componentes conexas, de nuevo muchas más de las que esperábamos. En la Figura 3.7 (centro) representamos $C_{2140.949}\,(\mathcal{X}_{5033})$. Podemos ver que nuestro estimador está muy fragmentado, incluso más que en el caso de noviembre de 2018, y además, posee demasiados puntos aislados. Entonces, al igual que antes, vamos a ver qué zona está provocando el rechazo de la hipótesis de r-convexidad si aumentamos el valor de r. Para ello, tomamos $r = 2200$ y aplicamos el contraste de 2200–convexidad, en la Figura 3.7 (derecha) se representa $C_{2200}\,(\mathcal{X}_{5033})$. De nuevo, si observamos los puntos que generan el rechazo de la hipótesis nula, representados en azul, estos se ubican en áreas sin muestra rodeada de zonas con alta densidad de puntos, en este caso en la zona de Oleiros.



$\mathcal{X}_{5033}$ julio de 2017 $C_{2140.949}\,(\mathcal{X}_{5033})$ $C_{2200}\,(\mathcal{X}_{5033})$

Figura 3.7: $C_r\,(\mathcal{X}_{2635})$ para los valores de r comentados para julio de 2017.

Estos dos casos nos llevan a conjeturar que nuestro contraste es sensible cuando existe cercanía entre zonas de alta y baja densidad de datos. Esto se puede ver claramente en los puntos que provocan el rechazo de la hipótesis nula en ambos casos. Por otro lado, este comportamiento también se puede deber a un calibrado deficiente del spacing maximal, el cual emplea el contraste de hipótesis basado en el Teorema 3.4.

Para estudiar si el calibrado es correcto o no, realizaremos un estudio de simulación en el siguiente capítulo. Para ello, calcularemos el porcentaje de veces que rechazamos H_0, es decir, que el spacing maximal es mayor que el punto crítico $c_{n,\alpha}$ definido en el Teorema 3.4. El Capítulo 5 se centrará en analizar con más detalle el comportamiento del estimador sobre nuestro conjunto de datos, aplicando las conclusiones obtenidas en el estudio de simulación.

Capítulo 4

Estudio de simulación

Como vimos en el capítulo anterior, el estimador propuesto por Rodríguez-Casal y Saavedra-Nieves (2019) no se comporta como cabía esperar al estimar el soporte de nuestro conjunto de datos. Por un lado, sospechamos que dicho estimador no se ajusta bien al soporte cuando existen zonas sin datos, rodeadas de otras con una alta densidad. Además, creemos que el calibrado del spacing maximal propuesto, basado en el artículo de Aaron *et al.* (2017), es deficiente. Esto puede deberse a que, al ser un calibrado asintótico, se necesite de un tamaño de muestra demasiado grande para ser válido. Para confirmar si dicha suposición es cierta, realizaremos varias pruebas de simulación a lo largo de este capítulo. En cada simulación, emplearemos el algoritmo que presentamos en el Capítulo 3 para chequear si, bajo H_0, el porcentaje de rechazos se aproxima al nivel de significación. Para la comprobación, calcularemos la proporción de veces que el spacing maximal es mayor que el punto crítico $c_{n,\alpha}$. Si dicha proporción se aproxima al nivel de significación fijado α, entonces, podemos concluir que el calibrado es correcto.

Para las distintas simulaciones emplearemos como modelos densidades que son mixturas de normales bivariantes, las cuales se extrajeron del artículo de Wand and Jones (1995). La elección de estas densidades se debe a que creemos que podrían servir como modelos para nuestros datos, pues como vimos en la Figura 1.2, la velutina aparece principalmente en dos focos. Además, estos modelos nos permiten estudiar diferentes escenarios, donde las modas se encuentran más o menos separadas y existe diferente grado de dispersión en los datos. En el Cuadro 4.1 se recogen los parámetros que definen cada uno de los modelos: w_i denota el peso de cada normal bivariante en la mixtura, μ_{ij} la componente j del vector de medias, σ_{ij}^2 la varianza de la componente j y ρ_i el coeficiente de correlación. Los contornos de probabilidad de cada densidad se representan en la Figura 4.1. Por otro lado, los parámetros que estableceremos en cada simulación son los siguientes: N denota el número de muestras generadas, n el tamaño muestral, r el parámetro de forma de la envoltura r–convexa y α el nivel de significación. En cada simulación calcularemos el porcentaje de rechazos de H_0 y comprobaremos si este se aproxima al nivel de significación α.

Densidad	$\boldsymbol{w_1\ N\left(\mu_{11},\mu_{12},\sigma^2_{11},\sigma^2_{12},\rho_1\right)+w_2\,N\left(\mu_{21},\mu_{22},\sigma^2_{21},\sigma^2_{22},\rho_2\right)}$
Bimodal I	$\frac{1}{2}N\left(-1,0,\frac{4}{9},\frac{4}{9},0\right)+\frac{1}{2}\,N\left(1,0,\frac{4}{9},\frac{4}{9},0\right)$
Bimodal II	$\frac{1}{2}N\left(-\frac{3}{2},0,\frac{1}{16},1,0\right)+\frac{1}{2}\,N\left(\frac{3}{2},0,\frac{1}{16},1,0\right)$
Bimodal III	$\frac{1}{2}N\left(-1,1,\frac{4}{9},\frac{4}{9},\frac{3}{5}\right)+\frac{1}{2}\,N\left(1,-1,\frac{4}{9},\frac{4}{9},\frac{3}{5}\right)$
Bimodal IV	$\frac{1}{2}N\left(1,-1,\frac{4}{9},\frac{4}{9},\frac{7}{10}\right)+\frac{1}{2}\,N\left(-1,1,\frac{4}{9},\frac{4}{9},0\right)$

Cuadro 4.1: Parámetros de los modelos de mixtura de normales.

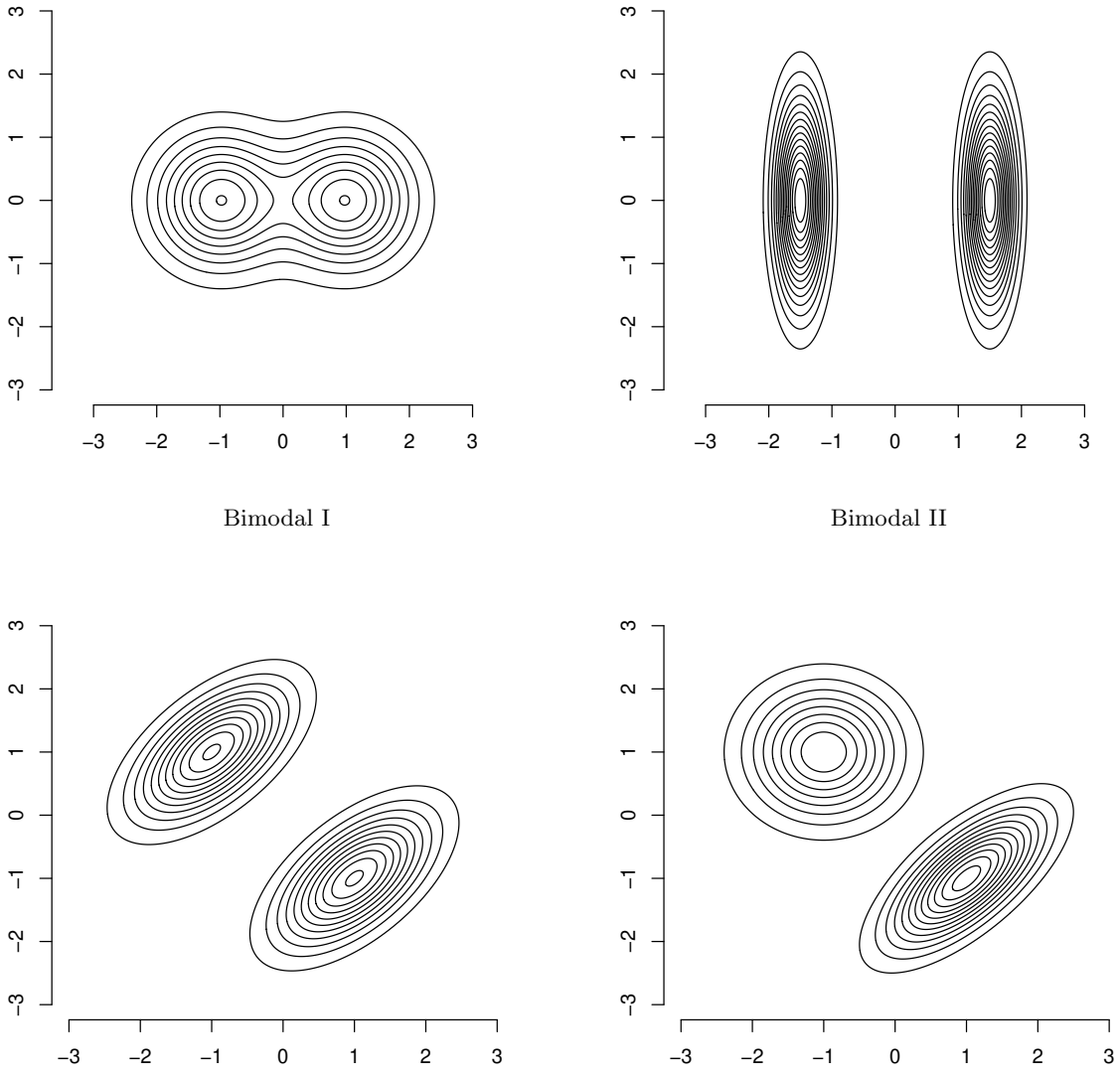


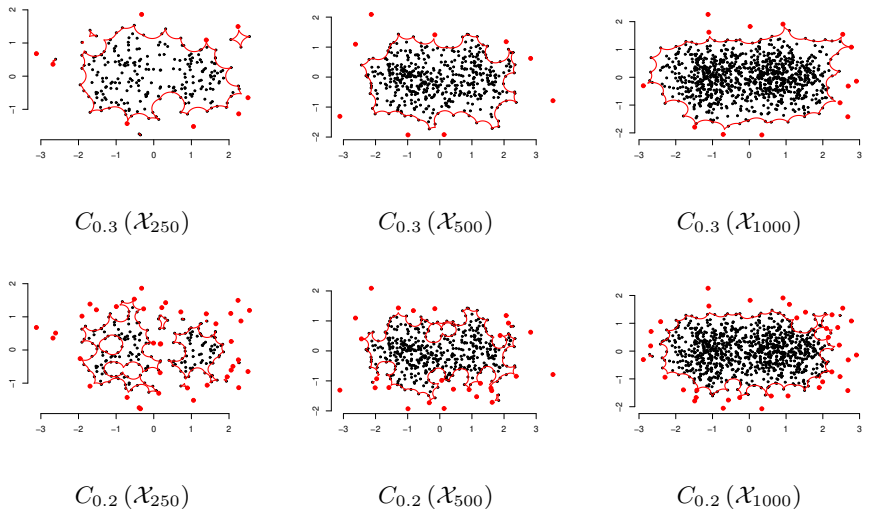
Figura 4.1: Contornos de probabilidad para los distintos modelos bimodales.

A continuación, recogemos los resultados para cada uno de los modelos considerados. Para cada valor del parámetro de forma $r = 0.3$ y 0.2 se han considerado niveles de significación $\alpha = 0.1$, 0.05 y 0.01 con muestras de tamaño $n = 250$, 500 y 1000 y simulaciones de $N = 250$ y 500 casos. En primer lugar, mostramos las proporciones de rechazo obtenidas para cada modelo en los Cuadros 4.2, 4.3, 4.4 y 4.5. A continuación, en las Figuras 4.2, 4.3, 4.4 y 4.5 se representan las envolturas r–convexas de una muestra de 250 datos, otra de 500 y una última de 1000 para los distintos valores de r considerados en las simulaciones.

Simulaciones modelo bimodal I

N	n	$r = 0.3$			$r = 0.2$		
		$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
250	250	0.008	0	0	0	0	0
	500	0.156	0.084	0.02	0	0	0
	1000	0.42	0.308	0.128	0.092	0.044	0.008
500	250	0.016	0	0	0	0	0
	500	0.19	0.088	0.022	0	0	0
	1000	0.42	0.296	0.094	0.096	0.036	0.004

Cuadro 4.2: Resultados de las simulaciones para la bimodal I.



$C_{0.3}(\mathcal{X}_{250})$ $C_{0.3}(\mathcal{X}_{500})$ $C_{0.3}(\mathcal{X}_{1000})$

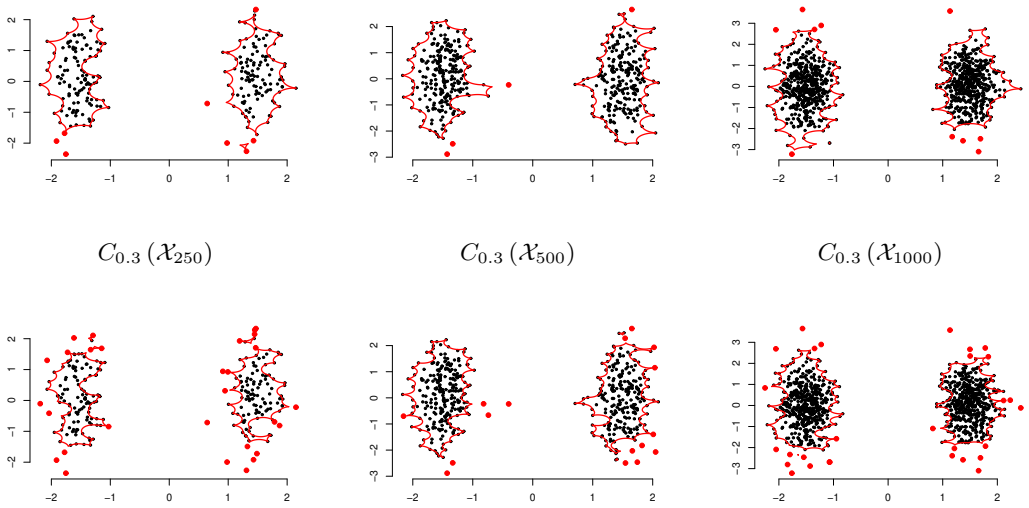
$C_{0.2}(\mathcal{X}_{250})$ $C_{0.2}(\mathcal{X}_{500})$ $C_{0.2}(\mathcal{X}_{1000})$

Figura 4.2: Envolturas r–convexas de las muestras $\mathcal{X}_{250}$, $\mathcal{X}_{500}$ y $\mathcal{X}_{1000}$ de la bimodal I.

Simulaciones modelo bimodal II

N	n	$r = 0.3$			$r = 0.2$		
		$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
250	250	0	0.004	0	0	0	0
	500	0.12	0.072	0.016	0.032	0.02	0
	1000	0.472	0.332	0.164	0.224	0.148	0.052
500	250	0.004	0	0	0	0	0
	500	0.11	0.068	0.014	0.014	0.002	0.002
	1000	0.462	0.318	0.146	0.242	0.16	0.044

Cuadro 4.3: Resultados de las simulaciones para la bimodal II.



$C_{0.3}(\mathcal{X}_{250})$ $C_{0.3}(\mathcal{X}_{500})$ $C_{0.3}(\mathcal{X}_{1000})$

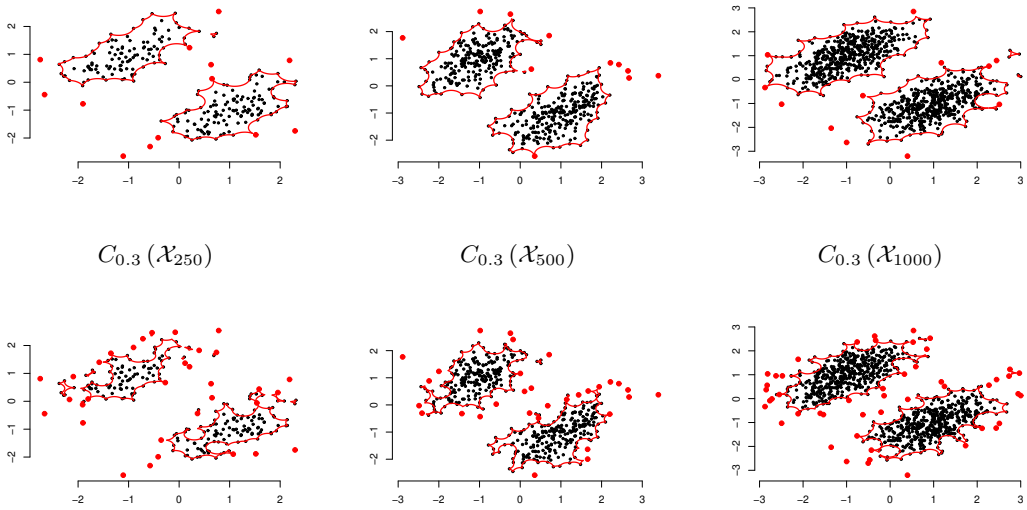
$C_{0.2}(\mathcal{X}_{250})$ $C_{0.2}(\mathcal{X}_{500})$ $C_{0.2}(\mathcal{X}_{1000})$

Figura 4.3: Envolturas r–convexas de las muestras $\mathcal{X}_{250}$, $\mathcal{X}_{500}$ y $\mathcal{X}_{1000}$ de la bimodal II.

Simulaciones modelo bimodal III

N	n	$r = 0.3$			$r = 0.2$		
		$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
250	250	0.008	0	0	0	0	0
	500	0.204	0.088	0.032	0.004	0	0
	1000	0.5	0.288	0.176	0.124	0.056	0.02
500	250	0.004	0	0	0	0	0
	500	0.154	0.1	0.016	0.004	0	0
	1000	0.496	0.364	0.14	0.11	0.07	0.018

Cuadro 4.4: Resultados de las simulaciones para la bimodal III.



$C_{0.3}(\mathcal{X}_{250})$ $C_{0.3}(\mathcal{X}_{500})$ $C_{0.3}(\mathcal{X}_{1000})$

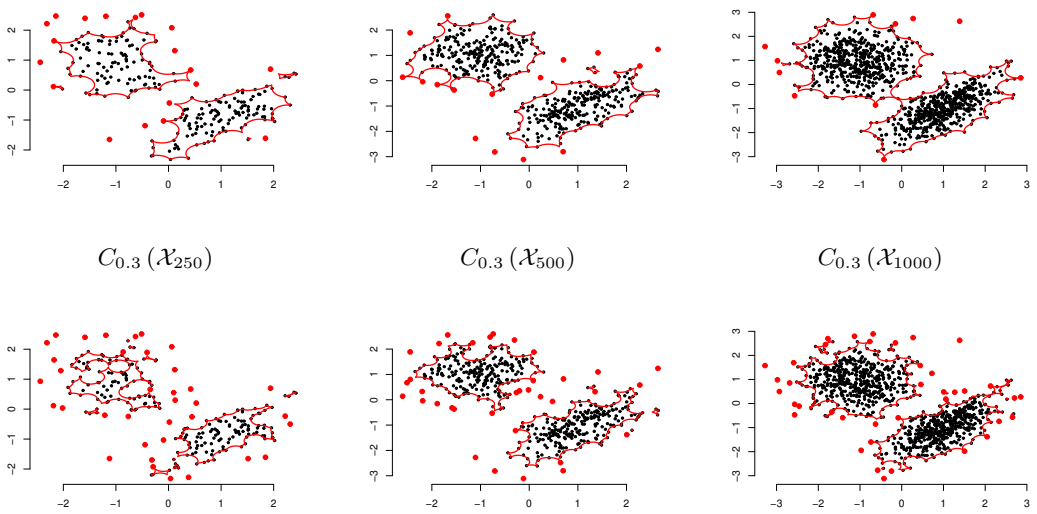
$C_{0.2}(\mathcal{X}_{250})$ $C_{0.2}(\mathcal{X}_{500})$ $C_{0.2}(\mathcal{X}_{1000})$

Figura 4.4: Envolturas r–convexas de las muestras $\mathcal{X}_{250}$, $\mathcal{X}_{500}$ y $\mathcal{X}_{1000}$ de la bimodal III.

Simulaciones modelo bimodal IV

N	n	$r = 0.3$			$r = 0.2$		
		$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
250	250	0.008	0.008	0	0	0	0
	500	0.152	0.096	0.016	0.004	0	0
	1000	0.576	0.424	0.232	0.140	0.072	0.016
500	250	0.002	0.006	0	0	0	0
	500	0.14	0.106	0.02	0	0	0
	1000	0.630	0.456	0.194	0.136	0.064	0.016

Cuadro 4.5: Resultados de las simulaciones para la bimodal IV.



$C_{0.3}(\mathcal{X}_{250})$ $C_{0.3}(\mathcal{X}_{500})$ $C_{0.3}(\mathcal{X}_{1000})$

$C_{0.2}(\mathcal{X}_{250})$ $C_{0.2}(\mathcal{X}_{500})$ $C_{0.2}(\mathcal{X}_{1000})$

Figura 4.5: Envolturas r–convexas de las muestras $\mathcal{X}_{250}$, $\mathcal{X}_{500}$ y $\mathcal{X}_{1000}$ de la bimodal IV.

A la vista de los resultados de los Cuadros 4.2, 4.3, 4.4 y 4.5, parece que el calibrado es deficiente. Esto resulta evidente en todos los modelos cuando el parámetro de forma r toma el valor de 0.3, siendo muy notable para las muestras de tamaño 1000 donde obtenemos proporciones tres o cuatro veces mayores al nivel de significación. Para $r = 0.2$, los resultados son mejores para $n = 1000$ en los modelos bimodales I, III y IV pues, los porcentajes obtenidos en el modelo bimodal II son mayores que dos veces el nivel de significación. Esta diferencia se puede deber a que en el modelo bimodal II,

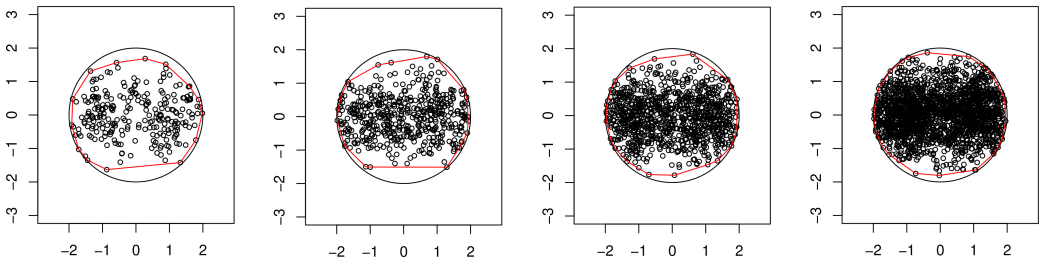
los datos se encuentra más concentrados entorno a las modas, esto provoca que la densidad se acerque más a zero en las zonas ubicadas entre las dos modas que en el resto de los modelos. En general, al igual que en el caso de la velutina, parece que las zonas con baja densidad de la muestra, en contraste con las zonas de alta, pueden estar generando una deficiencia en el calibrado.

4.1. Estudio de simulación con muestra restringida

Dados los resultados obtenidos hasta ahora, todo parece indicar que el comportamiento en la práctica del calibrado del spacing maximal es deficiente. Esto resulta más notable para valores grandes de r donde el estimador no se fragmenta tanto. Además, creemos que las zonas de elevada concentración de datos pueden estar causando esta deficiencia en el calibrado cuando existen otras de baja concentración.

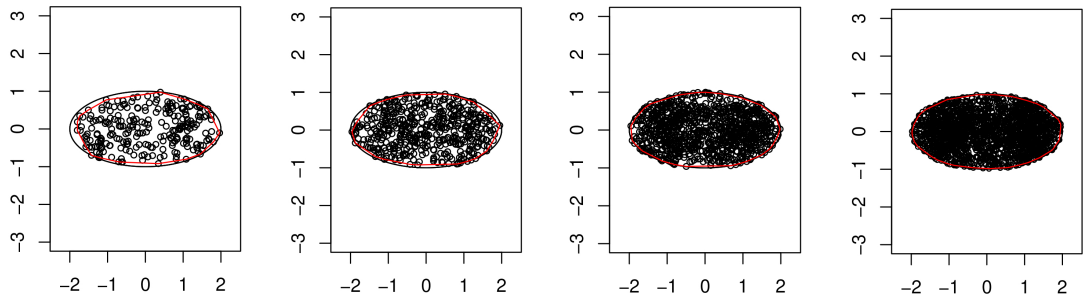
El calibrado que estamos comprobando es el que emplean Aaron *et al.* (2017) en el Teorema 4 para el caso de soporte desconocido y convexo. Entonces, para simplificar las simulaciones y evitar la influencia del parámetro r en el estimador, podemos emplear la envoltura convexa a la hora de chequear el calibrado. Para testar el calibrado del spacing maximal en este caso, realizaremos el mismo procedimiento de las simulaciones anteriores pero considerando como estimador la envoltura convexa. En las simulaciones, emplearemos el modelo bimodal I y tomaremos como S un conjunto convexo donde se cumplan las condiciones del Teorema 4.

Dados los contornos de dicha densidad, representados en la Figura 4.1, consideraremos dos escenarios distintos para S. En el primero de ellos, tomaremos como S_1 la bola de centro cero y radio 2. En el segundo, emplearemos como soporte S_2 la elipse $x^2/4 + y^2/1 = 1$, la cual tiene como eje mayor el eje OX y eje menor el eje OY, donde el eje mayor va de -2 a 2 y el eje menor de -1 a 1. En la Figura 4.6 se ilustran las envolturas convexas para los distintos tamaños muestrales considerados en S_1 y en la Figura 4.7 para S_2. Podemos observar que en el caso de la bola $B_2(0)$ las zonas de elevada concentración de datos son más acusadas respecto de otras de baja densidad mientras que, en el escenario de la elipse, el contraste entre estas zonas resulta más suave. Por tanto, la diferencia entre los resultados de estos dos casos nos permitirá chequear si la violación de la hipótesis de que la densidad es mayor o igual que f_0 afecta al calibrado del spacing.



$H(\mathcal{X}_{250})$ $H(\mathcal{X}_{500})$ $H(\mathcal{X}_{1000})$ $H(\mathcal{X}_{2000})$

Figura 4.6: Envolturas convexas de las muestras restringidas a la bola $B_2(0)$ de tamaño 250, 500, 1000 y 2000.



$H(\mathcal{X}_{250})$ $H(\mathcal{X}_{500})$ $H(\mathcal{X}_{1000})$ $H(\mathcal{X}_{2000})$

Figura 4.7: Envolturas convexas de las muestras restringidas a la elipse $x^2/4 + y^2/1 = 1$ de tamaño 250, 500, 1000 y 2000.

Para cada simulación generaremos muestras a partir de la distribución bimodal I de tamaños $n = 250, 500, 1000$ y 2000, restringiendo cada muestra al soporte correspondiente en cada caso. En el Cuadro 4.6 se recogen los resultados obtenidos para cada uno de los tamaños muestrales cuando consideramos como soporte S_1 y en el Cuadro 4.7 para el caso de S_2. En ambos casos, trabajaremos con niveles de significación del 50 %, 40 %, 30 %, 20 %, 10 %, 5 % y 1 %.

Simulaciones modelo bimodal I

N	n	$\alpha = 0.5$	$\alpha = 0.4$	$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
500	250	0.858	0.788	0.732	0.614	0.498	0.36	0.166
	500	0.886	0.844	0.792	0.668	0.528	0.362	0.216
	1000	0.892	0.852	0.76	0.716	0.51	0.396	0.202
	2000	0.862	0.81	0.724	0.642	0.468	0.308	0.19

Cuadro 4.6: Resultados de las simulaciones para la envoltura convexa en el caso de la densidad bimodal I restringida a la bola $B_2(0)$.

A la vista de los resultados del Cuadro 4.6, el nivel de significación no se aproxima bien cuando consideramos S_1, pues obtenemos proporciones de rechazo significativamente mayores al nivel de significación fijado α. Sin embargo, en el Cuadro 4.7 podemos ver que las proporciones de rechazo obtenidas se aproximan bastante a los niveles de significación fijados para el caso de S_2. Unos resultados completamente distintos a los que obteníamos para la bola $B_2(0)$. En S_2 la convergencia resulta bastante rápida y las oscilaciones en la segunda cifra decimal se deben a la variabilidad propia del proceso de simulación. Por tanto, resulta evidente que la elección del conjunto S influye en los resultados. En el caso de la bola, los datos no cubrían bien a S_1 y en las zonas donde la densidad f es muy baja no existían datos para el tamaño muestral fijado, provocando el rechazo del contraste de hipótesis en estas zonas. Sin embargo, al trabajar con la elipse, podemos ver que las datos recubren prácticamente por completo S_2 evitando así las zonas que provocaban el rechazo en el caso de la bola.

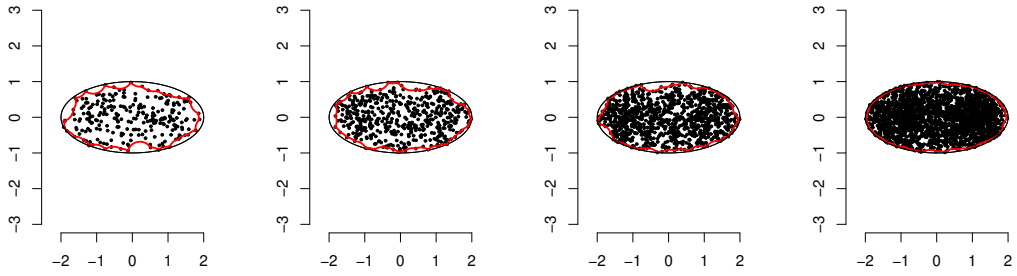
Simulaciones modelo bimodal I

N	n	$\alpha = 0.5$	$\alpha = 0.4$	$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
500	250	0.504	0.434	0.348	0.246	0.128	0.074	0.026
	500	0.482	0.462	0.346	0.242	0.098	0.066	0.024
	1000	0.506	0.48	0.3	0.216	0.12	0.074	0.01
	2000	0.482	0.348	0.31	0.186	0.108	0.056	0.014

Cuadro 4.7: Resultados de las simulaciones para la envoltura convexa en el caso de la densidad bimodal I restringida a la elipse $x^2/4 + y^2/1 = 1$.

4.1.1. Simulación de la envoltura r-convexa en el caso de la elipse

El siguiente paso es comprobar si el comportamiento obtenido cuando consideramos el soporte S_2 se mantiene si empleamos como estimador la envoltura r–convexa para distintos valores de r. Así, comprobaríamos el Teorema 3.1 de Rodríguez-Casal y Saavedra-Nieves (2019) para la estimación de la EOO. Para cada simulación consideraremos muestras de tamaño $n = 250, 500, 1000$ y 2000 generadas a partir de la distribución bimodal I y restringidas a la elipse descrita en el caso anterior. En cada simulación tomaremos $r = 0.55$, 0.5, 0.45, 0.4 y 0.35 y niveles de significación del 50 %, 40 %, 30 %, 20 %, 10 %, 5 % y 1 %. En el Cuadro 4.8 se recogen los resultados obtenidos y en las Figuras 4.8, 4.9 y 4.10 se ilustran las envolturas r–convexas para $r = 0.35$, 0.45 y 0.55 y los distintos tamaños muestrales considerados.



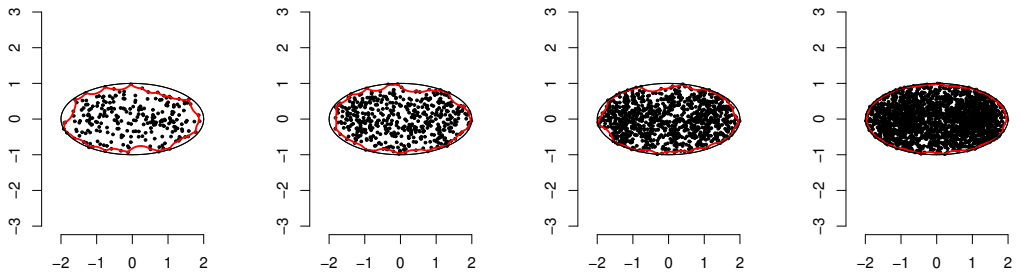
$C_{0.35}(\mathcal{X}_{250})$ $C_{0.35}(\mathcal{X}_{500})$ $C_{0.35}(\mathcal{X}_{1000})$ $C_{0.35}(\mathcal{X}_{2000})$

Figura 4.8: Envolturas 0.35–convexas de las muestras de tamaño 250, 500, 1000 y 2000.

Simulaciones modelo bimodal I

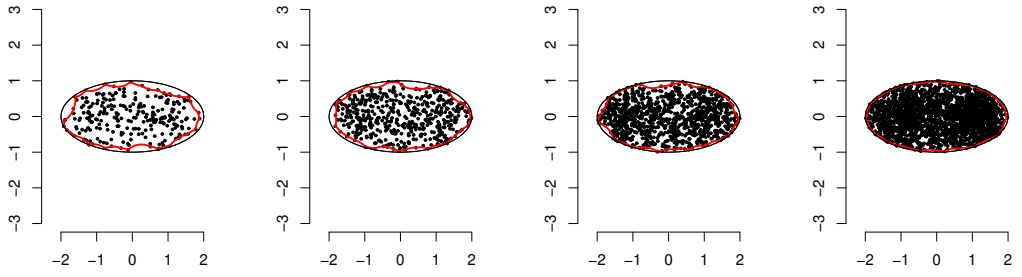
N	r	n	$\alpha = 0.5$	$\alpha = 0.4$	$\alpha = 0.3$	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
500	0.55	250	0.48	0.364	0.278	0.186	0.116	0.056	0.012
		500	0.516	0.368	0.318	0.244	0.124	0.062	0.008
		1000	0.464	0.372	0.308	0.23	0.106	0.07	0.006
		2000	0.43	0.348	0.292	0.162	0.104	0.054	0.006
	0.5	250	0.468	0.312	0.276	0.172	0.102	0.052	0.008
		500	0.418	0.362	0.31	0.218	0.09	0.084	0.016
		1000	0.456	0.352	0.282	0.2	0.102	0.054	0.01
		2000	0.438	0.37	0.296	0.208	0.13	0.066	0.01
	0.45	250	0.428	0.346	0.236	0.198	0.108	0.064	0.03
		500	0.426	0.364	0.322	0.218	0.124	0.08	0.012
		1000	0.458	0.38	0.31	0.202	0.122	0.056	0.006
		2000	0.474	0.418	0.258	0.182	0.098	0.042	0.014
	0.4	250	0.424	0.332	0.23	0.142	0.092	0.046	0.012
		500	0.492	0.33	0.296	0.196	0.104	0.054	0.02
		1000	0.5	0.386	0.272	0.18	0.098	0.054	0.012
		2000	0.496	0.378	0.278	0.192	0.118	0.04	0.012
	0.35	250	0.42	0.322	0.262	0.174	0.086	0.054	0.014
		500	0.446	0.368	0.3	0.2	0.116	0.052	0.008
		1000	0.458	0.386	0.29	0.182	0.14	0.052	0.016
		2000	0.438	0.35	0.284	0.164	0.09	0.042	0.018

Cuadro 4.8: Resultados de las simulaciones para la envoltura r–convexa en el caso de la densidad bimodal I restringida a la elipse $x^2/4 + y^2/1 = 1$.



$C_{0.45}(\mathcal{X}_{250})$ $C_{0.45}(\mathcal{X}_{500})$ $C_{0.45}(\mathcal{X}_{1000})$ $C_{0.45}(\mathcal{X}_{2000})$

Figura 4.9: Envolturas 0.45–convexas de las muestras de tamaño 250, 500, 1000 y 2000.



$C_{0.55}(\mathcal{X}_{250})$ $C_{0.55}(\mathcal{X}_{500})$ $C_{0.55}(\mathcal{X}_{1000})$ $C_{0.55}(\mathcal{X}_{2000})$

Figura 4.10: Envolturas 0.55–convexas de las muestras de tamaño 250, 500, 1000 y 2000.

Como cabía esperar, los resultados apenas han variado al pasar de la envoltura convexa a la envoltura r-convexa pues, como veíamos anteriormente, esta generaliza a la envoltura convexa. Recordemos que la envoltura convexa podía pensarse como una envoltura r-convexa cuando r tiende a infinito. A la vista de los resultados obtenidos en este estudio de simulación, parece que el método que empleamos para estimar el soporte sí se ve afectado por las zonas con poca densidad de datos. Por tanto, podemos concluir que el método propuesto por Rodríguez-Casal y Saavedra-Nieves (2019) calibra bien si en el conjunto a estimar no existen zonas sin datos y, en ese caso, no hacen falta tamaños muestrales muy grandes para conseguir una estimación correcta.

Capítulo 5

Análisis de los datos

En las simulaciones del Capítulo 4, hemos detectado que las zonas con baja densidad de datos provocan que el spacing maximal no esté bien calibrado. La existencia de dichas zonas pueden ser la razón por la que el estimador del soporte de la velutina, que calculábamos en la Sección 3.3, se fragmentara demasiado. Por tanto, todo parece indicar que a la hora de tratar de estimar la EOO en nuestro conjunto de datos de la velutina debemos de probar primero a recortar la muestra, centrándonos en eliminar las zonas con baja densidad de datos. Dicha limpieza de los datos la realizaremos mediante la muestra efectiva. Formalmente, la muestra efectiva de nivel τ con $\tau \in (0, 1)$, se define como el conjunto

$$L(\tau) = \left\{x \in \mathbb{R}^d \,:\, f(x) \geq f_\tau\right\}$$

donde

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

La forma en que se define este conjunto nos garantiza que su contenido en probabilidad es mayor o igual a $1-\tau$. A lo largo del capítulo, nos referiremos a la muestra efectiva de nivel τ como muestra efectiva del $(1-\tau)\times 100$ por ciento. Por ejemplo, para $\tau = 0.05$ hablaremos de la muestra efectiva del 95 %.

En primer lugar, en la Sección 5.1, analizaremos los resultados obtenidos tras aplicar la muestra efectiva, para distintos valores del nivel τ, a los datos de Galicia en noviembre de 2018, Figura 3.3. A continuación, estudiaremos las estimaciones resultantes de restringir la muestra a distintas regiones de Galicia en la Sección 5.2.

5.1. Estimación en noviembre de 2018

Para comenzar realizaremos pruebas con distintas proporciones de muestra para los datos de noviembre de 2018, Figura 5.1. Recordemos que estos nos daban un resultado poco satisfactorio en el Capítulo 3. Nos restringiremos a los datos de los 15 primeros días y estimaremos r_0 para las muestras efectivas correspondientes a los valores de $\tau = 0.4$, 0.3, 0.2, 0.1 y 0.05 finalizando con la estimación en la muestra completa. En cada caso hemos calculado $\widehat{r}_0$ para unos niveles de significación del 10 %, 5 % y 1 % y el número de componentes conexas, que denotaremos por *ncc*. En el Cuadro 5.1 se recogen los resultados obtenidos. A continuación, se representan los estimadores del soporte obtenidos para las muestras efectivas del 60 %, 80 % y 95 % en las Figuras 5.2, 5.3 y 5.4, respectivamente.

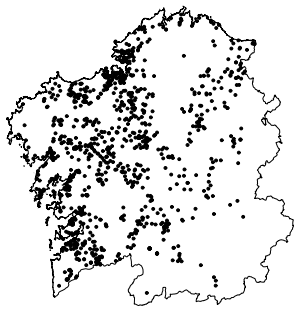
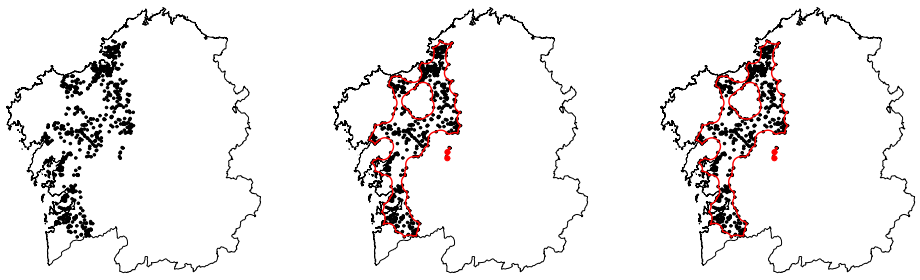


Figura 5.1: Muestra de velutina en los 15 primeros días de noviembre de 2018.

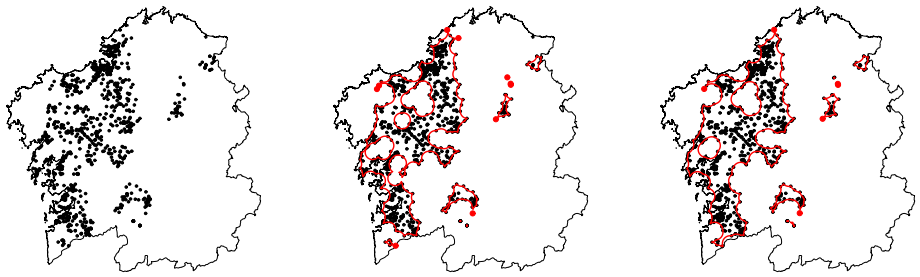
		Nivel de significación		
		10 %	5 %	1 %
Muestra 60 %	$\widehat{r}_0$	6500	6611.816	6611.816
	ncc	4	4	4
Muestra 70 %	$\widehat{r}_0$	5946.28	6105.957	6354.98
	ncc	7	7	6
Muestra 80 %	$\widehat{r}_0$	5964.844	5964.844	6354.98
	ncc	17	17	15
Muestra 90 %	$\widehat{r}_0$	5964.844	6247.559	6468.75
	ncc	17	14	12
Muestra 95 %	$\widehat{r}_0$	6000.977	6334.961	6468.75
	ncc	30	28	21
Muestra completa	$\widehat{r}_0$	6287.109	6499.512	7631.836
	ncc	52	44	28

Cuadro 5.1: Resultados para las estimaciones en los 15 primeros días de noviembre de 2018.



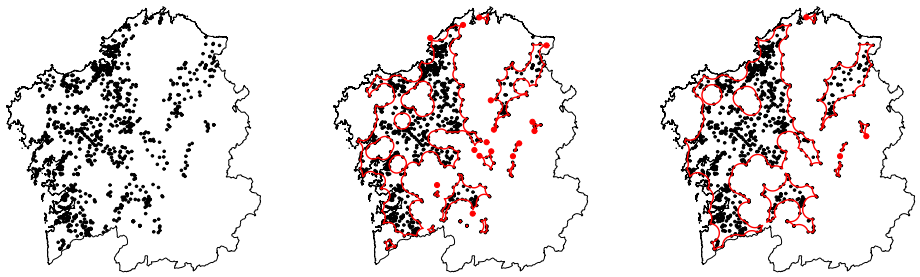
$\mathcal{X}_{669}$ $C_{6500}\left(\mathcal{X}_{669}\right)$ $C_{6611.816}\left(\mathcal{X}_{669}\right)$

Figura 5.2: Estimaciones para la muestra efectiva del 60 % de noviembre de 2018.



$\mathcal{X}_{892}$ $C_{5964.844}\left(\mathcal{X}_{892}\right)$ $C_{6505.859}\left(\mathcal{X}_{892}\right)$

Figura 5.3: Estimaciones para la muestra efectiva del 80 % de noviembre de 2018.



$\mathcal{X}_{1059}$ $C_{6000.977}\left(\mathcal{X}_{1059}\right)$ $C_{7772.949}\left(\mathcal{X}_{1059}\right)$

Figura 5.4: Estimaciones para la muestra efectiva del 95 % de noviembre de 2018.

Como podemos observar en el Cuadro 5.1 y las Figuras 5.2, 5.3 y 5.4, las estimaciones resultantes para las muestras efectivas del 60 %, 70 %, 80 % y 90 % son mejores y no tienden a fragmentarse tanto como en las pruebas del Capítulo 3. Los resultados obtenidos concuerdan con las conclusiones del Capítulo 4, eliminar datos en zonas de baja densidad mejora sustancialmente las estimaciones. Además, podemos ver que, cuando nos acercamos a la muestra completa estas siguen presentando una tendencia a ser un poco irregulares. De hecho, en el Cuadro 5.1 podemos ver como el número de componentes conexas del estimador prácticamente se duplica cuando pasamos de emplear la muestra efectiva del 90 % a la del 95 %.

5.2. Estimación en regiones

A la vista de la concentración de datos en determinadas zonas de la provincia de Pontevedra, una alternativa para evitar borrar datos, es centrarse en regiones de alta densidad de datos, como son las Rías Baixas. Para ello, vamos a estimar el soporte de la velutina en dos regiones distintas: la zona de Vigo y la ría de Muros y Noia.

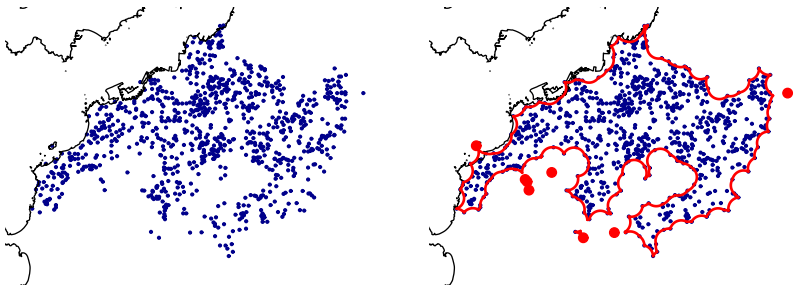
Regiones de Vigo y Mos

En primer lugar, estimaremos r_0 para las regiones de Vigo y Mos en los años 2017 y 2018. Estas regiones son vecinas, por tanto, esperamos que nuestro estimador del soporte se componga de una componente conexa. En el Cuadro 5.2 se recogen los resultados para cada año, en concreto, mostramos la estimación de r_0 y el número de componentes conexas de $C_{\widehat{r}_0}$, denotado por *ncc*.

		Nivel de significación		
		10 %	5 %	1 %
Año 2017	$\widehat{\boldsymbol{r}}_{\boldsymbol{0}}$	678.9062	678.9062	678.9062
	$\boldsymbol{ncc}$	10	10	10
Año 2018	$\widehat{\boldsymbol{r}}_{\boldsymbol{0}}$	749.1211	749.1211	778.418
	$\boldsymbol{ncc}$	6	6	6

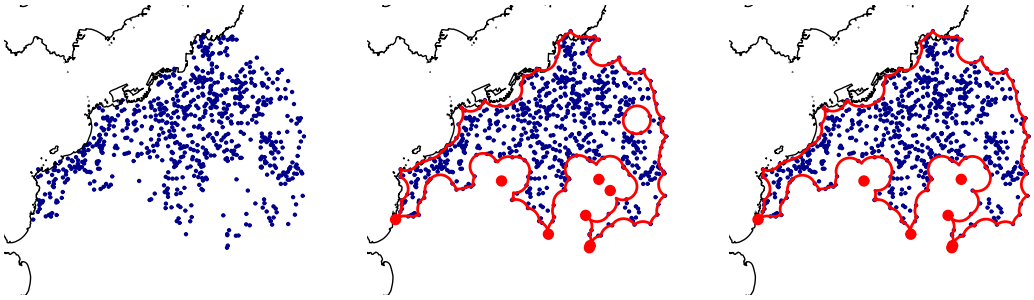
Cuadro 5.2: Estimaciones para las regiones de Vigo y Mos.

Como podemos observar en las Figuras 5.5 y 5.6, hemos obtenido unas estimaciones bastante satisfactorias, pues estas se llegan a parecerse a la forma de las regiones consideradas. Por tanto, en este caso no sería necesario recurrir a la muestra efectiva para realizar una limpieza de los datos ya que el muestreo en esta zona ha resultado bastante uniforme y no se observan esos contrastes de densidad entre distintas zonas.



$\mathcal{X}_{945}$ $C_{678.9062}(\mathcal{X}_{945})$

Figura 5.5: Estimación para las regiones de Vigo y Mos en 2017.



$\mathcal{X}_{757}$ $C_{749.1211}(\mathcal{X}_{757})$ $C_{778.418}(\mathcal{X}_{757})$

Figura 5.6: Estimaciones para las regiones de Vigo y Mos en 2018.

Regiones de Muros y Porto do Son

Para terminar, calcularemos $\widehat{r}_0$ en las regiones de Muros y Porto do Son. Dado que el volumen de datos en un único año no es considerablemente grande, vamos a agregar varios años, en este caso 2017, 2018 y 2019. Tras realizar dicha agregación conseguimos un total de 1063 coordenadas de nidos de velutina. Estas dos regiones se caracterizan por estar enfrentados en la ría y, por lo tanto, esperamos que nuestro estimador del soporte se componga al menos de dos componentes conexas. En el Cuadro 5.3 se recogen los resultados para, a continuación, representar los estimadores del soporte obtenidos en la Figura 5.7.

	Nivel de significación		
	10 %	5 %	1 %
$\widehat{\boldsymbol{r}}_{\mathbf{0}}$	484.082	484.082	612.9883
$\boldsymbol{ncc}$	53	53	36

Cuadro 5.3: Estimaciones para las regiones de Muros y Porto do Son.

Sin duda, y como era de esperar, la estimación menos fragmentada la obtenemos con el nivel de significación del 1 %. Aún así, el estimador se fragmenta demasiado. Esto se debe a que los datos no están muy concentrados y existen varias regiones sin muestra cerca de otras con más concentración. Además, la muestra se encuentra muy pegada a la costa y la envoltura r–convexa en franjas estrechas no es muy eficiente con tamaños muestrales moderados como es el caso. Entonces, a la vista de las conclusiones que obtuvimos en el Capítulo 4 y los resultados cuando estimamos el soporte en Galicia, vamos a limpiar la muestra estimando el soporte efectivo. Estimaremos r_0 para las muestras efectivas del 60 %, 70 %, 80 %, 90 % y del 95 %. Tomaremos unos niveles de significación del 10 %, 5 % y 1 % y calcularemos el número de componentes conexas, que denotaremos por *ncc*. Los resultados obtenidos se recogen en el Cuadro 5.4. La representación de los estimadores obtenidos para las muestras del 60 %, 80 % y 95 % se muestran en las Figuras 5.8, 5.9 y 5.10, respectivamente.

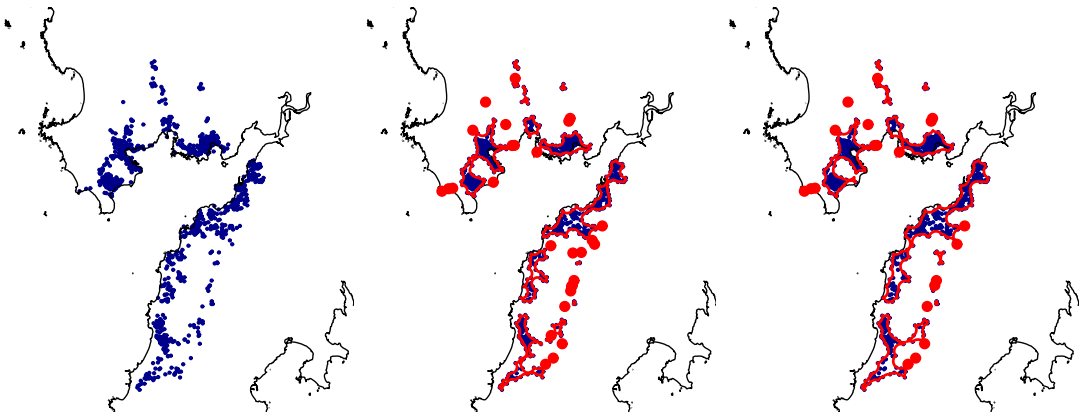


Figura 5.7: Estimaciones para las regiones de Muros y Porto do Son del 2017 al 2019.

		Nivel de significación		
		10 %	5 %	1 %
Muestra 60 %	$\widehat{r}_0$	755.5664	755.5664	755.5664
	ncc	5	5	5
Muestra 70 %	$\widehat{r}_0$	755.5664	755.5664	755.5664
	ncc	5	5	5
Muestra 80 %	$\widehat{r}_0$	484.082	484.082	484.082
	ncc	10	10	10
Muestra 90 %	$\widehat{r}_0$	484.082	484.082	561.9141
	ncc	17	17	8
Muestra 95 %	$\widehat{r}_0$	484.082	484.082	593.8477
	ncc	26	26	13

Cuadro 5.4: Estimaciones para las muestras efectivas de las regiones de Muros y Porto do Son del 2017 al 2019.

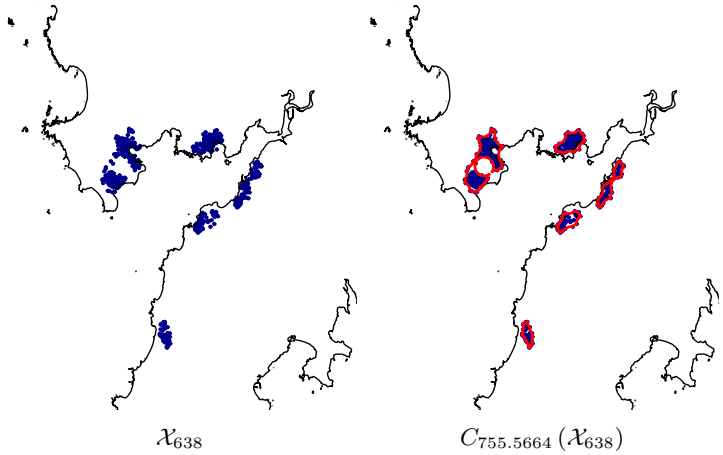


Figura 5.8: Estimaciones para la muestra efectiva del 60 % de Muros y Porto do Son del 2017 al 2019.

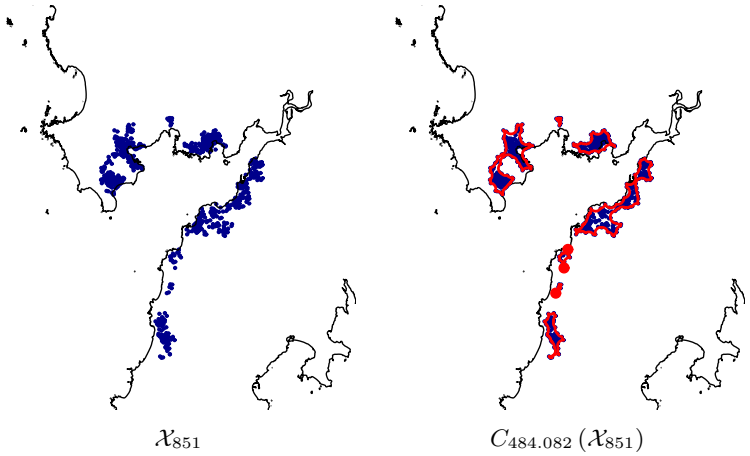


Figura 5.9: Estimaciones para la muestra efectiva del 80 % de Muros y Porto do Son del 2017 al 2019.

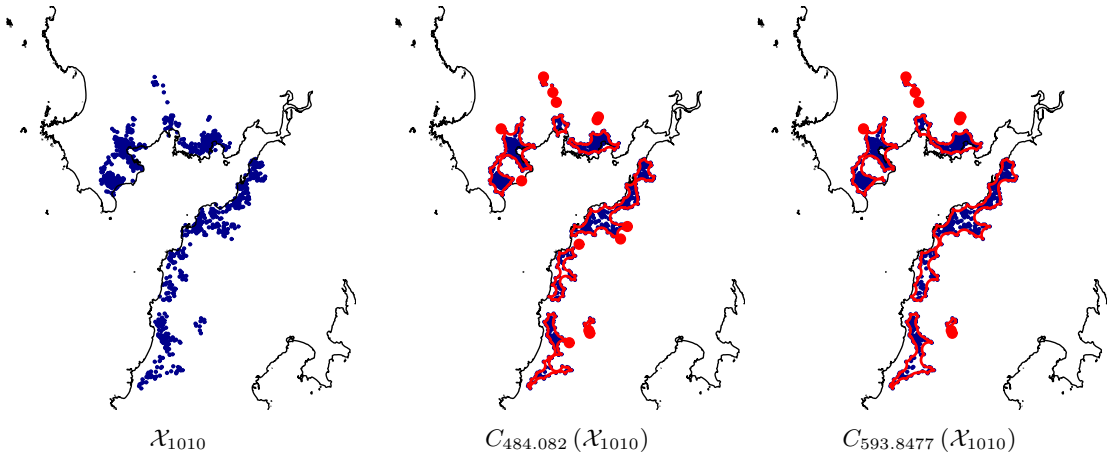


Figura 5.10: Estimaciones para la muestra efectiva del 95 % de Muros y Porto do Son del 2017 al 2019.

Como era de esperar, las estimaciones obtenidas para las muestras efectivas del 60 %, 70 % y 80 % son mejores y no tienden a fragmentarse tanto como ocurre cuando nos acercamos a la muestra completa. Al igual que ocurría en el caso de Galicia, eliminar datos evitando las zonas de baja densidad, ha mejorado considerablemente las estimaciones. Sin embargo, siguen teniendo más componentes conexas de las esperadas aunque esto parece ser un problema debido más al muestreo en esta zona que al método de estimación. Esto se puede observar claramente en la Figura 5.8, donde se ve que las poblaciones más grandes de la zona, como Muros o Esteiros, están sobre-representadas en nuestro conjunto de datos creando esa alta densidad en sus poblaciones. Por otro lado, según aumentamos la proporción de muestra con la que trabajamos, parece que la velutina se localiza principalmente en las zonas costeras de las regiones consideradas. Esto se podría deber a dos causas. La primera sería que la velutina se adapta mejor a las zonas costeras que a las de interior, lo cual explicaría la ausencia de datos en la provincia de Ourense. La segunda puede deberse a que los datos recogidos se restringen principalmente a núcleos poblacionales teniendo apenas datos de las zonas rurales y más despobladas.

Los resultados obtenidos en estas dos regiones ilustran las dos soluciones que hemos manejado ante el problema del calibrado del spacing maximal a lo largo del Capítulo 4, y la Sección 5.1 de este capítulo. En el caso de Vigo hemos comprobado como, al restringir la muestra a un conjunto donde la densidad es bastante homogénea el estimador del soporte ajusta bastante bien el conjunto, sin fragmentarse demasiado. Por otro lado, en la ría de Noia y Muros, los datos se encontraban bastante dispersos en algunas zonas y muy concentrados en otras. Para solventarlo, hemos aplicado la muestra efectiva consiguiendo mejores resultados en los casos del 60 %, 70 % y 80 %. Además, al considerar estas dos zonas hemos podido apreciar que la recogida de datos en el muestreo está muy sesgada a las zonas con alta densidad de población. Esto se puede apreciar al comparar la zona de Vigo, donde los datos son bastante homogéneos, con la zona de Muros y Porto do Son, donde se ve claramente como se concentran sobre todo en los núcleos poblacionales.

Capítulo 6

Conclusiones

Al comienzo de este trabajo hemos abordado la reconstrucción de la extensión de ocurrencia de la velutina en Galicia como un problema de estimación del soporte. Presentamos el conjunto de datos con el que vamos a trabajar y la forma de modelar matemáticamente el problema a tratar. A continuación, en el Capítulo 2, hemos realizado una pequeña revisión de los estimadores del soporte disponibles en la literatura para distintas restricciones de forma sobre dicho soporte. Comenzamos con el caso más general, donde no se asume ninguna condición de forma sobre S, para después abordar la estimación del soporte convexo y r–convexo. En este último, hemos visto que el estimador, la envoltura r-convexa $C_r(\mathcal{X}_n)$, depende de un parámetro desconocido en la práctica r, que deberemos estimar.

En el Capítulo 3, hemos presentado brevemente los métodos de estimación de r existentes y comentado con más detalle el método propuesto por Rodríguez-Casal y Saavedra-Nieves (2019). En la Sección 3.1, introducimos la teoría que sustenta dicho método para, a continuación, describir el algoritmo que permite su cálculo y terminar con su aplicación sobre nuestro conjunto de datos. En su aplicación, tras implementar el algoritmo en el software estadístico R, hemos obtenido estimadores muy fragmentados de la EOO. Para diagnosticar o estudiar las posibles causas, en el Capítulo 4 hemos llevado a cabo un estudio de simulación para comprobar si el calibrado que emplea el método es correcto y estudiar su comportamiento ante las zonas de baja densidad. Dicho estudio, nos ha permitido concluir que el calibrado será bueno siempre y cuando el conjunto a estimar no presente zonas sin datos o con baja densidad.

Para terminar, hemos realizado varias pruebas de estimación del soporte de la velutina teniendo en cuenta las conclusiones obtenidas en el estudio de simulación previo. Para ello, hemos empleado la muestra efectiva como herramienta que nos permite eliminar esas zonas con pocos datos, las cuales nos estaban generando problemas en la estimación de Galicia. Por otro lado, también hemos restringido nuestro conjunto de datos a la zona de la ría de Muros y Noia y a la zona de Vigo. Tras estudiar los datos en estas dos zonas podemos conjeturar que la muestra que disponemos está sesgada ya que, en poblaciones grandes como la de Vigo, observamos que la información recogida es más homogénea que en el caso de las poblaciones pequeñas como las de la región de Porto do Son. Esto resulta significativo en los datos correspondientes a la ría pues, cuando usamos la muestra efectiva, solo nos quedamos con los datos de las poblaciones más grandes. Por tanto, el próximo paso a realizar en un futuro sería intentar depurar la base de datos para eliminar dicho sesgo en el muestreo. Esto nos permitiría reducir ese contraste de densidades entre varias zonas de Galicia para así conseguir estimaciones más satisfactorias de la EOO de la velutina.

Bibliografía

Aaron, C., Cholaquidis, A. and Fraiman, R. (2017). A generalization of the maximal-spacings in several dimensions and a convexity test. Extremes 20(3), 605–634.

Berrendero, J. R., Cuevas, A. and Pateiro-López, B. (2012). A multivariate uniformity test for the case of unknown support. Statistics and Computing 22(1), 259–271.

Bland, L.M., Keith, D.A., Miller, R.M., Murray, N.J. and Rodríguez, J.P. (2016). Directrices para la aplicación de las Categorías y Criterios de la Lista Roja de Ecosistemas de UICN, Versión 1.0. Gland, Suiza: UICN.

Chevalier, J. (2010). Estimation du support et du contour du support d'une loi de probabilité. Annales de l'Institut Henri Poincaré, 12, 339-364.

Cuevas, A. (2010). Set estimation: Another bridge between statistics and geometry. BEIO 25(2), 71-85.

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

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. Advances in Applied Probability, 36, 340-354.

Deheuvels, P. (1983). Strong bounds for multidimensional spacings. Zeitchrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete 64(4), 411-424.

Devroye, L. and Wise, G. L. (1980). Detection of abnormal behaviour via non parametric estimation of the support. SIAM Journal on Applied Mathematics 38(3), 480–488.

Dümbgen, L. and Walther, G. (1996). Rates of convergence for random approximations of convex sets. Advances in Appplied Probabilit, 28, 384–393.

Eglen, S., Gebhard, A., Renka, R. J., White D. Zuyev S. (2016). Triangulation of Irregularly Spaced Data: The R package tripack. CRAN Repository. Disponible en `https://cran.r-project.org/web/packages/tripack/tripack.pdf`

Janson, S. (1987). Maximal spacings in several dimensions. The Annnals of Probability 15(1), 274–280.

Mandal, D. P. and Murthy, C. A. (1997). Selection of alpha for alpha-hull in $\mathbb{R}^2$. Pattern Recognition, 30(10), 1759-1767.

Monceau, K., Bonnard, O. and Thiéry, D. (2014). Vespa velutina: a new invasive predator of honeybees in Europe. Journal of Pest Science 87(1), 1–16.

Pateiro-López, B. and Rodríguez-Casal, A. (2010). Generalization of the convex hull of a sample: The R package alphahull. Journal of Statistical Software, 34, 1-28.

Rodríguez-Casal, A. (2007). Set estimation under convexity type assumptions. Annales de l'I.H.P.–Probabilités & Statistiques, 43, 763–774.

Rodríguez-Casal, A. Saavedra-Nieves, P. (2016). A fully data-driven method for estimating the shape of a point cloud. ESAIM: Probability and Statistics 20, 332-348.

Rodríguez-Casal, A. Saavedra-Nieves, P. (2019). Extent of occurrence reconstruction using a newdata-driven support estimator. ArXiv preprint:1907.08627v1

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

Serra, J. (1984). Image Analysis and Mathematical Morphology. Academic Press, London.

Wand, M. and Jones, M. (1995). Comparison of Smoothing Parameterizations in Bivariate Kernel Density Estimation. Journal of the American Statistical Association 88(422), 520–528.

Watlher, G. (1997). Granulometric smoothing. Annals of Statistics, 25, 2273-2299.

Watlher, G. (1999). On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. Mathematical methods in the Applied Sciences, 22, 301-316.