Trabajo Fin de Máster

Estimación no paramétrica de conjuntos de nivel

Alfredo Montero Fernández

Máster en Técnicas Estadísticas

Curso 2019-2020

Propuesta de Trabajo Fin de Máster

Título en galego: Estimación non paramétrica de conxuntos de nivel
Título en español: Estimación no paramétrica de conjuntos de nivel
English title: Nonparametric level set estimation
Modalidad: Modalidad A
Autor/a: Alfredo Montero Fernández, Universidad de Santiago de Compostela
Director/a: Paula Saavedra Nieves, Universidad de Santiago de Compostela
Breve resumen del trabajo: En este trabajo abordaremos el problema de la estimación no paramétrica de conjuntos de nivel mediante el método de la envoltura $r-$convexa. Como aplicación práctica, ilustraremos su comportamiento mediante la reconstrucción de determinados conjuntos de nivel para la localización de los nidos de avispa velutina en Galicia.
Recomendaciones: Cursar previamente Estadística No Paramétrica

Índice general

Resumen

Resumen en español

La avispa asiática (Vespa velutina) presenta una amenaza sobre la diversidad biológica europea desde su introducción accidental en el continente. A partir del año 2012 se ha ido expandiendo por Galicia llegando a colonizar todo el territorio. En este trabajo se estudiará la reconstrucción de los distintos focos de nidos de la avispa sobre la comunidad gallega mediante un problema de estimación no paramétrica de conjuntos de nivel. Se revisarán distintas metodologías para, dada una muestra aleatoria simple generada a partir de una distribución desconocida, la reconstrucción de un conjunto de nivel que contenga una probabilidad fijada. Se estudiará específicamente el método de la envoltura $r-$convexa. El principal problema de este, es la fuerte dependencia sobre el parámetro r. Se presentará y estudiará una versión mejorada de este método a partir de una propuesta para la estimación del valor r y lo aplicaremos sobre la distribución de los nidos de velutina en Galicia.

English abstract

The Asian hornet (Vespa velutina) has posed a threat to European biological diversity since its accidental introduction on the continent. Since 2012 it has been expanding throughout Galicia, colonizing the entire territory. In this work, the reconstruction of the different foci of the hornet nests on the Galician community will be studied by means of a non-parametric estimation problem of level sets. Different methodologies will be reviewed for, given a simple random sample generated from some unknown distribution, the reconstruction of a level set containing a fixed probability. The $r-$convex hull method will be specifically studied. The main problem with this is the strong dependency on the r parameter. An improved version of this method will be presented and studied based on a proposal for estimating the value r, and we will apply it to the distribution of velutina nests in Galicia.

Capítulo 1

Introducción

La *Vespa velutina nigrithorax*, conocida comúnmente como avispa negra, avispón asiático o avispa velutina, se ha convertido en el primer depredador Vespidae accidentalmente introducido desde Asia a Europa, a través de Francia en el año 2004 (Rortais et al., 2010). El origen de la introducción es incierto, y se cree que podrían haber sido importadas accidentalmente a través del comercio agrícola antes del 2004 (Villemant et al., 2006). En este continente se la considera una especie exótica invasora debido a que es una amenaza para la diversidad biológica nativa, principalmente para las abejas.

La expansión por Europa se fue realizando progresivamente, primero se detectó en España en el año 2010 (Castro and Pagola-Carte, 2010), en Bélgica y el norte de Portugal en el 2011, en Italia en el 2012, en Alemania en el 2014, Reino Unido en el 2016 y en Suiza en el 2017 (ver por ejemplo Rojas-Nossa et al. (2018)). En la Figura 1.1 se muestra un mapa obtenido de INPN (2020) en el que se presenta la distribución a día 4 de agosto del año 2020 de la velutina en Europa.

Los primeros avistamientos en la Península Ibérica tuvieron lugar en el año 2010 en Navarra y Guipúzcoa , seguidos cronológicamente por la detección de casos en el norte de Lugo, Vizcaya, Álava y el norte de Cataluña en 2012, el sur de Pontevedra en 2013, el oriente de Cantabria, el occidente de Asturias, La Rioja y el norte de Burgos en 2014 y la Isla de Mallorca en el 2015, aunque aquí se cree controlada al no haberse detectado ningún caso en el 2019. Por último, en el 2019 se avista por primera vez un nido en la provincia de Valladolid (se puede consultar Viejo (2020)).

Debido a la repercusión medioambiental, económica y de conservación de la biodiversidad, la especie fue incluida en el Anexo del Real Decreto 630/2013, de 2 de agosto, por el que se regula el Catálogo español de especies exóticas invasoras. A partir del 2014 se crea en Galicia el Protocolo De Vigilancia y Control Frente a la Avispa Asiática, en el cual, entre otras cosas, se trata cómo afecta la presencia de esta avispa a distintos ámbitos, principalmente al de la producción apícola y al ámbito medioambiental, debido a que son un depredador natural de las abejas. También provoca una alarma ciudadana, ya que, aún no siendo una avispa más violenta que las de otras especies, sí se vuelve agresiva cerca de los nidos, los cuales se encuentran, no solo en espacios rurales sino en espacios urbanos también, lo cual ha provocado casi una decena de muertes en Galicia (Xunta de Galicia, 2016).

En este trabajo tendremos el objetivo de realizar estimaciones no paramétricas de determinados conjuntos de nivel de la función de densidad de la localización de nidos de velutina en la comunidad autónoma de Galicia. Se trata de estimaciones no paramétricas debido a que no tendremos ninguna suposición de que la función de densidad f pertenezca a ninguna clase de funciones. Los conjuntos de nivel a estimar serán los valores x tales que la función de densidad aplicada en esos puntos sea

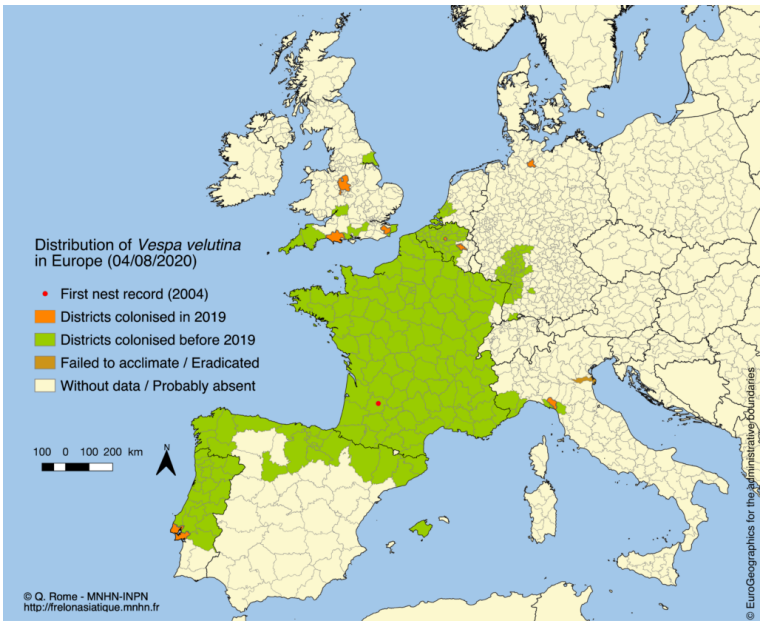


Figura 1.1: Mapa de la distribución de la velutina en Europa.

mayor que un determinado valor t. Esta forma de definirlos tiene el problema de que muchas veces no tenemos información sobre los valores que puede tomar $f(x)$, por lo tanto los conjuntos de nivel que definitivamente usaremos serán aquellos más pequeños cuya probabilidad total sea mayor que un parámetro $\tau \in (0, 1)$. De esta forma, podremos estimar las zonas en las que se encuentran los distintos focos de la avispa.

El conjunto de datos que utilizaremos, aportado por la Consellería de Medio Rural de la Xunta de Galicia, contiene la fecha y localización de los nidos de velutina registrados en Galicia entre los años 2014 y 2019, ambos inclusive. La localización viene data en coordenadas UTM y los representaremos directamente sobre la comunidad de Galicia. Los datos se pueden visualizar gráficamente en la Figura 1.2. Se puede observar la detección al principio de dos focos, en la zona más septentrional y en la parte suroeste de la comunidad, pero año tras año se ve cómo la detección de nidos se extiende por la comunidad hacia el interior, quedando cada vez más homogéngea la distribución. Los años de mayor incidencia fueron el 2017 y el 2018, viéndose reducido notablemente el número de casos en el 2019.

El problema de la estimación de los conjuntos de nivel de una distribución es un problema matemático perteneciente al estudio de la estimación de conjuntos, por lo tanto la geometría juega un papel esencial. El objetivo en general será la reconstrucción de ciertos conjuntos compactos en $\mathbb{R}^d$ a partir de una muestra aleatoria de puntos que tienen una relación directa con el conjunto a reconstruir. Cuando el valor $\tau = 0$, el problema se convierte en la reconstrucción del soporte de la distribución, que es, por lo tanto, un caso particular, por lo que se podrá abordar desde otros puntos de vista más concretos. El objetivo será conseguir buenos resultados teóricos para cualquier valor de τ bajo las menores restricciones posibles, pero siendo estas necesarias para que la reconstrucción sea correcta.

El contenido del presente Trabajo de Fin de Máster se distribuye de la siguiente manera. En el Capítulo 1 hemos introducido la idea del objetivo de este trabajo. La estimación no paramétrica de conjuntos de nivel, los cuales hemos tratado superficialmente. También hemos presentado los datos de los nidos de velutina en Galicia, datos que utilizaremos para apoyarnos a la hora de aplicar las estimaciones de dichos conjuntos.

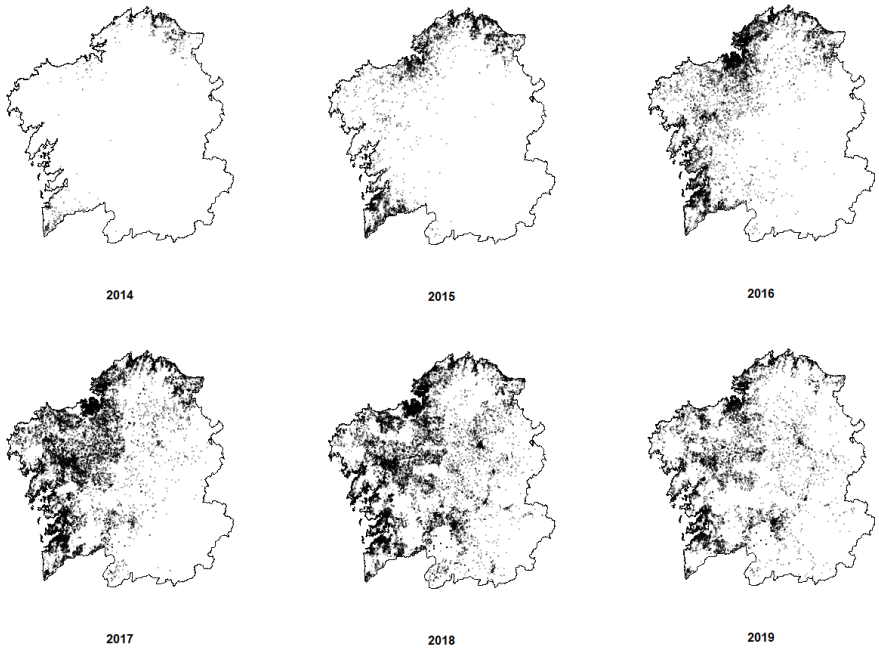


Figura 1.2: Localización de los nidos de velutina en Galicia entre los años 2014 y 2019.

En el Capítulo 2 estudiaremos con detalle los conjuntos comentados anteriormente, definiéndolos de forma concreta. Estudiaremos tres distintos métodos generales para la reconstrucción de los conjuntos de nivel, uno de los cuáles no tendrá restricciones geométricas sobre el conjunto a estimar, mientras que los otros dos requerirán de unas condiciones geométricas que definiremos anteriormente, siendo la más importante la $r-$convexidad, la cual depende de un parámetro r. Por último introduciremos las nociones de distancias entre conjuntos para su uso en los próximos Capítulos, teniendo especial relevancia en nuestro trabajo la distancia de Hausdorff.

Una vez presentados los métodos bajo condiciones geométricas, en el Capítulo 3 presentaremos el método propuesto por Saavedra (2014) para la estimación de conjuntos de nivel bajo la hipótesis de que dichos conjuntos son $r-$convexos. Primero introduciremos la noción del parámetro óptimo r_0 que nos proporcionará el mayor valor de r tal que el conjunto sea $r-$convexo. Una vez hecho esto estudiaremos el cálculo de una estimación de este parámetro y la construcción del conjunto de nivel estimado. Acto seguido detallaremos el algoritmo propuesto es una versión posterior revisada de Rodríguez-Casal and Saavedra-Nieves (2019) para la computación de los conjuntos deseados a partir de una muestra aleatoria simple y acabaremos el Capítulo con la aplicación a una muestra concreta de la velutina que es utilizada anteriormente en el Capítulo 2.

Para comprobar que el algoritmo anteriormente mencionado funciona bien, realizaremos, en el Capítulo 4, un estudio de simulación empleando varios modelos de mixturas de normales bidimensionales cuya función de densidad es conocida, pudiendo así comparar el conjunto estimado con el conjunto real para comprobar si el ajuste es bueno. Una vez comprobado que el algoritmo se comporta bien, en el Capítulo 5 lo aplicaremos sobre los datos de la velutina presentados al inicio de este Capítulo, de forma anual, para poder estudiar así cómo varía la concentración de nidos tras los años a lo largo del territorio gallego, y en la zona de Santiago de Compostela. Finalizaremos este Trabajo con el Capítulo 6, el cual servirá para la exposición final de las conclusiones sacadas a lo largo de los Capítulos y presentar una posible extensión del trabajo.

Capítulo 2

Estimación de conjuntos de nivel de la densidad

Dada una muestra aleatoria simple $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ de una variable aleatoria absolutamente continua X con distribución de probabilidad $\mathbb{P}_X$, puede ser interesante estudiar el soporte compacto y no vacío de la misma, pero cuando una parte importante de este sea casi vacío desde un punto de vista probabilístico, será más importante *estudiar determinados conjuntos de nivel.* Estos conjuntos de nivel serán de la forma:

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

con $t > 0$. Sin embargo, en muchas ocasiones no tenemos información a priori sobre los posibles valores de $f(t)$. En la práctica esto suele ser así, por lo que nos dedicaremos a, dado un valor $\tau \in (0, 1)$, estimar el conjunto de nivel que tiene una probabilidad contenida mayor o igual a $1 - \tau$, es decir, estimar el conjunto:

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

donde

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

Por lo tanto, nuestro objetivo será estimar el conjunto $L(\tau)$ a partir de una muestra aleatoria simple $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ de X con función de densidad f. La estimación no paramétrica de conjuntos de nivel ha sido considerada en la literatura desde tres perspectivas distintas. En la Sección 2.1 presentaremos la primera y más general de estas tres perspectivas: la metodología *plug-in*. En la Sección 2.2 estableceremos unas condiciones geométricas que tendrán bastante importancia a lo largo de este trabajo, sobre todo la condición de $r-$convexidad. Además presentaremos brevemente las otras dos perspectivas de la estimación de conjuntos de nivel de la función de densidad: la metodología del exceso de masa y la metodología híbrida. Por último estableceremos la noción de distancia entre conjuntos, lo cual nos servirá para determinar el error de estimación del conjunto a estimar.

2.1. Métodos plug-in

La forma más simple de estimar $L(\tau)$ cuando no se tienen condiciones geométricas sobre el conjunto resultan las estimaciones *plug-in*. Estos métodos proponen estimar $L(\tau)$ como

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

donde f_n es un estimador no paramétrico de la función de densidad f, en general el estimador no paramétrico de tipo núcleo

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

con H un matriz de ventanas $d-$dimensional, simétrica y definida positiva, $K_H(z) = |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}} z)$ y $K : \mathbb{R}^d \to \mathbb{R}$ una función tipo núcleo (por lo general será la densidad gaussiana); y $\hat{f}_\tau$ es un estimador de f_τ obtenido mediante métodos de integración numérica o, como propuso, Hyndman (1996) utilizando el $\tau-$cuantil de la distribución empírica de $f_n(X_1) \ldots, f(X_n)$.

Para distintos métodos *plug-in* de bases de datos específicos puede consultarse Baíllo and Cuevas (2006) o Singh et al. (2009). Para el caso unidimensional puede consultarse el método propuesto por Samworth and Wand (2010).

2.2. Métodos bajo condiciones geométricas

En los casos más generales no se le imponen condiciones geométricas a los conjuntos teóricos que han de ser estimados. Sin embargo, si se tiene alguna información adicional, se pueden conseguir estimadores más sofisticados. Ahora definiremos las propiedades de convexidad, de $r-$convexidad para un valor $r > 0$ y la propiedad de libre rodamiento que podremos utilizar para considerar otros estimadores más potentes de $L(\tau)$ bajo las condiciones establecidas.

2.2.1. Condiciones geométricas

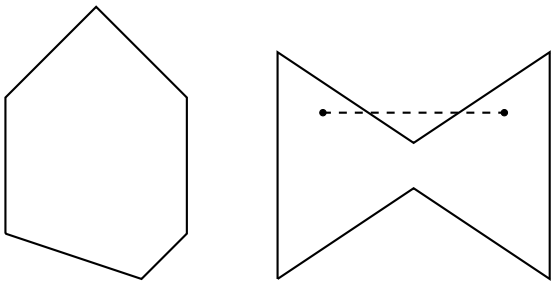
Definición 2.1. Un conjunto $A \subset \mathbb{R}^d$ se dice convexo si para cada par de puntos $x, y \in A$ y para todo $\gamma \in [0, 1]$ se verifica que $\gamma x + (a - \gamma) y \in A$.

Esto es, un conjunto será convexo si puedes llegar desde un punto cualquiera hasta otro en línea recta sin salirte del propio conjunto. En la Figura 2.1 se muestra un ejemplo de conjunto convexo y otro de conjunto no convexo en $\mathbb{R}^2$.

Definición 2.2. Sea $A \subset \mathbb{R}^d$ un conjunto. Se define la envoltura convexa de A como la intersección de todos los conjuntos convexos de $\mathbb{R}^d$ que contienen a A. Se denota por $H(A)$.

Esto es, $H(A)$ es el menor conjunto convexo que contiene a A. En la Figura 2.2 se puede observar la envoltura convexa del conjunto no convexo presentado en la figura 2.1.

En la práctica asumir convexidad suele ser muy restrictivo ya que cualquier conjunto con más de una componente conexa no es nunca convexo. Por ello vamos a definir una condición geométrica mucho menos restrictiva que generaliza la propiedad de convexidad como es la $r-$convexidad para un $r > 0$.



Conjunto convexo Conjunto no convexo

Figura 2.1: A la izquierda un conjunto convexo. A la derecha un conjunto no convexo.

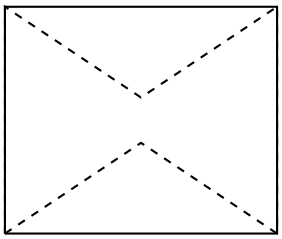


Figura 2.2: En continua la envoltura convexa del conjunto en discontinua.

Definición 2.3. Sea $A \subset \mathbb{R}^d$ un conjunto cerrado. Dado un $r > 0$ definimos la envoltura $r-$convexa de A como:

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

donde $B_r(x)$ representa la bola abierta centrada en x de radio r.

Definición 2.4. Un conjunto cerrado $A \subset \mathbb{R}^d$ se dice que es $r-$convexo para cierto $r > 0$ si

$$A = C_r(A).$$

Mientras que la envoltura convexa coincide con la intersección de todos los semiespacios cerrados que contienen a A, la envoltura $r-$convexa se calcula como la intersección de los complementos de las bolas abiertas de radio r que no intersecan con A. De forma intuitiva podemos pensar en la envoltura convexa de un conjunto A como la envoltura $r-$convexa cuando el valor de r tiende a infinito. En las Figuras 2.3 y 2.4 se puede observar la dependencia del valor r a la hora de que un mismo conjunto sea o no $r-$convexo.

La envoltura $r-$convexa satisface varias propiedades interesantes. Por ejemplo, $C_r(A) \subset C_{r*}(A)$ para todo $r \leq r^*$. También se puede probar que si A es un conjunto cerrado y convexo, entonces será $r-$convexo para todo $r > 0$. En la Figura 2.5 se puede ver esto con unos ejemplos ilustrados.

Por último introduciremos la condición de libre rodamiento.

Definición 2.5. 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 para cada punto frontera $b \in \partial A$ existe un $x \in \mathbb{R}^d$ tal que $b \in B_r[x] \subset A$.

En la Figura 2.6 se muestran dos conjuntos, uno con la propiedad de libre rodamiento y la otra sin ella. Posteriormente en este trabajo estableceremos la relación entre esta última condición y la $r-$convexidad.

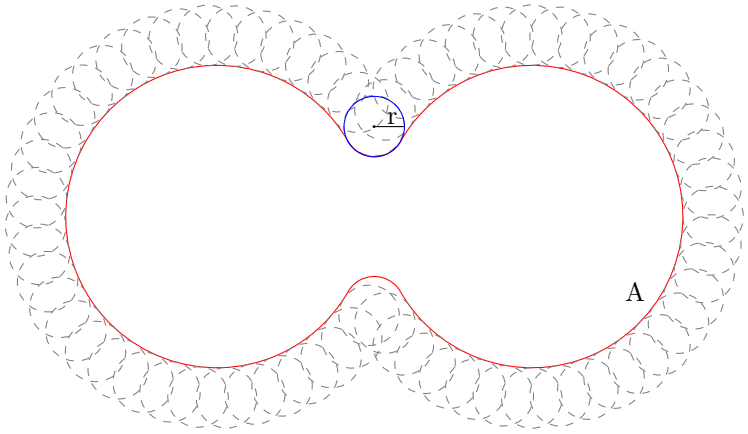


Figura 2.3: El conjunto A es igual a $C_r(A)$ y, por lo tanto, es $r-$convexo.

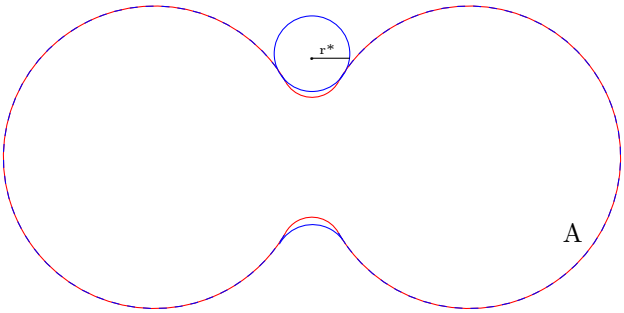


Figura 2.4: El conjunto A en rojo no es igual a $C_{r^*}(A)$ en azul y, por lo tanto, no es r^*-convexo.

2.2.2. Métodos bajo condiciones geométricas

Bajo la hipótesis de que el conjunto a estimar cumple alguna determinada restricción geométrica se pueden establecer nuevos modelos. La metodología de exceso de masa utiliza el hecho de que el conjunto de nivel $G(t)$, para un $t > 0$, maximiza el funcional

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

donde B es un conjunto de Borel, $\mathbb{P}$ es la medida de probabilidad inducida por f y μ es la medida de Lebesgue. Por lo tanto, si $\mathcal{B}$ es una clase de conjuntos, bajo la restricción $G(t) \in \mathcal{B}$, un estimador natural $\hat{G}(t)$ de $G(t)$ sería el maximizador, en $\mathcal{B}$, del exceso de masa empírico

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

donde $\mathbb{P}_n$ denota la probabilidad empírica inducida por la muestra $\mathcal{X}$.El caso cuando $\mathcal{B}$ es la clase de conjuntos convexos fue estudiado por Hartigan (1987) para el caso bidimensional y por Grübel (1988) para el caso unidimensional.

A la hora de estimar bajo las mismas condiciones el conjunto de nivel $L(\tau)$, el método empírico propuesto por Walther (1997) estima como $\hat{L}(\tau) = \hat{G}(\hat{f}_\tau)$, donde

$$\hat{f}_\tau = \text{máx}\left\{t > 0 : \mathbb{P}_n\left(\hat{G}(t)\right) \geq 1 - \tau\right\}.$$

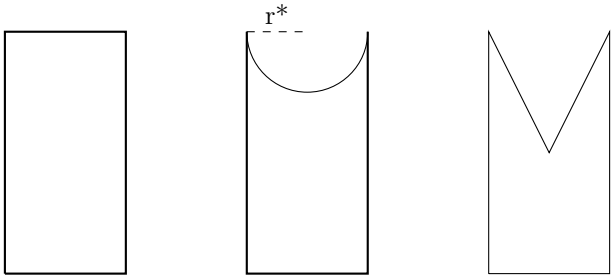


Figura 2.5: El conjunto de la izquierda es convexo y, por lo tanto, $r-$convexo para todo $r > 0$. El conjunto central no es convexo, pero es $r-$convexo para todo $r \leq r^*$. El conjunto de la derecha no es $r-$convexo para ningún $r > 0$.

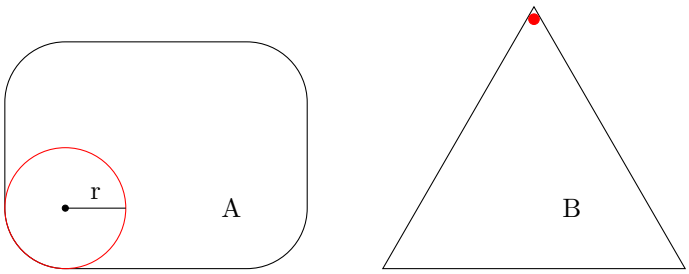


Figura 2.6: Una bola de radio r rueda libremente en A, pero ninguna bola de radio r^* rueda libremente en B con $r^* > 0$.

Los métodos que asumen tanto restricciones geométricas a priori sobre $L(\tau)$ como un estimador piloto no paramétrico de la densidad, son denominados métodos híbridos. Un método híbrido clásico es el método de suavizado granulométrico propuesto por Walther (1997) que asume que una bola de radio r rueda libremente sobre $L(\tau)$ y sobre la clausura de su complementario $\overline{L(\tau)^c}$. Este método depende de un parámetro r: para valores demasiado grandes de este parámetro el estimador obtenido puede ser igual al vacío, mientras que valores pequeños de r provoca estimadores separados.

Dos métodos propuestos en Saavedra (2014) son el método de la envoltura convexa y el método de la envoltura $r-$convexa.

El primero de estos métodos estima $L(\tau)$ asumiendo a priori que el conjunto de nivel es convexo. Para pequeños valores de τ se trata de una condición poco restrictiva. El método estima $L(\tau)$ de la siguiente forma:

1. Primero mediante un estimador no paramétrico tipo núcleo f_n se halla el estimador $\hat{f}_\tau$ como

$$\hat{f}_\tau = \text{máx}\left\{t : \mathbb{P}_n\left(\hat{G}(t)\right) \geq 1 - \tau\right\},$$

   donde $\hat{G}(t) = H(\mathcal{X}_{n,+}(t))$, siendo $\mathcal{X}_{n,+}(t) = \{X \in \mathcal{X}_n : f_n(X) \geq t\}$.

2. Se define el estimador del conjunto de nivel

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

El principal problema de este estimador es la asunción de que $L(\tau)$ es convexo, lo cual en la mayoría de las situaciones es muy restrictivo.

El segundo método estima $L(\tau)$ asumiendo a priori que el conjunto de nivel es $r-$convexo. Esta asunción es mucho menos restrictiva que la convexidad. El método estima $L(\tau)$ de la siguiente forma:

1. Primero mediante un estimador no paramétrico tipo núcleo f_n se halla el estimador $\hat{f}_\tau$ como

$$\hat{f}_\tau = \text{máx}\left\{t : \mathbb{P}_n\left(\hat{G}(t)\right) \geq 1-\tau\right\},$$

   donde $\hat{G}(t) = C_r(\mathcal{X}_{n,+}(t))$ para un $r > 0$ fijo.

2. Se define el estimador del conjunto de nivel

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

El principal problema de este método es la dependencia sobre el parámetro $r > 0$ que, por lo general, es desconocido. Para valores demasiado pequeños de r, el estimador resulta muy dividido, mientras que valores demasiado grandes de r producen estimadores muy parecidos a la envoltura convexa. En la Figura 2.7 se puede observa esto con bastante claridad. Para el valor $r = 200000$ la envoltura $r-$convexa unifica los dos grupos de puntos claramente separados, añadiendo muchos puntos del conjunto $X_{n,-}(\hat{f}_{0.3})$ a la estimación, siendo $\mathcal{X}_{n,-}(\hat{f}_\tau) = \mathcal{X} \setminus \mathcal{X}_{n,+}(\hat{f}_\tau)$, viéndose su parecido a la envoltura convexa, mientras que para el valor $r = 2000$, divide en muchos grupos pequeños la muestra $\mathcal{X}_{n,+}(\hat{f}_{0.3})$. En la Figura 2.8 se ve una versión ampliada de la zona norte en el caso $C_{8000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$ y en la Figura 2.9 una versión ampliada de $C_{2000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$.

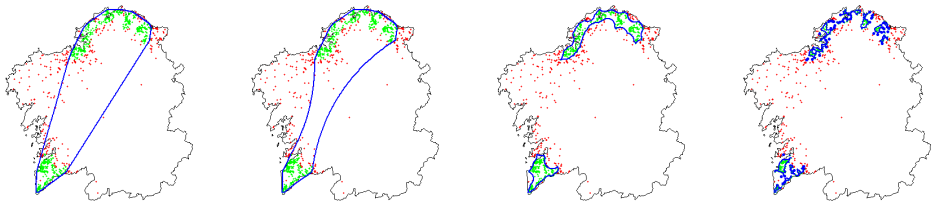


Figura 2.7: El conjunto $\mathcal{X}_{n,+}(\hat{f}_{0.3})$ en verde y $\mathcal{X}_{n,-}(\hat{f}_{0.3})$ en rojo para los datos de la velutina del mes de octubre del año 2015. En azul aparece representado $H(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$ en la figura de la izquierda, $C_{200000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$ en la figura del centro izquierda, $C_{8000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$ en la del centro derecha y $C_{2000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$ en la de la derecha.

2.2.3. Distancias entre conjuntos

Para poder evaluar cómo se ajusta el conjunto estimado $\hat{L}(\tau)$ a $L(\tau)$ debemos ser capaces de medir la *distancia* entre ambos conjuntos. Para esto uno podría pensar en la distancia euclidiana, pero esta

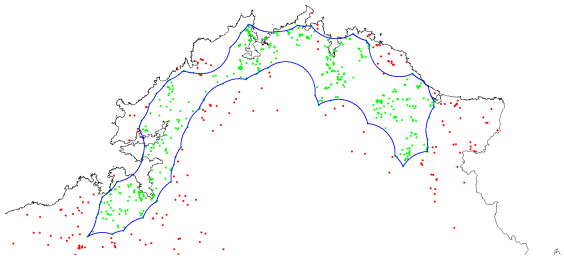


Figura 2.8: El conjunto $\mathcal{X}_{n,+}(\hat{f}_{0.3})$ en verde y $\mathcal{X}_{n,-}(\hat{f}_{0.3})$ en rojo para la zona norte de Galicia de los datos de la velutina del mes de octubre del año 2015. En azul aparece representado $C_{8000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$.

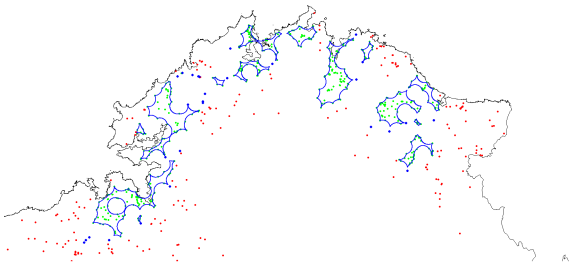


Figura 2.9: El conjunto $\mathcal{X}_{n,+}(\hat{f}_{0.3})$ en verde y $\mathcal{X}_{n,-}(\hat{f}_{0.3})$ en rojo para la zona norte de Galicia de los datos de la velutina del mes de octubre del año 2015. En azul aparece representado $C_{2000}(\mathcal{X}_{n,+}(\hat{f}_{0.3}))$.

tiene un gran problema y es que conjuntos bastante distintos pueden tener una distancia euclídea igual a cero como se muestra en la Figura 2.10.

Así pues necesitamos definir otro tipo de distancia sobre los conjuntos.

Definición 2.6. Sean A y C dos conjuntos de Borel acotados. La distancia en medida entre A y C se define como:

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

donde μ denota la medida de Lebesgue y $A \triangle C = (A \setminus C) \cup (C \setminus A)$.

La distancia en medida cuantifica la similitud en contenido. Si queremos una distancia que cuantifique la proximidad física entre dos conjuntos podemos usar la distancia de Hausdorff.

Definición 2.7. Sean $A,\, C \subset \mathbb{R}^d$ conjuntos compactos no vacíos. Se define la distancia de Hausdorff entre A y C, siendo $d(a, C) = \text{ínf}\{||a - c|| : c \in C\}$ como :

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

En la Figura 2.11 se ilustra un ejemplo de esta distancia. Intuitivamente se trata de la más larga de las distancias euclídeas que podemos establecer entre un punto y el punto más cercano del otro conjunto.

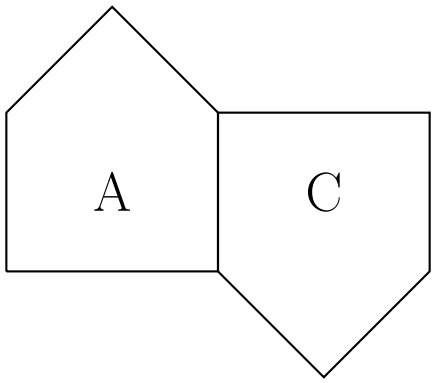


Figura 2.10: Los conjuntos A y C tienen un borde en común, por lo que su distancia euclídea es cero.

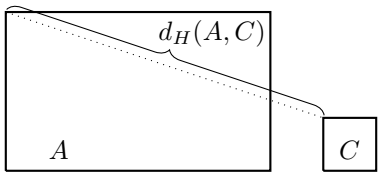


Figura 2.11: Distancia de Hausdorff entre los conjuntos A y C.

Con estas distancias tenemos pues, una forma de medir el error de estimación entre los conjuntos estimados y los teóricos. En el capítulo siguiente estableceremos el método introducido por Saavedra (2014) para estimar un valor óptimo de r que definiremos previamente y para calcular la estimación de $L(\tau)$ mediante una versión mejorada del método de la envoltura $r-$convexa.

Capítulo 3

Algoritmo para la estimación de conjuntos de nivel

Una vez presentado el método de la envoltura $r-$convexa y habiendo visto el problema que puede presentar la selección del parámetro r, en este capítulo presentaremos la versión mejorada tratada por Saavedra (2014) y presentaremos una versión revisada del algoritmo propuesto en Rodríguez-Casal and Saavedra-Nieves (2019) para la estimación de conjuntos de nivel de la función de densidad. Para ello primero tendremos que establecer un método automático para la selección del parámetro r.

En la Sección 3.1 definiremos el parámetro óptimo $r_0(t)$ y su respectivo estimador $\hat{r}_0(t)$ para el problema de estimar $G(t)$. En la Sección 3.2 estableceremos los órdenes de convergencia con respecto a las distancias entre conjuntos introducidas en el Capítulo 2 de los estimadores de $G(t)$ y de $L(\tau)$. Por último, en la Sección 3.3 se propondrá el algoritmo para la selección del parámetro r y acto seguido el algoritmo para la estimación de $L(\tau)$.

3.1. Estimación del parámetro óptimo

Dada una muestra aleatoria simple $\mathcal{X}_n$ queremos estimar un valor de r ideal para la estimación de $G(t)$ mediante un método de envoltura $r-$convexa. Para ello tenemos que empezar definiendo un parámetro óptimo $r > 0$, que será el mayor valor de $r > 0$ tal que $G(t)$ es $r-$convexo. Por simplificación en la exposición, primero supondremos que el conjunto $G(t)$ no es convexo pues, en caso contrario, el parámetro óptimo valdría infinito.

Definición 3.1. Sea $G(t)$ un conjunto de nivel compacto, no vacío, no convexo y $r-$convexo para cierto $r > 0$. Definimos el parámetro óptimo

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

El supremo anteriormente establecido resulta ser un máximo bajo la siguiente propiedad geométrica:

(R_λ^r) Una bola cerrada de radio $\lambda > 0$ rueda libremente en $G(t)$ y una bola cerrada de radio $r > 0$ rueda libremente en $\overline{G(t)^c}$.

Que se cumpla la condición (R_λ^r) es una propiedad natural en general de los conjuntos de nivel de una función de densidad. Ahora estableceremos las hipótesis del Teorema 2 de Walther (1997):

A.
1. El umbral t de $G(t)$ pertenece a $[l,u]$ con $-\infty < l \leq u < sup(f) < \infty$
2. $f \in \mathcal{C}^p(U)$, $p \geq 1$ donde U es un conjunto abierto y acotado que contiene a $\overline{G(l-\zeta)} \backslash \text{Int}(G(u+\zeta))$ para cierto $\zeta > 0$, donde $G(u+\zeta)$ es acotado.
3. El gradiente de f, ∇f, satisface $|\nabla f| \geq m > 0$ y la condición Lipschitz en U:

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

Bajo estas condiciones se verifica (R_λ^r) para $r = \lambda = m/k$. La hipótesis (A) garantiza que $G(t)$ es un conjunto compacto y no vacío y $r_0(t) \geq m/k$, para todo $t \in [l,u]$. La consideración adicional de la condición de forma (R_λ^r) nos proporciona varias propiedades interesantes. En particular se puede demostrar que bajo la condición (R_λ^r), $G(t)$ es $r-$convexo. Más específicamente, bajo (R_λ^r), la propiedad de que una bola de radio r ruede libremente implica $r-$convexidad, lo cual no ha de ocurrir siempre como se indica en Cuevas et al. (2012). Un ejemplo de esto se muestra en la Figura 3.1.

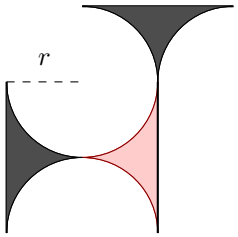


Figura 3.1: Para el conjunto mostrado en negro se cumple que una bola de radio r rueda libremente sobre el conjunto. Sin embargo el conjunto no es $r-$convexo ya que su envoltura $r-$convexa es la unión del conjunto negro con el rojo.

En el método de la envoltura $r-$convexa dividíamos la muestra original $\mathcal{X}_n$ en dos subconjuntos $\mathcal{X}_{n,+}$ y $\mathcal{X}_{n,-}$ de los cuales solo utilizábamos el primero. De esta forma no considerábamos ninguna información que nos pudiera aportar el complementario de $G(t)$. Para resolver este problema, modificando el método anteriormente comentado, empezaremos por proponer un estimador del parámetro óptimo r_0, para lo cual tenemos que considerar una secuencia D_n que cumpla la siguiente asunción:

D. D_n es igual a $M(\log n/n)^{p/(d+2p)}$ un valor suficientemente grande de la constante $M > 0$.

Definición 3.2. Sea $G(t)$ un conjunto de nivel compacto, no vacío y no convexo. Bajo las asunciones (A) y (D), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de distribución f. Un estimador de r_0 puede ser definido como:

$$\hat{r}_0(t) = \sup\left\{\gamma > 0 : C_\gamma\left(\mathcal{X}_n^+(t)\right) \cap \mathcal{X}_n^-(t) = \varnothing\right\},$$

donde

$$\mathcal{X}_n^+(t) = \{X \in \mathcal{X}_n : f_n(X) \geq t + D_n\} \text{ y } \mathcal{X}_n^-(t) = \{X \in \mathcal{X}_n : f_n(X) < t - D_n\}.$$

De esta forma dividimos a la muestra $\mathcal{X}_n$ en tres submuestras, $\mathcal{X}_n^+(t)$, $\mathcal{X}_n^-(t)$ y $\mathcal{X}_n \setminus \mathcal{X}_n^+(t) \cup \mathcal{X}_n^-(t)$. Desde un punto de vista intuitivo, $\mathcal{X}_n^+(t)$ debería estar contenido en $G(t)$ y $\mathcal{X}_n^+(t)$ en $G(t)^c$. Esto ocurre así, incluso para conjuntos convexos, como enunciaremos formalmente más adelante. También se puede probar que $\mathcal{X}_n^+(t) \neq \varnothing$ y que, si $G(t)$ no es convexo, con probabilidad uno y para n lo suficientemente

grande, el conjunto $\sup\{\gamma > 0 : C_\gamma(\mathcal{X}_n^+(t)) \cap \mathcal{X}_n^-(t) = \varnothing\}$ es no vacío y acotado superiormente, por lo que el estimador propuesto está bien definido. Para garantizar que $\hat{r}_0$ satisface estas interesantes propiedades tenemos que considerar dos propiedades sobre el estimador tipo núcleo f_n de f. Se puede consultar Walther (1997) o Giné and Guillou (2002) para más detalles:

K.
1. La función tipo núcleo K es continua de al menos orden p con soporte acotado y varianza finita.
2. El parámetro ventana h es del orden $(\log n/n)^{1/(d+2p)}$.

Proposición 3.3. *Sea $G(t)$ un conjunto de nivel compacto y no vacío. Bajo las asunciones (A), (D) y (K), sea $\mathcal{X}_n$ un muestra aleatoria generada a partir de una distribución con función de densidad f. Entonces,*

$$\mathbb{P}\left(\mathcal{X}_n^+(t) \subset G(t), \mathcal{X}_n^-(t) \subset G(t)^c, \text{ eventualmente }\right) = 1.$$

Proposición 3.4. *Sea $G(t)$ un conjunto de nivel compacto y no vacío. Bajo las asunciones (A) , (D) y (K), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de densidad f. Entonces, para todo $\varepsilon > 0$ se verifica que*

$$\mathbb{P}\left(\sup_{t\in[l,u]} \sup_{x\in G(t)} d(x, \mathcal{X}_n^+(t)) \leq \varepsilon, \text{ eventualmente }\right) = 1.$$

Esta proposición garantiza que, eventualmente, de forma casi segura, $\mathcal{X}_n^+$ es no vacío. El siguiente corolario nos muestra, en particular, que $\mathcal{X}_n^+$ es un estimador consistente de $G(t)$ uniformemente en t con la distancia de Hausdorff.

Corolario 3.5. *Sea $G(t)$ un conjunto de nivel compacto y no vacío. Bajo las asunciones (A) , (D) y (K), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de densidad f. Entonces, para todo $\varepsilon > 0$ se verifica que*

$$\mathbb{P}\left(d_H(G(t), \mathcal{X}_n^+(t)) \leq \varepsilon \text{ eventualmente }\right) = 1.$$

El siguiente teorema nos garantiza la convergencia uniforme en $t \in [l, u]$ para el estimador del parámetro $r_0(t)$.

Teorema 3.6. *Sea $G(t)$ un conjunto de nivel compacto, no vacío y no convexo. Bajo las asunciones (A) , (D) y (K), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de densidad f. Entonces, para todo $\varepsilon > 0$*

$$\mathbb{P}\left(\sup_{t\in[l,u]} |\hat{r}_0(t) - r_0(t)| \leq \varepsilon, \text{ eventualmente }\right) = 1.$$

3.2. Ordenes de convergencia del estimador

Una vez estudiada la consistencia del parámetro $\hat{r}_0(t)$, sería natural considerar $C_{\hat{r}_0(t)}(\mathcal{X}_n^+(t))$ como un estimador del conjunto de nivel $G(t)$. Sin embargo, en general, la consistencia de este estimador no puede ser garantizada. Saavedra (2014) propone $\hat{G}(t) = C_{r_n(t)}(\mathcal{X}_n^+(t))$ como el estimador de $G(t)$, donde $r_n(t) = \nu\hat{r}_0(t)$ para un valor fijo de $\nu \in (0, 1)$. Como $G(t)$ es r^*-convexo si es $r-$convexo, se tiene que $G(t) = C_r(G(t)) = C_{r^*}(G(t))$, por lo que cabe esperar que para n suficientemente grande , $G(t)$ sea $r_n(t)-$convexo. Esto lo estableceremos en la siguiente proposición y, acto seguido, proveeremos los órdenes de convergencia en forma de Teorema.

Proposición 3.7. *Sea $G(t)$ un conjunto de nivel compacto, no vacío y no convexo. Bajo las asunciones (A) , (D) y (K), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de densidad f. Sea $\nu \in (0,1)$ un valor fijado y sea $r_n(t) = \nu \hat{r}_0(t)$. Entonces,*

$$\mathbb{P}\left(C_{r_n(t)}(\mathcal{X}_n^+(t)) \subset G(t), \text{ eventualmente }\right) = 1.$$

Teorema 3.8. *Sea $G(t)$ un conjunto de nivel compacto, no vacío y no convexo. Bajo las asunciones (A) , (D) y (K), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de densidad f. Sea $\nu \in (0,1)$ un valor fijado y sea $r_n(t) = \nu \hat{r}_0(t)$. Entonces,*

$$\sup_{t\in[l,u]} d_H\left(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t)\right) = \mathcal{O}\left(\text{máx}\left\{D_n, \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}\right\}\right), \text{ de forma casi segura.}$$

El mismo orden de convergencia se mantiene para $d_\mu(C_{r_n(t)}(\mathcal{X}_n^+(t)), G(t))$.

Todas las demostraciones de los resultados anteriores se pueden consultar en Saavedra (2014).

Al principio del Capítulo 2 habíamos comentado que en la práctica resulta natural estimar, no el conjunto de nivel $G(t)$ para cierto valor t, sino el conjunto $L(\tau)$ para algún $\tau \in (0,1)$. En el siguiente teorema establecemos los órdenes de convergencia del estimador natural de $L(\tau)$, $\hat{L}(\tau)$ a través del estimador establecido anteriormente de $G(t)$.

Teorema 3.9. *Sea $G(t)$ un conjunto de nivel compacto, no vacío y no convexo. Bajo las asunciones (A) , (D) y (K), sea $\mathcal{X}_n$ una muestra aleatoria generada a partir de una distribución con función de densidad f. Sea $\nu \in (0,1)$ un valor fijado y sea $r_n(t) = \nu \hat{r}_0(t)$. Sea $\hat{f}_\tau$ el estimador de f_τ definido como*

$$\hat{f}_\tau = \text{máx}\left\{t : \mathbb{P}_n\left(C_{r_n(t)}(\mathcal{X}_n^+(t))\right) \geq 1-\tau\right\},$$

y $\mathbb{P}_n$ la medida de probabilidad empírica inducida por $\mathcal{X}_n$. Si $\underline{\tau} > \bar{\tau}$ son tales que $l < f_{\underline{\tau}}$, $f_{\bar{\tau}} < u$, y $\hat{L}(\tau) = C_{r_n(\hat{f}_n)}(\mathcal{X}_n^+(\hat{f}_n))$, entonces

$$\sup_{\tau\in[\underline{\tau},\bar{\tau}]} d_H\left(\hat{L}(\tau), L(\tau)\right) = \mathcal{O}\left(\text{máx}\left\{D_n, \left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}\right\}\right), \text{ de forma casi segura.}$$

El mismo orden de convergencia se mantiene para $d_\mu\left(\hat{L}(\tau), L(\tau)\right)$.

La demostración puede encontrarse en Rodríguez-Casal and Saavedra-Nieves (2019).

3.3. Algoritmo para la estimación de conjuntos de nivel

En este apartado propondremos el algoritmo revisado introducido en Rodríguez-Casal and Saavedra-Nieves (2019) para la estimación del conjunto de nivel $L(\tau)$. Para ello hemos de calcular los conjuntos $\mathcal{X}_n^+$ y $\mathcal{X}_n^-$, que dependen a su vez de la secuencia D_n. En el algoritmo no será necesario la estimación directa de D_n.

Dado un valor de $\tau \in (0,1)$, seleccionaremos el valor de $\bar{\tau}$ más grande, cumpliendo $0 < \bar{\tau} \leq \tau$, tal que el estimador de conjuntos de nivel asociado contenga una proporción de puntos mayor o igual que $1-\tau$.

Para un valor $\bar{\tau}$ se establecen los estimadores $\mathcal{X}_n^+(\hat{f}_{\bar{\tau}}^+)$ y $\mathcal{X}_n^-(\hat{f}_{\bar{\tau}}^-)$ a partir de la muestra $\mathcal{X}_n$, donde $\hat{f}_{\bar{\tau}}^+$ y $\hat{f}_{\bar{\tau}}^-$ son calculados mediante remuestreo bootstrap como se indica a continuación.

A partir de f_n se generan B remuestras bootstrap $\mathcal{X}_{1,n}^*, \ldots, \mathcal{X}_{B,n}^*$ de tamaño n. Para cada una, se determina la proporción de puntos dentro y fuera del conjunto de nivel $\{f_n \geq (f_n)_{\bar{\tau}}\}$. Para un valor fijo de un parámetro p se calculan dos valores $\bar{\tau}_+$ y $\bar{\tau}_-$ y a partir de ellos se calculan los umbrales $\hat{f}_{\bar{\tau}}^+$ y $\hat{f}_{\bar{\tau}}^-$ como se describe detalladamente en el algoritmo posterior. Si la proporción de puntos de $\mathcal{X}_n$ contenidos en $\hat{L}(\bar{\tau})$ es mayor o igual que $1-\tau$ finaliza el algoritmo. En otro caso se ha de considerar un valor más pequeño de $\bar{\tau}$. D_n podría ser computado como máx$\left\{\hat{f}_{\bar{\tau}}^+ - \hat{f}_\tau, \hat{f}_\tau - \hat{f}_{\bar{\tau}}^-\right\}$, pero esta computación explícita no es necesaria.

A continuación se presenta un algoritmo por dicotomía para el cálculo del parámetro $\hat{r}_0(\hat{f}_\tau)$:

Algoritmo 3.10. *Dados los argumentos de entrada: $\mathcal{X}_n^+(\hat{f}_\tau^+)$, $\mathcal{X}_n^-(\hat{f}_\tau^-)$, $r_m \in (0,\infty)$, $r_M \in (r_m,\infty)$ y $J \in \mathbb{N}$:*

1. *Para cada iteración y mientras el número de ellas sea menor que J:*
   a) $r = (r_m + r_M)/2$
   b) *Si $\mathcal{X}_n^-(\hat{f}_\tau^-) \cap C_r(\mathcal{X}_n^+(\hat{f}_\tau^+)) \neq \varnothing$, entonces $r_M = r$.*
   c) *En el otro caso, $r_m = r$.*
2. $\hat{r}_0(\hat{f}_\tau) = r_m$

Computacionalmente resulta sencillo el cálculo de las envolturas $r-$convexas para el caso bidimensional como se muestra en Pateiro-López and Rodríguez-Casal (2010).

A continuación presentamos de forma detallada el algoritmo para el cómputo del estimador del conjunto de nivel $L(\tau)$, $\hat{L}(\tau)$.

Algoritmo 3.11. *Dados los argumentos de entrada $\tau, \mathcal{X}_n, B, p$ y el paso s:*

1. *Calculamos el estimador de la densidad tipo núcleo f_n de $\mathcal{X}_n$.*
2. *Establecemos $P = 0$ y $\bar{\tau} = \tau$.*
3. *Mientras que $P < (1-\tau)$ hacemos:*
   - *Determinamos el umbral $(f_n)_{\bar{\tau}}$.*
   - *Para cada $i \in \{1, \ldots, B\}$:*
     - *Generamos una remuestra bootstrap $\mathcal{X}_{i,n}^*$ de tamaño n a partir de f_n.*
     - *Determinamos $f_n(\mathcal{X}_{i,n}^*)$.*
     - *Calculamos los parámetros*

$$\bar{\tau}_{i,+}^* = \frac{\#\{f_n(\mathcal{X}_{i,n}^*) \geq (f_n)_{\bar{\tau}}\}}{n} \text{ y } \bar{\tau}_{i,-}^* = \frac{\#\{f_n(\mathcal{X}_{i,n}^*) < (f_n)_{\bar{\tau}}\}}{n}.$$

   - *Determinamos $\bar{\tau}_-$ y $\bar{\tau}_+$ como los $p-$cuantiles de los vectores $\bar{\tau}_{i,-}^*$ y $\bar{\tau}_{i,+}^*$ respectivamente.*
   - *Calculamos los umbrales $\hat{f}_{\bar{\tau}}^-$ y $\hat{f}_{\bar{\tau}}^+$ como los $\bar{\tau}_-$ y $(1-\bar{\tau}_+)-$cuantiles de $f_n(\mathcal{X}_n)$, respectivamente.*

- *Obtenemos los subconjuntos de* $\mathcal{X}_n$*:*

$$\mathcal{X}_n^+(\hat{f}_{\bar{\tau}}^+) = \{X \in \mathcal{X}_n : f_n(X) \geq \hat{f}_{\bar{\tau}}^+\} \text{ y } \mathcal{X}_n^-(\hat{f}_{\bar{\tau}}^-) = \{X \in \mathcal{X}_n : f_n(X) < \hat{f}_{\bar{\tau}}^-\}.$$

- *Computamos* $\hat{r}_0(\hat{f}_{\bar{\tau}})$ *a partir de los conjuntos establecidos en el paso anterior mediante el algoritmo 3.10.*
- *Estimamos* $L(\bar{\tau})$ *como* $\hat{L}(\bar{\tau}) = C_{\hat{r}_0(\hat{f}_{\bar{\tau}})}(\mathcal{X}_n^+(\hat{f}_{\bar{\tau}}^+))$.
- *Actualizamos:*
  - P *como la proporción de puntos de* $\mathcal{X}_n$ *en* $\hat{L}(\bar{\tau})$.
  - $\bar{\tau} = \bar{\tau} - s$

4. El estimador obtenido es $\hat{L}(\bar{\tau})$.

Para el cálculo del parámetro $\hat{r}_0(\hat{f}_{\bar{\tau}})$ tenemos que indicar qué parámetros J , r_m y r_M utilizaremos. Es natural establecer el valor r_m lo más pequeño posible, por lo que utilizaremos $r_m = 0$. Para que no pueda haber el problema de seleccionar un $\hat{r}_0 = 0$ haremos que, en caso de que $\hat{r}_0(\hat{f}_{\bar{\tau}}) = 0$ en el paso 2 del algoritmo 3.10, cambiaremos el valor por $\hat{r}_0(\hat{f}_{\bar{\tau}}) = (r_m + r_M)/2$. Para el valor de r_M cogeremos un valor tal que $C_{r_M(\hat{f}_{\bar{\tau}})}(\mathcal{X}_n)$ sea casi igual a la envoltura convexa $H(\mathcal{X}_n)$. Para ello cogeremos un valor excesivamente grande. En particular, en el caso bidimensional escogeremos el mayor valor de las componentes x de la muestra menos el mínimo de estas y le sumaremos el máximo valor de las componentes y menos la menor. Hecho esto es necesario un valor de J que garantice que ante un r_M muy grande se siga ajustando bien. Cogeremos el valor $J = 60$ ya que un mayor valor de J implica solo un pequeño coste computacional mayor sobre el algoritmo 3.11, y el error sobre $\hat{r}_0$ será menor que $\frac{r_M}{2^{60}}$, lo cual nos garantiza una precisión de siete cifras decimales para $r_M < 10^{12}$. En el próximo Capítulo estudiaremos mediante simulaciones el comportamiento del algoritmo 3.11 en función de sus parámetros de entrada y del tamaño muestral.

Aplicando el algoritmo a la muestra de octubre del año 2015 de los datos de la Velutina en Galicia con los parámetros $B = 250, p = 0.25, s = 0.01$ y $\tau = 0.3$ obtenemos un valor $\hat{r}_0(\hat{f}_\tau) = 5386.2920$. En las Figura 3.2 y 3.3 se muestra la estimación $\hat{L}(\tau)$ obtenida. Puede resultar interesante compararla con las Figuras 2.7, 2.8 y 2.9 de la Sección 2.2.2.

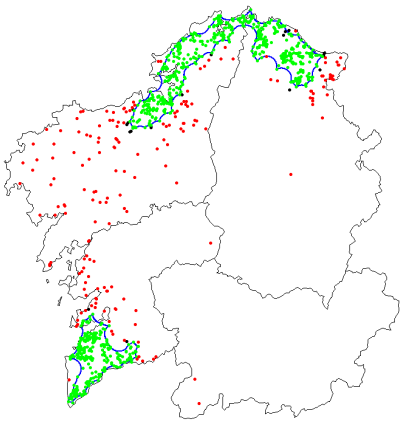


Figura 3.2: El conjunto $\mathcal{X}_n^+(\hat{f}_{0.3})$ en verde , $\mathcal{X}_n^-(\hat{f}_{0.3})$ en rojo y $\mathcal{X}_n \setminus \left(\mathcal{X}_n^+(\hat{f}_{0.3}) \cup \mathcal{X}_n^-(\hat{f}_{0.3})\right)$ con puntos negros, obtenidos sobre los datos de la velutina del mes de octubre del año 2015. En azul aparece representado el conjunto de nivel $\hat{L}(0.3) = C_{5386.2920}(\mathcal{X}_n^+(\hat{f}_{0.3}))$ obtenido mediante el algoritmo 3.11.

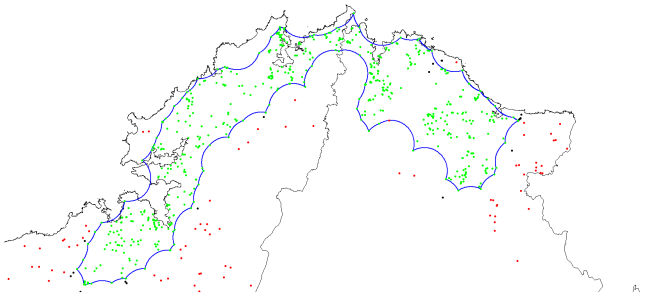


Figura 3.3: El conjunto $\mathcal{X}_n^+(\hat{f}_{0.3})$ en verde , $\mathcal{X}_n^-(\hat{f}_{0.3})$ en rojo y $\mathcal{X}_n \setminus \left(\mathcal{X}_n^+(\hat{f}_{0.3}) \cup \mathcal{X}_n^-(\hat{f}_{0.3})\right)$ con puntos negros para la zona norte de Galicia obtenidos sobre los datos de la velutina del mes de octubre del año 2015. En azul aparece representado el conjunto de nivel $\hat{L}(0.3) = C_{5386.2920}(\mathcal{X}_n^+(\hat{f}_{0.3}))$ obtenido mediante el algoritmo 3.11.

Capítulo 4

Estudio de simulación

El algoritmo descrito en Rodríguez-Casal and Saavedra-Nieves (2019) para estimar $L(\tau)$ depende de los los parámetros B , p , s y de la muestra $\mathcal{X}_n$. Es, por lo tanto, interesante estudiar cómo se comporta el algoritmo en función de estos valores. Para las simulaciones utilizaremos cuatro modelos distintos, obtenidos como mixturas de normales, extraídos de Wand and Jones (1995). Se trata de dos modelos bimodales y dos modelos trimodales con distintas separaciones de las modas y con distintos valores de la función de densidad en estas. En la Figura 4.1 aparecen representados los cuatro modelos, mientras que en el Cuadro 4.1 aparecen descritos los cuatro modelos, siendo cada uno representado de la forma:

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

con $k = 2$ para los dos primeros modelos y $k = 3$ para los dos últimos, donde w es un vector de pesos tal que $\sum_{i=1}^{k} w_i = 1$, μ_{ij} es la componente j del vector de medias de la i-ésima normal de la mixtura , σ_{ij}^2 es la componente j de la diagonal de la matriz de covarianza de la i-ésima normal y ρ_i el coeficiente de correlación de la i-ésima normal.

En la Sección 4.1 presentaremos los resultados para un estudio sobre la influencia de los parámetros B, p y s. Para ello, fijando una muestra y τ , calcularemos el estimador $\hat{L}(\tau)$ y $d_H(\hat{L}(\tau), L(\tau))$, variando los tres parámetros sobre los que vamos a realizar el estudio. En la Sección 4.2 estudiaremos la influencia del tamaño muestral en el algoritmo, calculando la media de las distancias de Hausdorff entre $\hat{L}(\tau)$ y $L(\tau)$ a través de muchas muestras con el mismo tamaño muestral y utilizando los mismos parámetros B , p y s; para varios tamaños muestrales, observando así una clara disminución de la distancia de Hausdorff a medida que aumenta el valor de n. También comprobaremos si el algoritmo es capaz de detectar las distintas componentes conexas estudiando la proporción de veces que se estiman bien. Por último, en la Sección 4.3 estableceremos las conclusiones de los resultados obtenidos en el estudio de simulación. Para todos los resultados utilizaremos los valores de J, r_m y r_M descritos al final del Capítulo 3.

4.1. Influencia de los parámetros de entrada

Para cada modelo empezaremos por comprobar cómo influyen los parámetros B , s y p sobre una misma muestra para muestras de distinto tamaño n y para distintos valores de τ, a la hora de calcular la distancia de Hausdorff entre $L(\tau)$ y $\hat{L}(\tau)$. Más específicamente, para cada modelo crearemos 5 muestras,

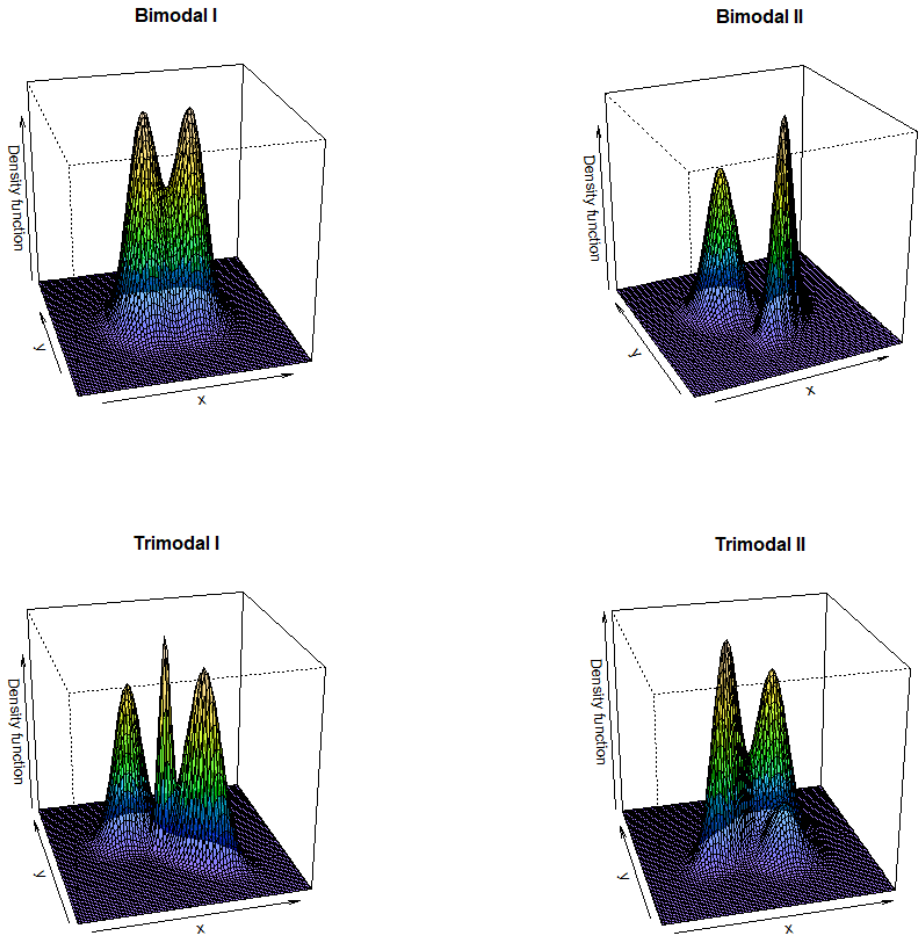


Figura 4.1: Representación gráfica de los cuatro modelos.

cada una de tamaño $n = \{250, 500, 1000, 2500, 5000\}$ y dados $\tau \in \{0.3, 0.5, 0.7, 0.9\}$, $B \in \{50, 100, 250\}$, $p \in \{0.1, 0.25, 0.5\}$ y $s \in \{0.01, 0.04\}$ calcularemos $\hat{L}(\tau)$ mediante el algoritmo 3.11, simularemos una gran cantidad de puntos de su frontera y calcularemos la distancia de Hausdorff entre estos puntos y la frontera de $L(\tau)$.

En los Cuadros 4.10, 4.11, 4.12 y 4.13 al final del Capítulo se pueden observar los resultados obtenidos para el modelo Bimodal I. En la Figura 4.7 aparecen representados gráficamente los conjuntos estimados para $B = 250$, $p = 0.25$, $s = 0.01$, $n \in \{250, 100, 5000\}$ y $\tau \in \{0.9, 0.7, 0.5, 0.3\}$. Por lo general no parece que el valor de B ni de p influyan significativamente en los resultados. La mayor variación se da en el caso de $\tau = 0.9$, $n = 250$ y $s = 0.01$, ya que con $p = 0.5$ se obtienen mejores resultados. Esto se debe a que para este valor de p es capaz de detectar las dos modas mientras que en el resto de los casos detecta una única componente conexa como se puede observar en la Figura 4.2. Aún así, por lo general, no parece influir, sobre todo para valores mayores de n, donde los resultados se vuelven más estables. No podemos hacer una comparativa de los resultados con distinto tamaño muestral debido a que los resultados se pueden ver fuertemente influidos por la muestra concreta. En

Modelo	$\sum_{i=1}^{k} w_i \; N\left(\mu_{i1}, \mu_{i2}, \sigma_{i1}^2, \sigma_{i2}^2, \rho_i\right)$
Bimodal I	$\frac{1}{2}N\left(-1, 0, \left(\frac{2}{3}\right)^2, \left(\frac{2}{3}\right)^2, 0\right) + \frac{1}{2}N\left(1, 0, \left(\frac{2}{3}\right)^2, \left(\frac{2}{3}\right)^2, 0\right)$
Bimodal II	$\frac{1}{2}N\left(1, -1, \left(\frac{2}{3}\right)^2, \left(\frac{2}{3}\right)^2, \frac{7}{10}\right) + \frac{1}{2}N\left(-1, 1, \left(\frac{2}{3}\right)^2, \left(\frac{2}{3}\right)^2, 0\right)$
Trimodal I	$\frac{9}{20}N\left(-\frac{6}{5}, \frac{6}{5}, \left(\frac{3}{5}\right)^2, \left(\frac{3}{5}\right)^2, \frac{3}{10}\right) + \frac{9}{20}N\left(\frac{6}{5}, -\frac{6}{5}, \left(\frac{3}{5}\right)^2, \left(\frac{3}{5}\right)^2, -\frac{3}{5}\right) + \frac{1}{10}N\left(0, 0, \left(\frac{1}{4}\right)^2, \left(\frac{1}{4}\right)^2, \frac{1}{5}\right)$
Trimodal II	$\frac{3}{7}N\left(-1, 0, \left(\frac{3}{5}\right)^2, \left(\frac{7}{10}\right)^2, \frac{3}{5}\right) + \frac{3}{7}N\left(1, \frac{2\sqrt{3}}{3}, \left(\frac{3}{5}\right)^2, \left(\frac{7}{10}\right)^2, 0\right) + \frac{1}{7}N\left(1, -\frac{2\sqrt{3}}{3}, \left(\frac{3}{5}\right)^2, \left(\frac{7}{10}\right)^2, 0\right)$

Cuadro 4.1: Parámetros de los 4 modelos obtenidos como mixtura de normales.

un estudio posterior estudiaremos este efecto sobre el resultado. Para menores valores de τ también se observan mejores resultados en general. Esto es normal ya que el algoritmo funciona mejor cuanto menor sea el valor de τ, sobre todo cuando $L(\tau)$ consiste en una única componente conexa, o cuando a medida que se disminuye τ no aparecen modas nuevas.

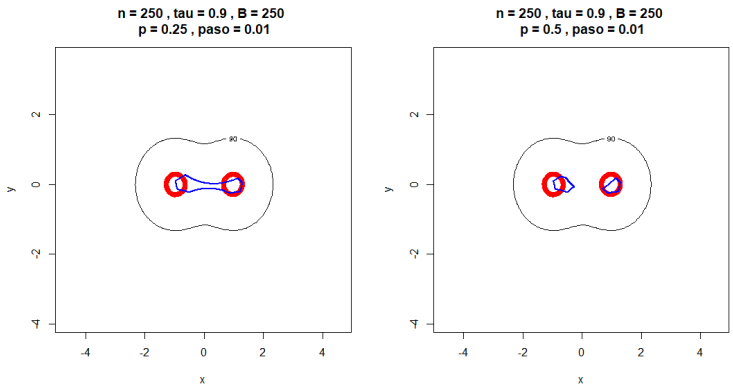


Figura 4.2: En rojo la frontera del conjunto $L(\tau)$ del modelo Bimodal I para $\tau = 0.9$. En azul la frontera del conjunto de nivel estimado $\hat{L}(\tau)$ para el mismo modelo y el mismo valor de τ , con el valor $B = 250$, $s = 0.01$, el modelo de tamaño $n = 250$ y $p = 0.25$ en la primera gráfica y $p = 0.5$ en la segunda.

En los Cuadros 4.14, 4.15, 4.16 y 4.17 al final del Capítulo se pueden observar los resultados obtenidos para el modelo Bimodal II. En la Figura 4.8 aparecen representados gráficamente los conjuntos estimados para $B = 250$, $p = 0.25$, $s = 0.01$, $n \in \{250, 100, 5000\}$ y $\tau \in \{0.9, 0.7, 0.5, 0.3\}$. Para $\tau = 0.9$, el modelo de tamaño $n = 250$ no estima bien $L(\tau)$ debido a que detecta una componente conexa en cada una de las dos modas, cuando debería detectar una única componente conexa. Esto se puede observar con claridad en la Figura 4.3. En la siguiente sección estudiaremos con qué frecuencia ocurre esto. No parece que influya significativamente la variación de los valores de B, p y s para ningún caso.

En los Cuadros 4.18, 4.19, 4.20 y 4.21 al final del Capítulo se pueden observar los resultados obtenidos para el modelo Trimodal I. En la Figura 4.9 aparecen representados gráficamente los conjuntos estimados para $B = 250$, $p = 0.25$, $s = 0.01$, $n \in \{250, 100, 5000\}$ y $\tau \in \{0.9, 0.7, 0.5, 0.3\}$. Los resultados para los cinco modelos con $\tau = 0.9$ no son del todo satisfactorios. Como se observa en la Figura 4.4, esto es debido a que $L(0.89)$ tiene una tercera componente conexa en la tercera moda que $L(0.9)$ no tiene. Aún así, de aquí no podemos inferir cómo afecta el tamaño muestral a la buena

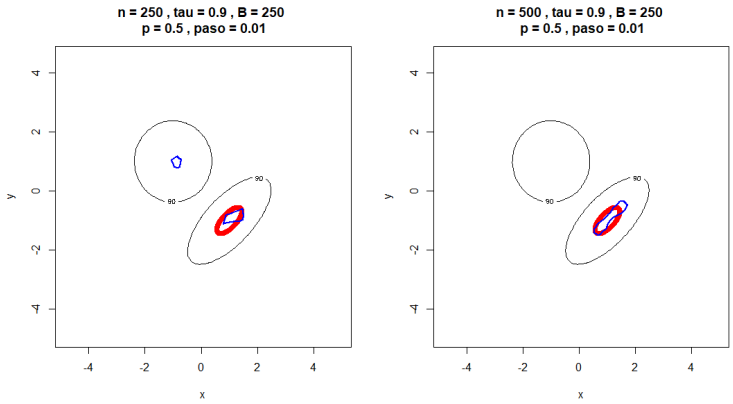


Figura 4.3: En rojo la frontera del conjunto $L(\tau)$ del modelo Bimodal II para $\tau = 0.9$. En azul la frontera del conjunto de nivel estimado $\hat{L}(\tau)$ para el mismo modelo y el mismo valor de τ , con el valor $B = 250$, $s = 0.01$ y $p = 0.5$ y el modelo de tamaño $n = 250$ en la primera gráfica y el de tamaño $n = 0.5$ en la segunda.

detección, en este caso, de la no existencia de una tercera componente conexa pues puede verse fuertemente influenciado por la muestra concreta como ya comentamos anteriormente. En la Sección 4.2 reflejaremos esto.
Como en los anteriores modelos, parece que no influyen de manera significativa los valores de B y de p. El cambio de valor en s tampoco parece influir mucho salvo en algún caso como el mostrado en la Figura 4.5. Esto se debe a que al ser el paso mayor, $\hat{L}(\tau)_{s=0.01} \subset \hat{L}(\tau)_{s=0.04}$ siempre, por lo que ante componentes conexas cercanas intenta unificar tanto o más las distintas componentes.

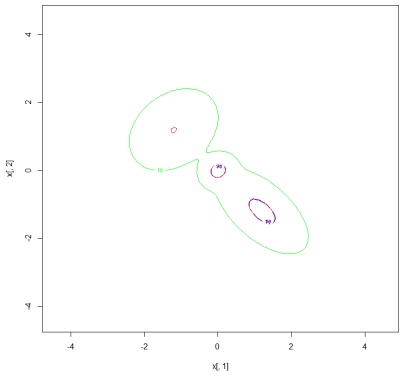


Figura 4.4: En verde la frontera de $L(0.1)$, en rojo la frontera de $L(0.89)$ y en azul la frontera de $L(0.9)$ para el modelo Trimodal I.

Por último, en los Cuadros 4.22, 4.23, 4.24 y 4.25 al final del Capítulo se pueden observar los resultados obtenidos para el modelo Trimodal II. En la Figura 4.10 aparecen representados gráficamente los conjuntos estimados para $B = 250$, $p = 0.25$, $s = 0.01$, $n \in \{250, 100, 5000\}$ y $\tau \in \{0.9, 0.7, 0.5, 0.3\}$. En este último modelo tampoco parece que los valores de B , p y s, tengan mucha relevancia. Las distancias de Hausdorff de la muestra de tamaño $n = 500$ en el caso de $\tau = 0.9$ y $n = 250$ en el caso de $\tau = 0.3$ dan malos resultados debido a que no son capaces de detectar una segunda y una tercera componente conexa respectivamente, como se muestra en la Figura 4.6.

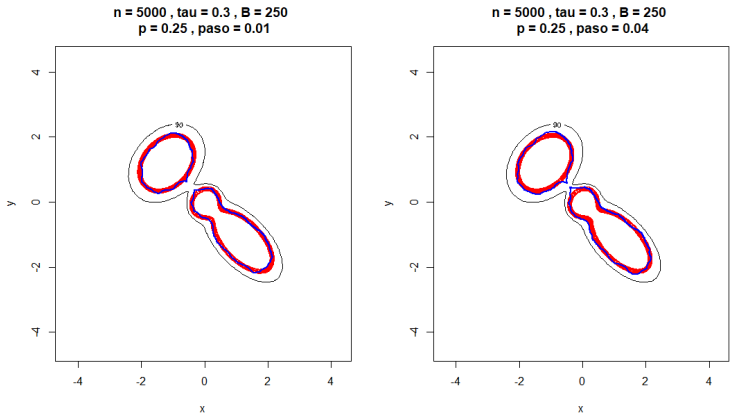


Figura 4.5: En rojo la frontera del conjunto $L(\tau)$ del modelo Trimodal I para $\tau = 0.3$. En azul el conjunto de nivel estimado $\hat{L}(\tau)$ para el mismo valor de τ , con el modelo de tamaño $n = 5000$, $B = 250$, $p = 0.25$ y el valor de $s = 0.01$ para la gráfica de la izquierda y de $s = 0.04$ para la gráfica de la derecha

.

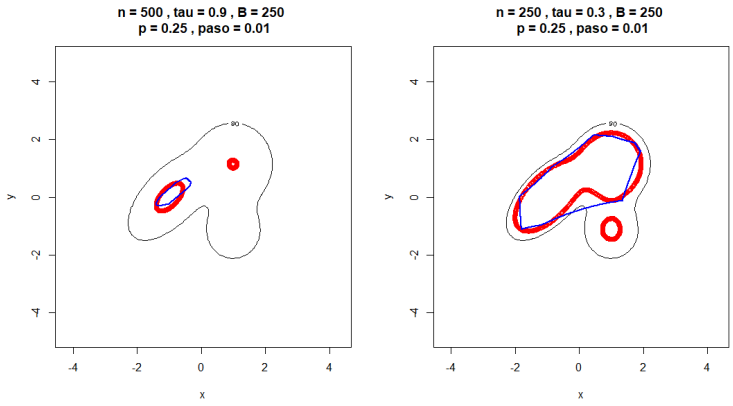


Figura 4.6: En rojo la frontera del conjunto $L(\tau)$ del modelo Trimodal II para $\tau = 0.9$ en el gráfico de la izquierda y $\tau = 0.3$ en el gráfico de la derecha. En azul el conjunto de nivel estimado $\hat{L}(\tau)$ para los mismos valores de τ , con el modelo de tamaño $n = 500$, $B = 250$, $p = 0.25$ y $s = 0.01$ en el gráfico de la izquierda y el modelo de tamaño $n = 250$, $B = 250$, $p = 0.25$ y $s = 0.01$ para la gráfica de la derecha

.

Parece ser, por lo tanto, que en los modelos presentados no tiene a priori relevancia significativa el valor de p que escojamos, siempre y cuando no sea muy alto ya que puede provocar errores en el algoritmo a la hora de generar la sucesión D_n. Con respecto al valor de las remuestras bootstrap B, tampoco parece que merezca la pena el coste computacional añadido de pasar de $B = 100$ a $B = 250$. Aún así hay que tener en cuenta que valores excesivamente bajos de B sí provocan resultados irregulares. Con respecto al valor del paso s, tener en cuenta que, aún en general siendo los resultados muy parecidos con $s = 0.04$ y con $s = 0.01$, en el primer caso la proporción de puntos de la muestra contenidos en el conjunto $\hat{L}(\tau)$ es siempre igual o mayor que en el segundo caso, por lo que los resultados serán igual o peores.

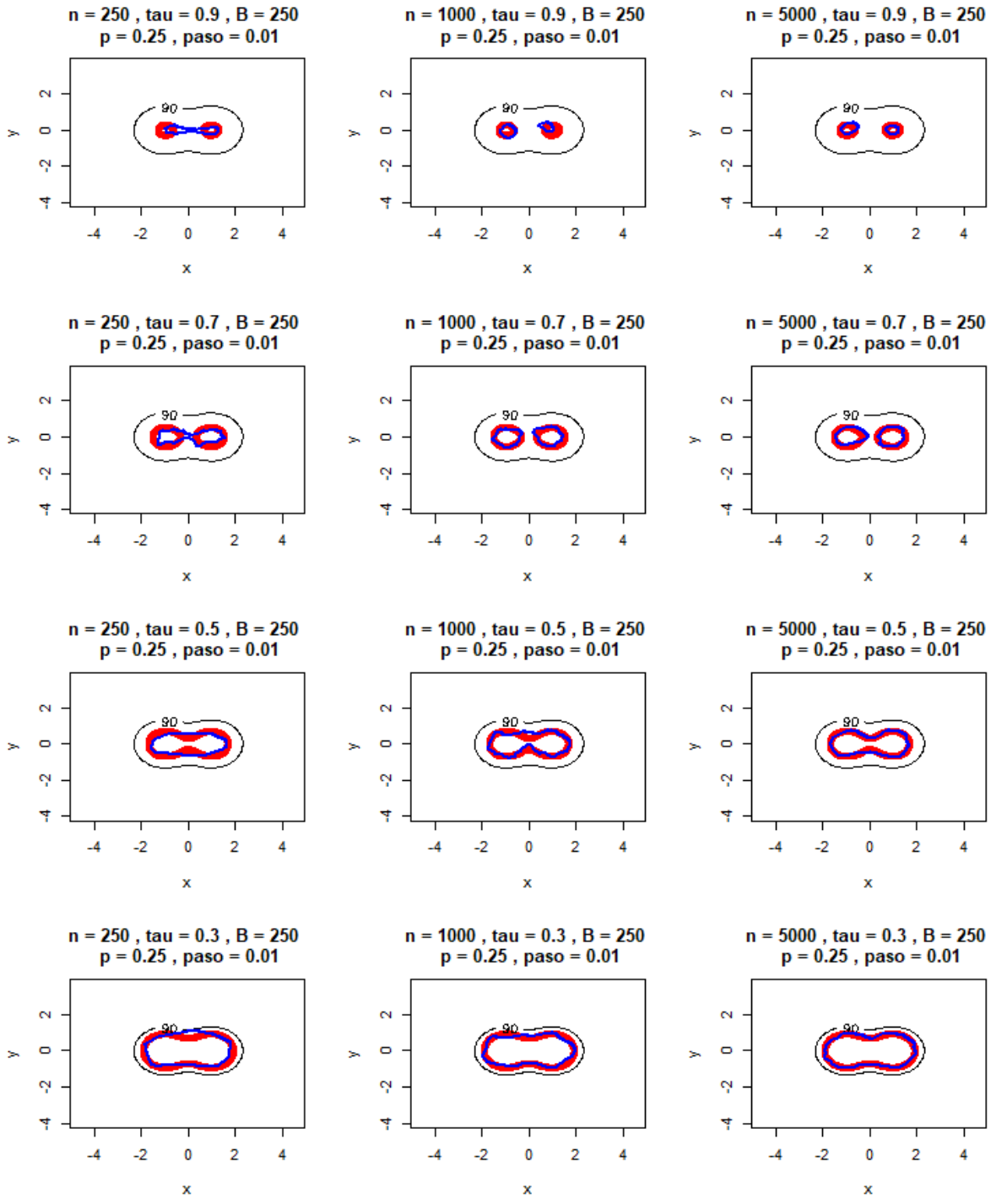


Figura 4.7: En rojo la frontera del conjunto $L(\tau)$ del modelo Bimodal I para $\tau = 0.9$ en la primera fila, $\tau = 0.7$ en la segunda fila, $\tau = 0.5$ en la tercera fila y $\tau = 0.3$ en la cuarta fila. En azul la frontera del conjunto de nivel estimado $\hat{L}(\tau)$ del mismo modelo para los mismos valores de τ , con el valor $B = 250$, $s = 0.01$, $p = 0.25$ y el modelo de tamaño $n = 250$ en la primera columna , $n = 1000$ en la segunda columna y $n = 5000$ en la tercera columna.

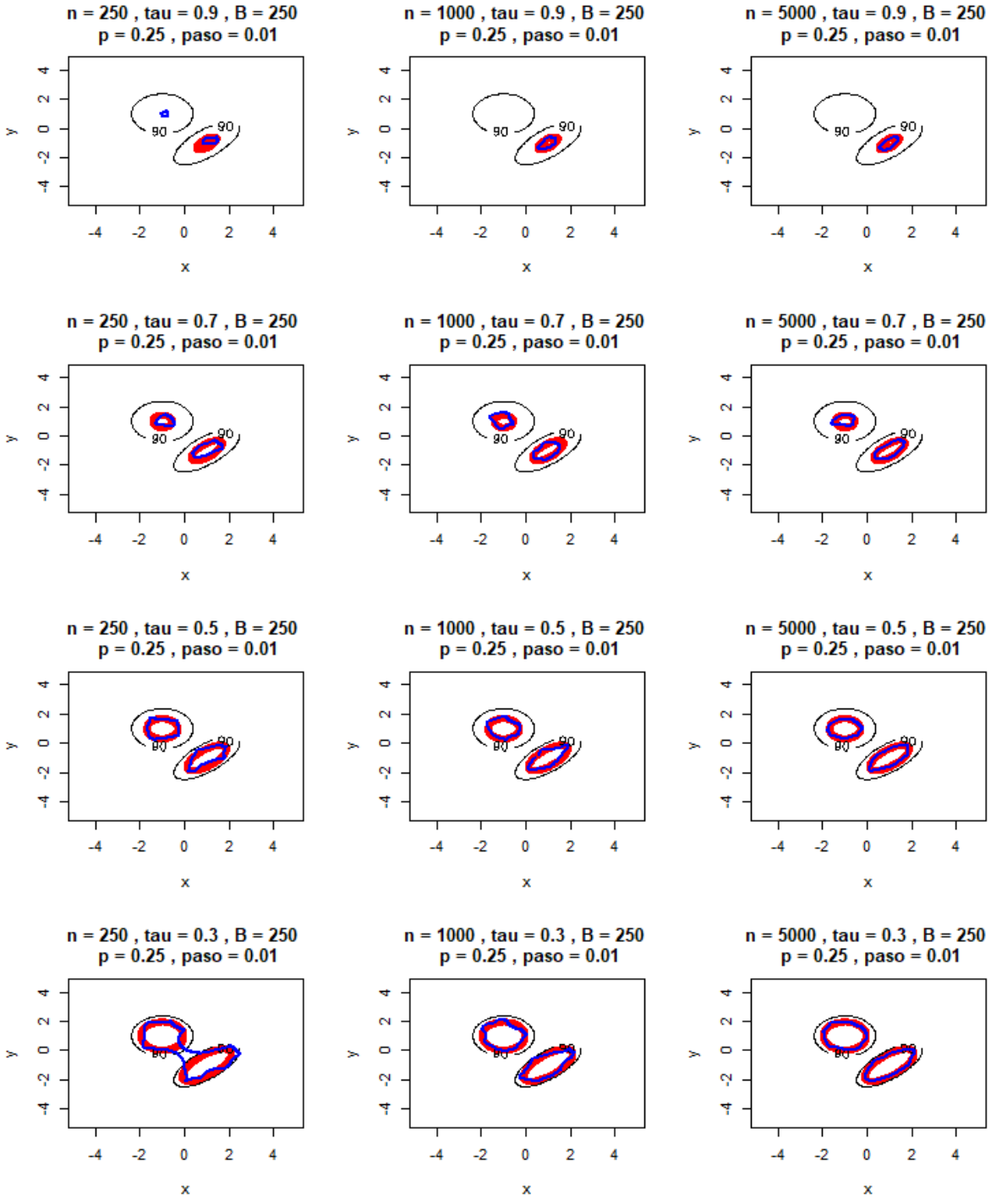


Figura 4.8: En rojo la frontera del conjunto $L(\tau)$ del modelo Bimodal II para $\tau = 0.9$ en la primera fila, $\tau = 0.7$ en la segunda fila, $\tau = 0.5$ en la tercera fila y $\tau = 0.3$ en la cuarta fila. En azul la frontera del conjunto de nivel estimado $\hat{L}(\tau)$ del mismo modelo para los mismos valores de τ , con el valor $B = 250$, $s = 0.01$, $p = 0.25$ y el modelo de tamaño $n = 250$ en la primera columna , $n = 1000$ en la segunda columna y $n = 5000$ en la tercera columna.

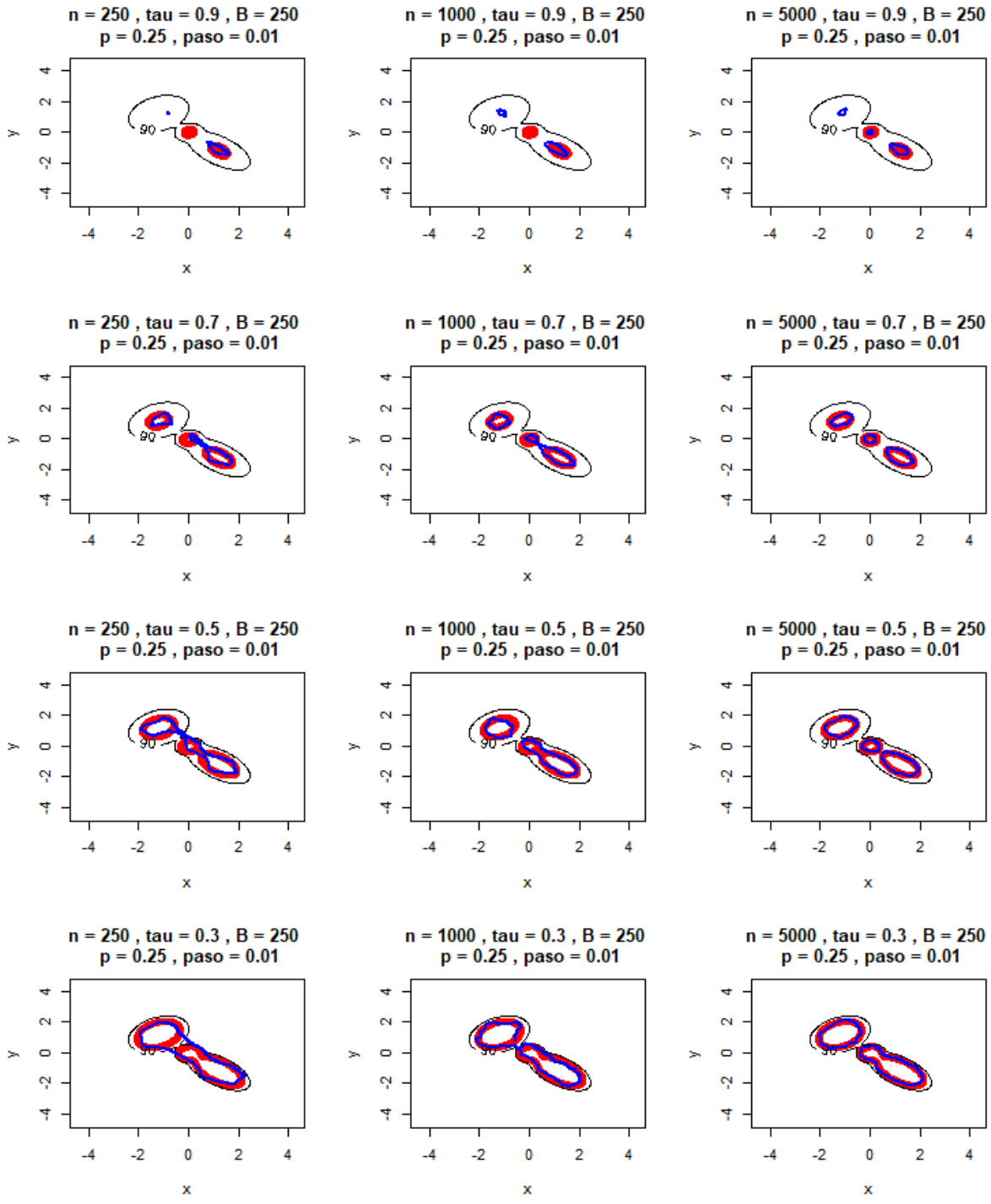


Figura 4.9: En rojo la frontera del conjunto $L(\tau)$ del modelo Trimodal I para $\tau = 0.9$ en la primera fila, $\tau = 0.7$ en la segunda fila, $\tau = 0.5$ en la tercera fila y $\tau = 0.3$ en la cuarta fila. En azul la frontera del conjunto de nivel estimado $\hat{L}(\tau)$ del mismo modelo para los mismos valores de τ , con el valor $B = 250$, $s = 0.01$, $p = 0.25$ y el modelo de tamaño $n = 250$ en la primera columna , $n = 1000$ en la segunda columna y $n = 5000$ en la tercera columna.

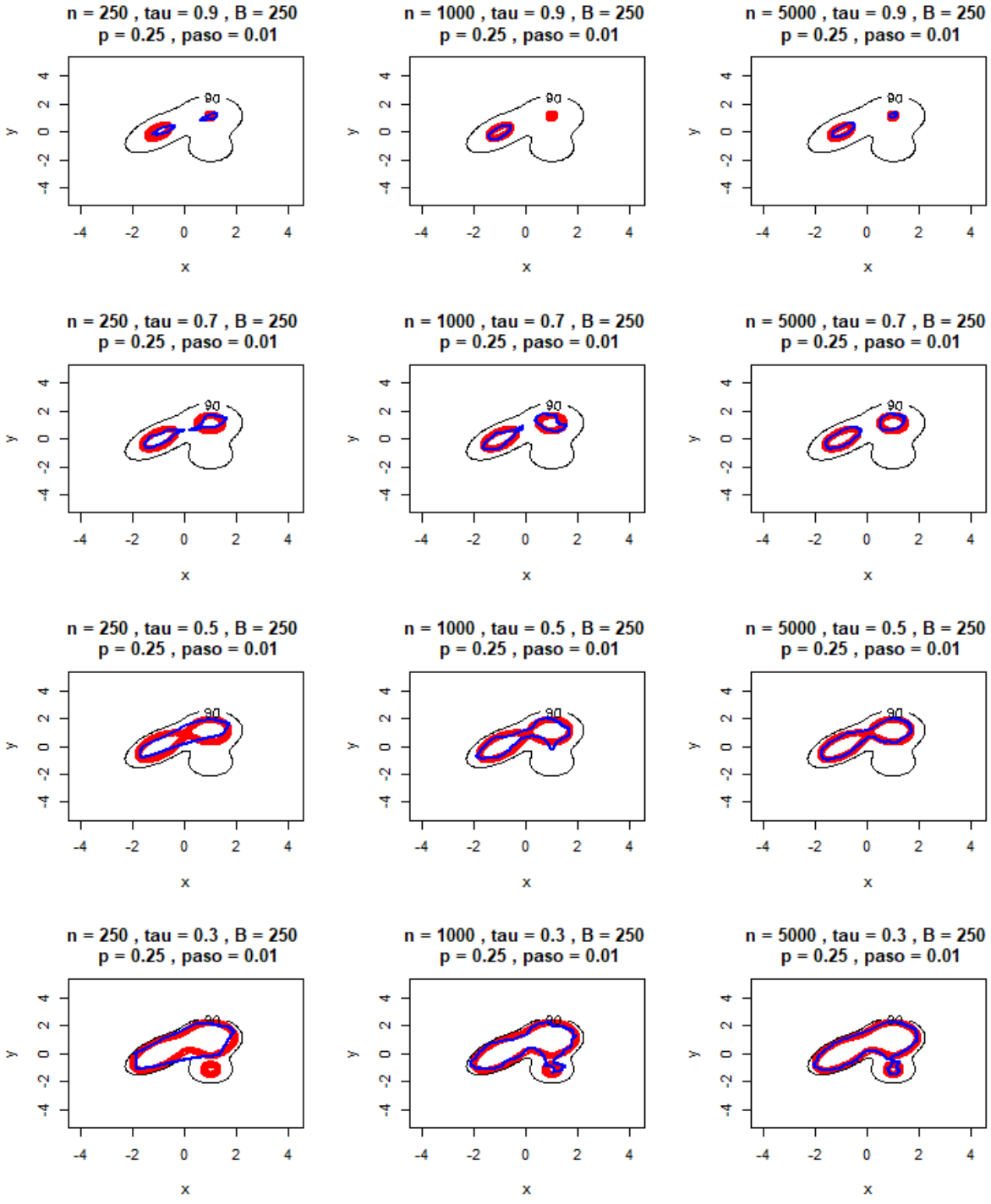


Figura 4.10: En rojo la frontera del conjunto $L(\tau)$ del modelo Trimodal II para $\tau = 0.9$ en la primera fila, $\tau = 0.7$ en la segunda fila, $\tau = 0.5$ en la tercera fila y $\tau = 0.3$ en la cuarta fila. En azul la frontera del conjunto de nivel estimado $\hat{L}(\tau)$ del mismo modelo para los mismos valores de τ , con el valor $B = 250$, $s = 0.01$, $p = 0.25$ y el modelo de tamaño $n = 250$ en la primera columna , $n = 1000$ en la segunda columna y $n = 5000$ en la tercera columna.

4.2. Influencia del tamaño muestral

Una vez estudiado cómo afectan los parámetros de entrada, resulta interesante estudiar la eficacia del algoritmo en función del tamaño muestral n. Para ello generaremos sobre cada uno de los cuatro modelos anteriores 300 muestras para cada tamaño $n \in \{250, 1000, 2500\}$ y le aplicaremos el algoritmo con los parámetros $B = 100$, $p = 0.25$ y $s = 0.01$ para cada $\tau \in \{0.3, 0.5, 0.7, 0.9\}$. Hecho esto calcularemos la distancia de Hausdorff entre la frontera de $L(\tau)$ y la frontera de $\hat{L}(\tau)$ para así calcular la distancia media. También calcularemos el número de componentes conexas formadas en cada $\hat{L}(\tau)$ para compararlo con el número real de componentes conexas de $L(\tau)$. En el Cuadro 4.2 se pueden observar los resultados obtenidos con respecto a la distancia de Hausdorff en el modelo Bimodal I. Se observa una clara mejoría en el ajuste al aumentar el tamaño muestral como era de esperar, al igual que al disminuir el valor de τ. El algoritmo establecido por lo general ajusta un mejor modelo cuanto menor sea el valor de τ. Esto no ocurre siempre pues depende fuertemente del número de componentes conexas que forman el conjunto de nivel a estimar. En el Cuadro 4.3 se observan los resultados con respecto a la proporción de componentes conexas estimadas correctamente sobre el total. El algoritmo tiende a unir componentes conexas cercanas, esto se observa bien para el valor de $\tau = 0.7$. Debido a la cercanía de las dos modas, más de la mitad de las veces las junta en una sola, incluso para el valor de $n = 2500$. Realizando en este caso 100 simulaciones con un tamaño muestral de $n = 10000$ obtenemos un resultado de una proporción de 0.96 de cantidad de veces que se estima correctamente la cantidad de componentes conexas.

n \ τ	0.9	0.7	0.5	0.3
250	0.6435	0.461	0.4032	0.3431
1000	0.3661	0.3633	0.2579	0.2153
2500	0.2915	0.315	0.1805	0.1672

Cuadro 4.2: Media de las distancia de Hausdorff entre $\hat{L}(\tau)$ y $L(\tau)$ para 300 muestras con cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Bimodal I.

τ \ n	250	1000	2500	Nº de componentes conexas reales
0.9	0.5833	0.8667	0.9633	2
0.7	0.1967	0.3233	0.5033	2
0.5	0.9867	0.9867	0.9933	1
0.3	1	1	1	1

Cuadro 4.3: Proporción de las estimaciones de $\hat{L}(\tau)$ de las 300 muestras generadas para cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Bimodal I que están formadas por el mismo número de componentes conexas que $L(\tau)$. En la última columna se indica este último valor.

En el Cuadro 4.4 se observan los resultados para el modelo Bimodal II. De nuevo, como era de

esperar, se disminuye la distancia de Hausdorff al aumentar el valor del tamaño muestral n. Los resultados para $\tau = 0.9$ no son del todo satisfactorios debido a que, como se puede observar en el Cuadro 4.5, estima mal el número de componentes conexas que, en este caso, están relativamente alejadas entre sí, al detectar en muchos casos una segunda componente en la segunda moda, lo cual provoca una distancia entre conjuntos bastante grande. Con un paso s más pequeño los modelos estimarían mejor pues, cuanto mayor sea s, mayor resulta la proporción P final del algoritmo 3.11. Un ejemplo de esto se podía observar en el conjunto estimado de la izquierda de la Figura 4.3. Realizando para este valor $\tau = 0.9$, 100 simulaciones con un tamaño muestral de $n = 5000$ obtenemos un resultado de una proporción de 0.73 veces que se estima de forma correcta la presencia de una única componente conexa.

n \ τ	0.9	0.7	0.5	0.3
250	1.8816	0.4321	0.5373	0.6218
1000	1.8982	0.241	0.2468	0.336
2500	1.1716	0.1845	0.2086	0.2175

Cuadro 4.4: Media de las distancia de Hausdorff entre $\hat{L}(\tau)$ y $L(\tau)$ para 300 muestras con cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Bimodal II.

τ \ n	250	1000	2500	Nº de componentes conexas reales
0.9	0.28	0.3567	0.5967	1
0.7	0.9633	1	0.99	2
0.5	0.8533	0.9933	0.9967	2
0.3	0.2167	0.9233	0.98	2

Cuadro 4.5: Proporción de las estimaciones de $\hat{L}(\tau)$ de las 300 muestras generadas para cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Bimodal II que están formadas por el mismo número de componentes conexas que $L(\tau)$. En la última columna se indica este último valor.

En el Cuadro 4.6 se observan los resultados obtenidos para el modelo Trimodal I. Como en los dos casos anterior, el aumento del parámetro n provoca una disminución de la distancia de Hausdorff media entre $L(\tau)$ y $\hat{L}(\tau)$. Puede resultar curioso que a la hora de estimar la proporción de componentes conexas, para $\tau = 0.9$ se haga mejor con una muestra más pequeña, pero esto no es del todo correcto ya que, aún detectando más veces 2 componentes conexas el conjunto estimado cuando $n = 250$, las dos componentes que halla no son las dos que debería. Con $n = 2500$ aunque la proporción de casos en las que estima 2 componentes es menor, es mayor la proporción de veces que las estima de forma correcta (aún así sigue habiendo veces que estima la moda que no es, incluso más veces que con $n = 250$, solo que suele estimarla junto a las dos correctas, interpretando así que $L(0.9)$ esta formado por tres componentes conexas). Esto ocurría en el ejemplo presentado en la Figura 4.9. La razón la habíamos explicado en la sección anterior, se debía a la existencia real de una tercera componente conexa en $L(0.89)$, como se podía ver en la Figura 4.5. Para el valor de $\tau = 0.5$ ocurre lo contrario. Se tienen tres

componentes conexas con dos de ellas muy cercanas y, debido a que la envoltura $r-$convexa generada por el algoritmo tiende a unificar componentes cercanas, no es capaz en general de estimar la existencia de las tres distintas componentes.

n \ τ	0.9	0.7	0.5	0.3
250	1.5417	0.4737	0.4447	0.5351
1000	1.9623	0.3128	0.2845	0.3304
2500	1.5187	0.201	0.2234	0.2575

Cuadro 4.6: Media de las distancia de Hausdorff entre $\hat{L}(\tau)$ y $L(\tau)$ para 300 muestras con cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Trimodal I.

τ \ n	250	1000	2500	Nº de componentes conexas reales
0.9	0.5967	0.34	0.25	2
0.7	0.17	0.6533	0.9833	3
0.5	0.0367	0.0167	0.1	3
0.3	0.0033	0.0967	0.5467	2

Cuadro 4.7: Proporción de las estimaciones de $\hat{L}(\tau)$ de las 300 muestras generadas para cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Trimodal I que están formadas por el mismo número de componentes conexas que $L(\tau)$. En la última columna se indica este último valor.

Por último, el el Cuadro 4.8 están los resultados realizados sobre el modelo Trimodal II. La media de las distancias de Hausdorff también disminuye como era esperable con el aumento del tamaño muestral. Los resultados a la hora de estimar el número de componentes conexas de $L(\tau)$ también son satisfactorios para $\tau \in \{0.9, 0.7, 0.5\}$ como se puede ver en el Cuadro 4.9. Para el valor de $\tau = 0.3$ falla bastante más el algoritmo por razones similares a las que provocaban la mala estimación en el caso anterior sobre el modelo Trimodal I con $\tau = 0.5$.

4.3. Conclusiones del estudio

El método propuesto por Rodríguez-Casal and Saavedra-Nieves (2019) parece, por todo lo anterior, ser capaz de estimar, por lo general, el conjunto $L(\tau)$ eficazmente sin necesitar tamaños muestrales muy grandes. Parece tener problemas a la hora de estimar conjuntos de nivel $L(\tau)$ cuando el conjunto $L(\tau + \alpha)$ tiene nuevas componentes conexas, con α relativamente pequeño, y a la hora de estimar conjuntos de nivel $L(\tau)$ cuando hay componentes conexas próximas entre sí. En definitiva, los resultados simulados muestran que $d_H\left(L(\tau), \hat{L}(\tau)\right)$ tiende a 0 a medida que n aumenta. Con respecto a los parámetros B , p y s, parece que una pequeña variación en ellos no afecta en gran medida al algoritmo.

El número de remuestras bootstrap B, cuanto mayor mejor, pero hay que tener en cuenta hasta qué punto merece la pena aumentarlo a cambio de un mayor coste computacional. Lo mismo ocurre con el parámetro s, cuanto más pequeño más cercana será la proporción de puntos en el conjunto estimado del valor $1-\tau$, pero a cambio de un mayor coste computacional. El parámetro p afecta principalmente a la estimación de $\mathcal{X}_n^+(\hat{f}_\tau^+)$, pero la cantidad de puntos que forman este conjunto no varía mucho dependiendo de su valor.

n \ τ	0.9	0.7	0.5	0.3
250	0.9938	0.4823	0.6359	0.763
1000	0.4365	0.3396	0.3526	0.433
2500	0.4257	0.2181	0.2567	0.3511

Cuadro 4.8: Media de las distancia de Hausdorff entre $\hat{L}(\tau)$ y $L(\tau)$ para 300 muestras con cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Trimodal II.

τ \ n	250	1000	2500	Nº de componentes conexas reales
0.9	0.6967	0.8867	0.87	2
0.7	0.2467	0.87	0.9833	2
0.5	0.9067	0.97	0.9767	1
0.3	0.1067	0.26	0.35	2

Cuadro 4.9: Proporción de las estimaciones de $\hat{L}(\tau)$ de las 300 muestras generadas para cada valor de $n \in \{250, 1000, 2500\}$, $\tau = \{0.9, 0.7, 0.5, 0.3\}$ con $B = 100$, $s = 0.01$ y $p = 0.25$ en el modelo Trimodal II que están formadas por el mismo número de componentes conexas que $L(\tau)$. En la última columna se indica este último valor.

$\tau = 0.9$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.6518	0.6464	0.6474	0.6456	0.3908	0.6435
	100	0.6513	0.6509	0.6465	0.6509	0.4409	0.4587
	250	0.6409	0.6448	0.6465	0.6448	0.3897	0.4577
500	50	0.6753	0.6728	0.6756	0.676	0.6741	0.6715
	100	0.6703	0.6754	0.6762	0.6738	0.615	0.6723
	250	0.6761	0.6736	0.6753	0.6751	0.676	0.6713
1000	50	0.3085	0.3763	0.2588	0.3565	0.3085	0.3085
	100	0.3085	0.3757	0.3085	0.2589	0.3085	0.3085
	250	0.259	0.3565	0.3085	0.2588	0.3085	0.3085
2500	50	0.2872	0.2539	0.2539	0.2541	0.2883	0.2543
	100	0.2872	0.2872	0.2539	0.2539	0.2883	0.2543
	250	0.2872	0.2874	0.2539	0.2539	0.2883	0.2543
5000	50	0.2444	0.2777	0.236	0.2714	0.2253	0.2589
	100	0.2444	0.2644	0.2511	0.2715	0.2252	0.2589
	250	0.2444	0.274	0.2361	0.2666	0.2255	0.259

Cuadro 4.10: Resultados de las simulaciones para $\tau = 0.9$ del modelo Bimodal I.

$\tau = 0.7$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.4191	0.5641	0.4195	0.419	0.4191	0.4188
	100	0.4192	0.5073	0.4193	0.4193	0.419	0.4186
	250	0.4193	0.5636	0.4194	0.4191	0.4188	0.4189
500	50	0.4572	0.4572	0.514	0.4696	0.4713	0.4054
	100	0.5044	0.5044	0.4696	0.4696	0.4713	0.4054
	250	0.4572	0.4572	0.4696	0.4696	0.4696	0.4054
1000	50	0.241	0.2409	0.2588	0.2408	0.2407	0.2408
	100	0.2726	0.2409	0.3085	0.2408	0.2408	0.2411
	250	0.2412	0.2409	0.3085	0.2406	0.2407	0.2405
2500	50	0.3239	0.3453	0.3284	0.3457	0.3285	0.345
	100	0.3266	0.3327	0.2918	0.3442	0.3297	0.3451
	250	0.3261	0.3325	0.328	0.3452	0.3296	0.346
5000	50	0.2929	0.3285	0.2962	0.2883	0.2929	0.2886
	100	0.2929	0.3291	0.2963	0.2962	0.2879	0.2924
	250	0.2924	0.3293	0.2963	0.2963	0.2877	0.2928

Cuadro 4.11: Resultados de las simulaciones para $\tau = 0.7$ del modelo Bimodal I.

$\tau = 0.5$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.3544	0.4211	0.3543	0.3545	0.3742	0.3743
	100	0.4210	0.4218	0.3545	0.4214	0.3742	0.4211
	250	0.3549	0.4212	0.3543	0.4210	0.3744	0.4213
500	50	0.3010	0.3385	0.2997	0.3009	0.3316	0.3239
	100	0.2997	0.3011	0.3009	0.3380	0.3339	0.3239
	250	0.2998	0.3385	0.2999	0.3338	0.3338	0.3338
1000	50	0.3110	0.3089	0.3037	0.3073	0.3037	0.3037
	100	0.3075	0.3089	0.3037	0.3037	0.3037	0.3037
	250	0.3089	0.3089	0.3037	0.3037	0.3037	0.3037
2500	50	0.1828	0.2017	0.1831	0.1830	0.1837	0.1586
	100	0.1763	0.1977	0.1846	0.1830	0.1585	0.1661
	250	0.1763	0.1846	0.1831	0.1830	0.1566	0.1836
5000	50	0.1863	0.2112	0.1746	0.1746	0.1859	0.1745
	100	0.1863	0.2112	0.1761	0.1746	0.1736	0.1745
	250	0.1650	0.2222	0.1746	0.1746	0.1846	0.1846

Cuadro 4.12: Resultados de las simulaciones para $\tau = 0.5$ del modelo Bimodal I.

$\tau = 0.3$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.4426	0.4426	0.3929	0.3933	0.3924	0.4422
	100	0.3931	0.4426	0.3924	0.3929	0.3924	0.4422
	250	0.3925	0.4426	0.3924	0.3925	0.4422	0.4422
500	50	0.2877	0.2862	0.2877	0.2877	0.2971	0.2877
	100	0.2862	0.2862	0.2878	0.2877	0.2971	0.2877
	250	0.2878	0.2862	0.2877	0.2863	0.2971	0.2878
1000	50	0.1960	0.1964	0.1832	0.1838	0.1889	0.1607
	100	0.1961	0.1747	0.1885	0.1838	0.1889	0.1587
	250	0.1842	0.1851	0.1886	0.1840	0.1869	0.1603
2500	50	0.1865	0.2093	0.1710	0.2212	0.1710	0.2211
	100	0.1865	0.2133	0.1710	0.2212	0.1866	0.2212
	250	0.1711	0.2093	0.1710	0.2212	0.1865	0.2212
5000	50	0.1374	0.1515	0.1374	0.1711	0.1252	0.1711
	100	0.1329	0.1527	0.1398	0.1711	0.1410	0.1709
	250	0.1365	0.1528	0.1381	0.1710	0.1253	0.1711

Cuadro 4.13: Resultados de las simulaciones para $\tau = 0.3$ del modelo Bimodal I.

$\tau = 0.9$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	2.6537	2.6533	2.6534	2.6532	2.6534	2.6539
	100	2.6534	2.6531	2.6538	2.6534	2.6536	2.6537
	250	2.6536	2.6537	2.6534	2.6538	2.6537	2.6538
500	50	0.3246	0.3247	0.3247	0.3245	0.3247	0.3247
	100	0.3246	0.3246	0.3246	0.3246	0.3246	0.3245
	250	0.3245	0.3245	0.3245	0.3247	0.3245	0.3246
1000	50	0.1568	0.1273	0.1573	0.1573	0.1495	0.1496
	100	0.1467	0.1273	0.1495	0.1573	0.1495	0.1496
	250	0.1573	0.1273	0.1573	0.1573	0.1496	0.1496
2500	50	0.2504	0.2503	0.2503	0.2320	0.2503	0.2331
	100	0.2340	0.2509	0.2338	0.2337	0.2503	0.2327
	250	0.2506	0.2504	0.2503	0.2337	0.2503	0.2320
5000	50	0.0870	0.1127	0.0870	0.1127	0.0870	0.1128
	100	0.0870	0.1028	0.0870	0.1128	0.0899	0.1128
	250	0.0870	0.1028	0.0870	0.1126	0.0899	0.1128

Cuadro 4.14: Resultados de las simulaciones para $\tau = 0.9$ del modelo Bimodal II.

$\tau = 0.7$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.2196	0.2196	0.2196	0.2196	0.2196	0.2196
	100	0.2196	0.2196	0.2196	0.2197	0.2197	0.2196
	250	0.2196	0.2196	0.2196	0.2196	0.2196	0.2479
500	50	0.4008	0.4008	0.4008	0.4013	0.4012	0.4012
	100	0.4008	0.4008	0.4008	0.4012	0.4012	0.4013
	250	0.4008	0.4008	0.4008	0.4012	0.4012	0.4013
1000	50	0.2928	0.2928	0.2941	0.2928	0.2938	0.2938
	100	0.2928	0.2928	0.2941	0.2928	0.2944	0.3196
	250	0.2928	0.2928	0.2943	0.2832	0.2938	0.2938
2500	50	0.2368	0.2168	0.2368	0.2169	0.2368	0.2368
	100	0.2368	0.2168	0.2306	0.2169	0.2368	0.2368
	250	0.2368	0.2168	0.2306	0.2169	0.2368	0.2368
5000	50	0.1687	0.2077	0.1805	0.2077	0.2004	0.2004
	100	0.1688	0.2077	0.1829	0.1829	0.2004	0.2004
	250	0.1688	0.2076	0.1829	0.1805	0.2004	0.2004

Cuadro 4.15: Resultados de las simulaciones para $\tau = 0.7$ del modelo Bimodal II.

$\tau = 0.5$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.3108	0.3104	0.3108	0.3112	0.3103	0.3184
	100	0.2342	0.3111	0.3108	0.2350	0.2355	0.3182
	250	0.3109	0.3103	0.2321	0.3106	0.2323	0.3181
500	50	0.3487	0.3009	0.3492	0.2782	0.3491	0.2580
	100	0.3532	0.3009	0.3491	0.2782	0.3013	0.2571
	250	0.3532	0.2761	0.3013	0.2782	0.3013	0.2570
1000	50	0.2018	0.2018	0.2000	0.1998	0.2021	0.1795
	100	0.2017	0.2018	0.2012	0.2013	0.2021	0.1796
	250	0.2017	0.2019	0.2012	0.1998	0.2022	0.1796
2500	50	0.1545	0.1684	0.1545	0.1545	0.1548	0.1548
	100	0.1542	0.1729	0.1545	0.1545	0.1385	0.1385
	250	0.1684	0.1543	0.1545	0.1545	0.1548	0.1385
5000	50	0.0975	0.1113	0.0976	0.0753	0.0975	0.0813
	100	0.0975	0.1113	0.0976	0.1189	0.0815	0.0814
	250	0.0814	0.1113	0.0991	0.1191	0.0815	0.0815

Cuadro 4.16: Resultados de las simulaciones para $\tau = 0.5$ del modelo Bimodal II.

$\tau = 0.3$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.6535	0.5749	0.6477	0.6329	0.5731	0.6539
	100	0.6415	0.5736	0.6404	0.6354	0.6477	0.6356
	250	0.6469	0.5706	0.6497	0.6354	0.6563	0.6270
500	50	0.6230	0.6229	0.6219	0.6279	0.6215	0.6192
	100	0.6226	0.6253	0.6229	0.6288	0.6245	0.6198
	250	0.6140	0.6259	0.6253	0.6222	0.6244	0.6249
1000	50	0.2442	0.2450	0.2442	0.2430	0.2442	0.2464
	100	0.2451	0.2447	0.2442	0.2430	0.2442	0.2471
	250	0.2430	0.2450	0.2450	0.2430	0.2447	0.2474
2500	50	0.1421	0.1567	0.1563	0.1727	0.1726	0.1728
	100	0.1565	0.1566	0.1561	0.1728	0.1424	0.1728
	250	0.1562	0.1567	0.1561	0.1727	0.1725	0.1728
5000	50	0.1362	0.2117	0.1378	0.2113	0.1378	0.1362
	100	0.1362	0.2116	0.1378	0.2119	0.1362	0.1362
	250	0.1362	0.2115	0.1378	0.2119	0.1362	0.1361

Cuadro 4.17: Resultados de las simulaciones para $\tau = 0.3$ del modelo Bimodal II.

$\tau = 0.9$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	1.3826	1.3823	1.3828	1.3823	1.3821	1.3824
	100	1.1548	1.3819	1.1536	1.3825	1.3825	1.4694
	250	1.4699	1.3815	1.3814	1.3824	1.3823	1.4702
500	50	1.7637	1.7638	1.7636	1.7637	1.7634	1.7814
	100	1.7638	1.7635	1.7635	1.7637	1.7637	1.7813
	250	1.7636	1.7633	1.7631	1.7636	1.7637	1.7814
1000	50	1.7536	1.8056	1.7535	1.7537	1.7536	1.7537
	100	1.7537	1.8053	1.7536	1.7537	1.7537	1.7537
	250	1.7537	1.8057	1.7537	1.7536	1.7536	1.7537
2500	50	1.5708	1.6006	1.5709	1.6006	1.5541	1.5541
	100	1.5708	1.6006	1.5739	1.6006	1.5541	1.5541
	250	1.5708	1.6006	1.5739	1.6006	1.5541	1.5541
5000	50	1.6468	1.6678	1.6508	1.6679	1.6468	1.6697
	100	1.6509	1.6641	1.6508	1.6679	1.6468	1.6678
	250	1.6507	1.6641	1.6508	1.6678	1.6507	1.6697

Cuadro 4.18: Resultados de las simulaciones para $\tau = 0.9$ del modelo Trimodal I.

$\tau = 0.7$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.3931	0.3918	0.3951	0.3902	0.3931	0.5321
	100	0.3936	0.3916	0.3969	0.3949	0.3879	0.3855
	250	0.3834	0.3928	0.3882	0.3911	0.3908	0.3935
500	50	0.2997	0.6407	0.2993	0.2992	0.2992	0.3325
	100	0.2987	0.6366	0.6360	0.2989	0.2992	0.3325
	250	0.2996	0.6402	0.2997	0.2996	0.2992	0.3330
1000	50	0.3613	0.3629	0.3623	0.3625	0.3621	0.3622
	100	0.3613	0.3613	0.3614	0.3614	0.3216	0.3623
	250	0.3625	0.3632	0.3604	0.3607	0.2651	0.3622
2500	50	0.2677	0.2677	0.2677	0.2583	0.2627	0.2583
	100	0.2677	0.3551	0.2677	0.2627	0.2627	0.2583
	250	0.2677	0.2677	0.2720	0.2674	0.2720	0.2583
5000	50	0.1325	0.1471	0.1114	0.1651	0.1205	0.1651
	100	0.1324	0.1464	0.1204	0.1650	0.1113	0.1649
	250	0.1116	0.1464	0.1204	0.1648	0.1115	0.1651

Cuadro 4.19: Resultados de las simulaciones para $\tau = 0.7$ del modelo Trimodal I.

$\tau = 0.5$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.5024	0.5008	0.5165	0.5145	0.5020	0.5003
	100	0.4993	0.5093	0.5008	0.5148	0.5031	0.5030
	250	0.4990	0.5059	0.5039	0.5153	0.5006	0.5042
500	50	0.2767	0.2767	0.2773	0.2769	0.2786	0.2773
	100	0.2767	0.2767	0.2773	0.2773	0.2785	0.2785
	250	0.2767	0.2767	0.2773	0.2769	0.2785	0.2773
1000	50	0.2513	0.2851	0.2599	0.2539	0.2513	0.2513
	100	0.2513	0.2845	0.2599	0.2599	0.2561	0.2513
	250	0.2513	0.2853	0.2599	0.2523	0.2467	0.2513
2500	50	0.2475	0.2629	0.2410	0.2629	0.2460	0.2625
	100	0.2434	0.2609	0.2459	0.2622	0.2482	0.2607
	250	0.2456	0.2602	0.2488	0.2617	0.2483	0.2623
5000	50	0.1559	0.1561	0.1340	0.1673	0.1344	0.1709
	100	0.1559	0.1552	0.1340	0.1674	0.1344	0.1711
	250	0.1517	0.1558	0.1340	0.1673	0.1344	0.1711

Cuadro 4.20: Resultados de las simulaciones para $\tau = 0.5$ del modelo Trimodal I.

$\tau = 0.3$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.5767	0.5538	0.5538	0.5830	0.5171	0.5170
	100	0.5170	0.5538	0.5597	0.5830	0.5170	0.5170
	250	0.5539	0.5170	0.5538	0.5830	0.5170	0.5170
500	50	0.3057	0.3081	0.3067	0.3048	0.3058	0.3050
	100	0.3104	0.3708	0.3069	0.3029	0.3062	0.3066
	250	0.3054	0.3080	0.3037	0.3180	0.3050	0.3047
1000	50	0.3344	0.3360	0.3224	0.4168	0.2995	0.2992
	100	0.3005	0.3381	0.3015	0.4168	0.2980	0.2985
	250	0.2986	0.3231	0.2996	0.4172	0.2993	0.2983
2500	50	0.2576	0.2252	0.2409	0.2409	0.2251	0.2576
	100	0.2576	0.2251	0.2409	0.2409	0.2409	0.2576
	250	0.2409	0.2409	0.2409	0.2409	0.2251	0.2576
5000	50	0.1240	0.2252	0.1244	0.2270	0.1032	0.1687
	100	0.1233	0.1605	0.1401	0.1970	0.1031	0.1691
	250	0.1348	0.1608	0.1398	0.2328	0.1031	0.1686

Cuadro 4.21: Resultados de las simulaciones para $\tau = 0.3$ del modelo Trimodal I.

$\tau = 0.9$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.3915	0.3625	0.3607	0.3910	0.3632	0.3629
	100	0.3633	0.3631	0.3624	0.3891	0.3630	0.3620
	250	0.3630	0.3631	0.3632	0.3632	0.3625	0.3633
500	50	1.5700	1.5221	1.5702	1.5219	1.5706	1.5704
	100	1.5217	1.5705	1.5710	1.5224	1.5702	1.5700
	250	1.5701	1.5221	1.5699	1.5704	1.5703	1.5700
1000	50	0.4199	0.4242	0.4202	0.4497	0.3716	0.4198
	100	0.4198	0.4242	1.8700	0.4495	0.3712	1.8700
	250	0.3920	0.4242	1.8700	0.4495	1.8699	0.4496
2500	50	0.1865	0.1864	0.1978	0.1846	0.1846	0.1870
	100	0.1845	0.1948	0.1977	0.1846	0.1846	0.1870
	250	0.1864	0.1864	0.1977	0.1846	0.1845	0.1870
5000	50	0.2045	0.2459	0.1690	0.2241	0.1784	0.1971
	100	0.1705	0.2331	0.1690	0.2239	0.1786	0.2042
	250	0.2059	0.2337	0.1690	0.2238	0.1782	0.2038

Cuadro 4.22: Resultados de las simulaciones para $\tau = 0.9$ del modelo Trimodal II.

$\tau = 0.7$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.4826	0.4822	0.4726	0.4723	0.4879	0.4728
	100	0.4877	0.5000	0.4705	0.4887	0.4698	0.4866
	250	0.4859	0.4995	0.4718	0.4706	0.4765	0.4839
500	50	0.2335	0.3204	0.2195	0.3078	0.2195	0.3116
	100	0.2240	0.3207	0.2334	0.2307	0.2241	0.3588
	250	0.2195	0.3138	0.2195	0.3075	0.2195	0.3585
1000	50	0.3560	0.4649	0.3561	0.3692	0.3561	0.3692
	100	0.3562	0.4805	0.3559	0.3693	0.3693	0.4203
	250	0.3685	0.4802	0.3560	0.3692	0.3688	0.4203
2500	50	0.1419	0.1635	0.1423	0.1637	0.1428	0.1742
	100	0.1419	0.1637	0.1423	0.1637	0.1428	0.1742
	250	0.1419	0.1636	0.1423	0.1637	0.1428	0.1742
5000	50	0.1765	0.2588	0.1578	0.1607	0.1764	0.1762
	100	0.1835	0.2583	0.1607	0.1578	0.1765	0.1765
	250	0.1723	0.2652	0.1607	0.1543	0.1765	0.1764

Cuadro 4.23: Resultados de las simulaciones para $\tau = 0.7$ del modelo Trimodal II.

$\tau = 0.5$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	0.5323	0.5383	0.5335	0.5693	0.4730	0.5592
	100	0.5194	0.5381	0.5337	0.5791	0.4727	0.5737
	250	0.5194	0.5336	0.5335	0.5783	0.4728	0.5595
500	50	0.3635	0.3560	0.2901	0.3561	0.2905	0.3282
	100	0.3152	0.3746	0.3746	0.3562	0.2898	0.3290
	250	0.3151	0.3747	0.2906	0.3559	0.2905	0.3291
1000	50	0.4813	0.4835	0.5273	0.5533	0.5270	0.5817
	100	0.4837	0.5273	0.5529	0.5532	0.5272	0.5820
	250	0.5270	0.5267	0.5535	0.5533	0.5273	0.5811
2500	50	0.3123	0.3554	0.3123	0.3374	0.2836	0.3430
	100	0.3537	0.3509	0.3124	0.3192	0.2842	0.3454
	250	0.3123	0.3510	0.3124	0.3374	0.2842	0.3453
5000	50	0.1612	0.2282	0.1674	0.2037	0.1590	0.2037
	100	0.1683	0.2604	0.1674	0.2037	0.1590	0.2037
	250	0.2050	0.2090	0.1674	0.2037	0.1590	0.2037

Cuadro 4.24: Resultados de las simulaciones para $\tau = 0.5$ del modelo Trimodal II.

$\tau = 0.3$		p = 0.1		p = 0.25		p = 0.5	
n	B	s = 0.01	s = 0.04	s = 0.01	s = 0.04	s = 0.01	s = 0.04
250	50	1.2440	1.2440	1.2441	1.2440	1.4533	1.2587
	100	1.2440	1.2589	1.2440	1.2575	1.4530	1.2441
	250	1.2440	1.2441	1.2575	1.4530	1.2575	1.2441
500	50	0.3192	0.3124	0.3158	0.3332	0.3292	0.3320
	100	0.3185	0.3743	0.3156	0.3335	0.3341	0.3341
	250	0.3187	0.3167	0.3152	0.3331	0.3288	0.3354
1000	50	0.3633	0.3655	0.3645	0.3650	0.4034	0.3948
	100	0.4036	0.3643	0.3649	0.3647	0.3646	0.3949
	250	0.4034	0.3759	0.3651	0.3947	0.3641	0.3947
2500	50	0.3544	0.3512	0.3485	0.3484	0.3321	0.3872
	100	0.3465	0.3491	0.3501	0.3472	0.3313	0.3497
	250	0.3390	0.3512	0.3466	0.3483	0.3327	0.3499
5000	50	0.3071	0.3883	0.3068	0.3975	0.3282	0.3975
	100	0.3065	0.3883	0.3064	0.3970	0.2858	0.3974
	250	0.3072	0.3884	0.3062	0.3966	0.2858	0.3975

Cuadro 4.25: Resultados de las simulaciones para $\tau = 0.3$ del modelo Trimodal II.

Capítulo 5

Análisis de los datos

En el Capítulo 1 presentamos los datos donde se recogen las localizaciones de nidos de velutinas registrados en Galicia entre los años 2014 y 2019, ambos inclusive. En este Capítulo estimaremos distintos conjuntos de nivel empleando el método visto en el Capítulo 3 para la avispa asiática, tras haber comprobado en el Capítulo 4 su correcto funcionamiento general, pudiendo de esta forma estudiar la distribución de los focos en Galicia de forma anual en la Sección 5.1, y en la zona de Santiago de Compostela en la Sección 5.2.

5.1. Estimación anual

Empezaremos por estimar el conjunto de nivel estimado $\hat{L}(\tau)$ para los distintos valores de $\tau \in \{0.95, 0.9, 0.8, 0.7, 0.5\}$ para cada uno de los seis años desde el 2014 hasta el 2019. Para ello utilizaremos unos valores en el algoritmo de $J = 60$, $r_m = 0$ y $r_M = 502275$ (escogido como indicamos al final del Capítulo 3) para el cálculo del parámetro $\hat{r}_0$; y de $B = 100$, $p = 0.25$ y $s = 0.01$ para el cálculo de $\hat{L}(\tau)$. En las Figuras 5.4, 5.5, 5.6, 5.7, 5.8 y 5.9, al final de este Capítulo, se pueden observar gráficamente los resultados obtenidos para los años $2014, 2015, 2016, 2017, 2018$ y 2019, respectivamente, mientras que los valores estimados $\hat{r}_0(\hat{f}_\tau)$ se pueden contemplar en el Cuadro 5.1, donde marcamos como infinito aquellos valores iguales al valor r_M. Como habíamos comentado en el Capítulo 1, la avispa se localizó por primera vez en la zona norte de la provincia de Lugo y al sur de Pontevedra en el 2012, aún así, habiéndose expandido en ambas zonas desde el principio, se puede observar cómo hasta, y durante, el año 2014, la expansión fue mayoritaria en la zona norte. En el año 2015 se observa un repunte de casos en la zona del sur de la ciudad de Vigo, quedando esta contenida en el conjunto estimado $\hat{L}(0.95)$. También se detecta la presencia de una moda en la ciudad de A Coruña y un aumento en la zona norte de Lugo. En el año 2016 el número de nidos reportados se duplicó con respecto a su año pasado. El conjunto $\hat{L}(0.95)$ presenta una única componente conexa en la zona de A Coruña lo cual indica un mayor incremento en la detección (y posible localización) de nidos de velutina en esa zona comparado con el resto de Galicia. $\hat{L}(0.5)$ nos muestra que en el 2016 la presencia de los nidos empieza a expandirse no solamente por la costa de la comunidad, particularmente detecta una componente en la ciudad de Santiago de Compostela y la de la zona de A Coruña se extiende hasta el sur de la comarca de Betanzos. Con respecto al año 2017, los casos se volvieron a duplicar. Este año $\hat{L}(0.95)$ detecta dos componentes. Una en A Coruña y una en Santiago de Compostela. De esto no se puede inferir que las avispas asiáticas tengan prevalencia por las zonas urbanas debido a que en estas, de una misma cantidad de nidos se reportan más al ser zonas más pobladas. En el año 2018, habiéndose

τ \ Año	2014	2015	2016	2017	2018	2019
0.95	∞	∞	9210	25827	4620	36120
0.9	14619	40678	6456	3072	1711	6153
0.8	∞	2137	3598	1969	3512	2766
0.7	∞	3562	1929	2054	1671	3917
0.5	4401	2275	2041	1757	1746	2505

Cuadro 5.1: Valor de los parámetros estimados $\hat{r}_0(\hat{f}_\tau)$ para los valores de $\tau \in \{0.95, 0.9, 0.8, 0.7, 0.5\}$ para las estimaciones realizadas sobre los datos de la velutina de los años desde el 2014 hasta el 2019, ambos inclusive.

reportado una cantidad similar de casos que en el 2017, la gran cantidad de componentes conexas que se detecta para todos los niveles estudiados del valor τ nos indican que la cantidad de focos es cada vez mayor, pero a la vez que la distribución de los nidos se vuelve un poco más homogénea a lo largo del territorio gallego. Se puede ver como progresivamente se ha ido expandiendo hacia la zona oriental de la comunidad, quedando la ciudad de Ourense contenida en el conjunto $\hat{L}(0.9)$ y la ciudad de Lugo en el conjunto $\hat{L}(0.8)$. Por último, en el 2019 se registró una caída a la mitad de detecciones de nidos. $\hat{L}(0.95)$ detecta tres componentes: una en la ciudad de Ourense, una en la ciudad de Pontevedra y la mayor de todas en la ciudad de La Coruña y el sur de esta. La ciudad de Santiago de Compostela ve mayor disminución en los casos quedando su área no contenida en $\hat{L}(0.9)$. Los focos pasan a estar principalmente en las ciudades y por las zonas rurales se distribuye de forma más homogénea. En la Figura 5.1 presentamos el conjunto $\hat{L}(0.25)$ por año. Esto es la estimación de las áreas de Galicia más pequeñas que contienen al 75 % de los nidos de velutina reales de la comunidad en cada año. Se puede ver de esta forma muy bien cómo los dos focos en los que se inició la invasión de la velutina en Galicia van perdiendo importancia en comparación a otras zonas donde el aumento es mucho más significativo.

De forma general podemos comprobar cómo aún expandiéndose por toda la comunidad, a diferencia de la avispa común, la velutina construye sus nidos de forma habitual en las ciudades. Con los conjuntos estimados podemos establecer prioridades en distintas zonas para la búsqueda de nidos, así que lo mismo puede resultar interesante a la hora de estudiar la distribución en las ciudades.

5.2. Estimación en Santiago de Compostela

Una vez comprobada la distribución de los focos de nidos de velutina a lo largo de los años en Galicia y, visto que estos se dan en ciudades, por último en este trabajo, vamos a comprobar en qué parte de la ciudad de Santiago de Compostela se focalizan los casos. Para ello restringiremos los datos de la velutina a los que se localizan en la comarca de Santiago de Compostela, tomando los 6 años de la muestra a la vez. Utilizaremos los valores de $B = 250$, $p = 0.25$, $s = 0.01$ y para los valores del algoritmo del cálculo de $\hat{r}_0$ utilizaremos $J = 60$, $r_m = 0$ y $r_M = 38545$. A la izquierda de la Figura 5.2 mostramos el conjunto estimado $\hat{L}(0.9)$. El valor estimado $\hat{r}_0(\hat{f}_{0.9}) = 38545 = r_M$, lo cual nos indica que el conjunto $\hat{L}(0.9)$ es, en esencia, convexo. Se ve cómo la Catedral de Santiago se encuentra en el interior del conjunto, por lo que parece que no es solo que haya gran concentración de nidos en las ciudades sino que se encuentran en le centro de estas. Hay que recordar que esto puede ser porque en

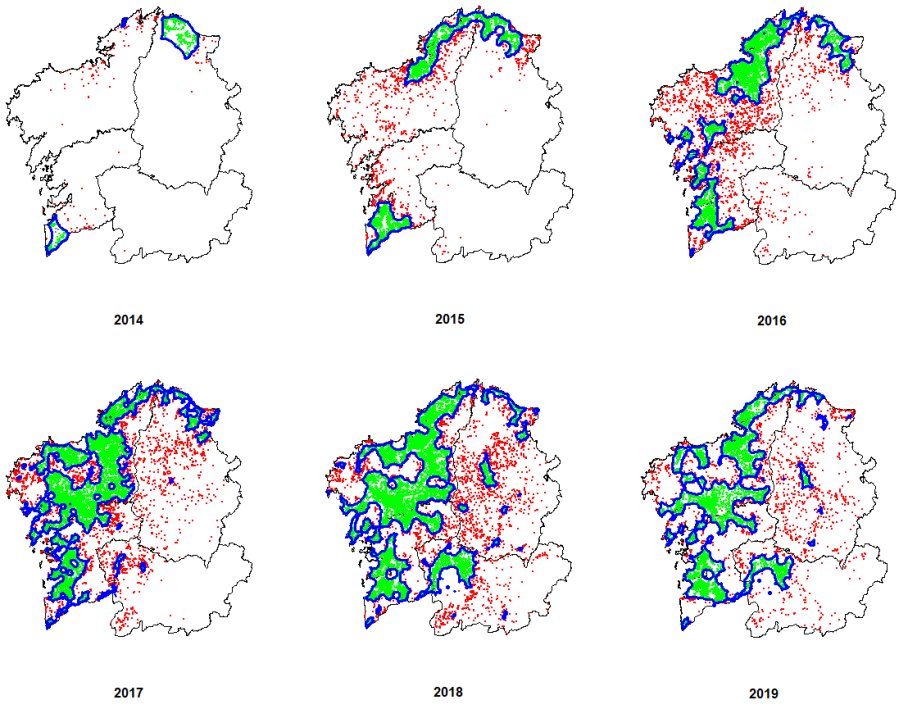


Figura 5.1: El conjunto $\mathcal{X}_n^+(\hat{f}_{0.25})$ en verde , $\mathcal{X}_n^-(\hat{f}_{0.25})$ en rojo y $\mathcal{X}_n \setminus \left(\mathcal{X}_n^+(\hat{f}_{0.25}) \cup \mathcal{X}_n^-(\hat{f}_{0.25})\right)$ con puntos negros, obtenidos sobre los datos de la velutina para cada año entre el 2014 y el 2019. En azul aparece representado el conjunto de nivel $\hat{L}(0.25)$ obtenido mediante el algoritmo 3.11.

las zonas con más densidad de población es más probable que se encuentre un nido en el caso de que hubiera la misma densidad de nidos. En el lado derecho de la Figura 5.2 mostramos la reconstrucción del conjunto $\hat{L}(0.7)$. El parámetro $\hat{r}_0(\hat{f}_{0.7}) = 927.0046$, quedando así un conjunto más ajustado que si supusiéramos convexidad. Vemos en ella cómo los datos se concentran en la ciudad de Santiago de Compostela, viendo que la Facultad de Matemáticas de la USC está en el interior del conjunto de nivel, zona que ya no es considerada centro de la ciudad, con una menor densidad de población.

Con lo anterior, viendo que efectivamente está concentrada la dispersión de los nidos en el centro de la ciudad, vamos a comprobar si a un nivel $\tau = 0.95$ sigue estando la catedral en el interior del conjunto $L(\tau)$. De esta forma obtenemos un valor de $\hat{r}_0(\hat{f}_{0.95}) = 2670.071$, siendo este conjunto, a diferencia de $\hat{L}(0.9)$, no convexo. A la izquierda de la Figura 5.3 observamos que efectivamente ocurre. Como ya habíamos mencionado es posiblemente debido a que en las zonas de alta densidad poblacional haya más gente que pueda percatarse de la existencia de un nido, y no por un tipo de predilección de la velutina por las áreas altamente pobladas.
En mayo del año 2020 ha fallecido un hombre de 54 años en la parroquia de Villestro, en el municipio de Santiago de Compostela, debido a una reacción alérgica a la picadura de una velutina. Podemos comprobar si esta zona estaba dentro del conjunto de nivel de la distribución de la velutina del 50 %, o si era una zona más aislada, en los 6 años anteriores. Al realizar la estimación obtenemos un valor $\hat{r}_0(\hat{f}_{0.5}) = 477.1192$. A la derecha de la Figura 5.3 comprobamos que no es así, que la zona de Villestro no se encuentra en ningún foco de la comarca de Santiago de Compostela, lo cual en cierto modo es esperable ya que la mayoría de las notificaciones ocurren en áreas urbanas, mientras que gran parte de las defunciones son provocadas en zonas rurales debido a la mayor dificultad de la detección de los nidos, así como en las zonas dedicadas a la apicultura.

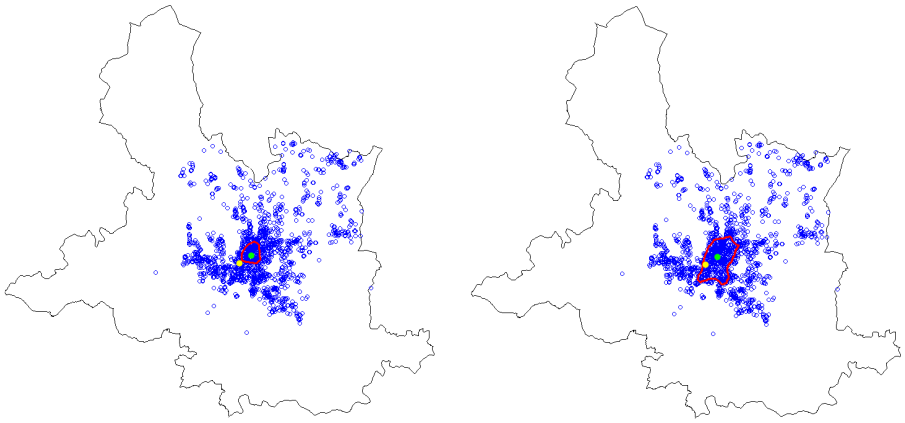


Figura 5.2: En azul la muestra de la velutina en la comarca de Santiago de Compostela. En rojo la frontera del conjunto de nivel estimado $\hat{L}(0.9)$ en la izquierda y del conjunto de nivel estimado $\hat{L}(0.7)$ en la derecha. En verde la localización de la Catedral de Santiago de Compostela y en amarillo la localización de la Facultad de Matemáticas de la USC.

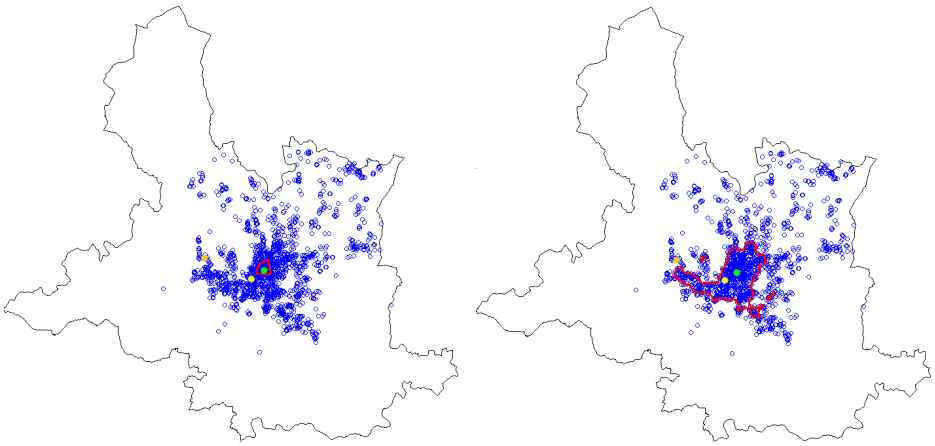


Figura 5.3: En azul la muestra de la velutina en la comarca de Santiago de Compostela. En rojo la frontera del conjunto de nivel estimado $\hat{L}(0.95)$ en la izquierda y del conjunto de nivel estimado $\hat{L}(0.5)$ en la derecha. En verde la localización de la Catedral de Santiago de Compostela, en amarillo la localización de la Facultad de Matemáticas de la USC y en naranja la localización de la parroquia de Villestro.

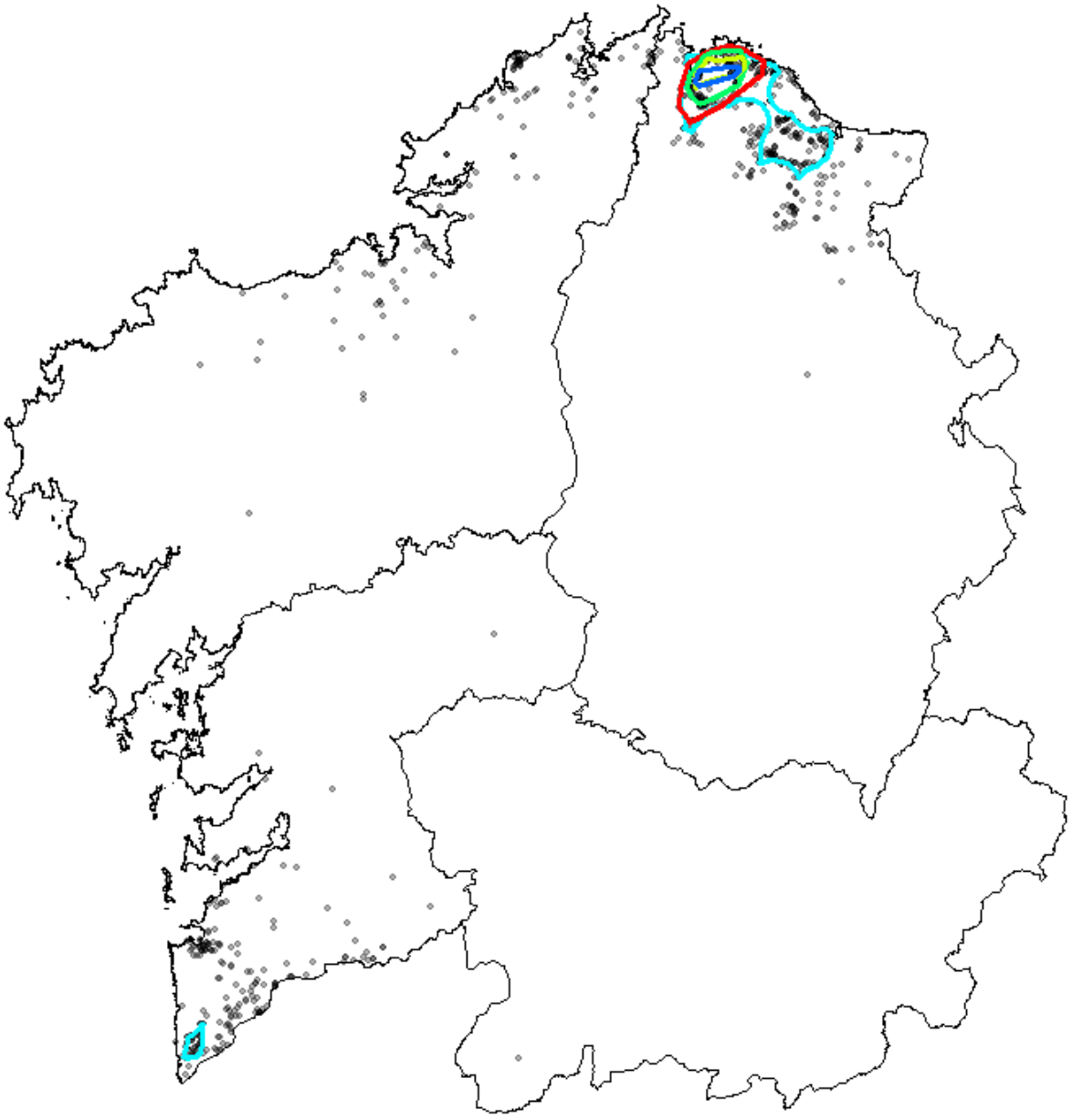


Figura 5.4: La muestra de nidos de velutina del año 2014 en Galicia en negro. Los conjuntos de nivel estimados para $\tau = 0.95, 0.9, 0.8, 0.7$ y 0.5 en color azul (oscuro), amarillo, verde, rojo y turquesa, respectivamente.

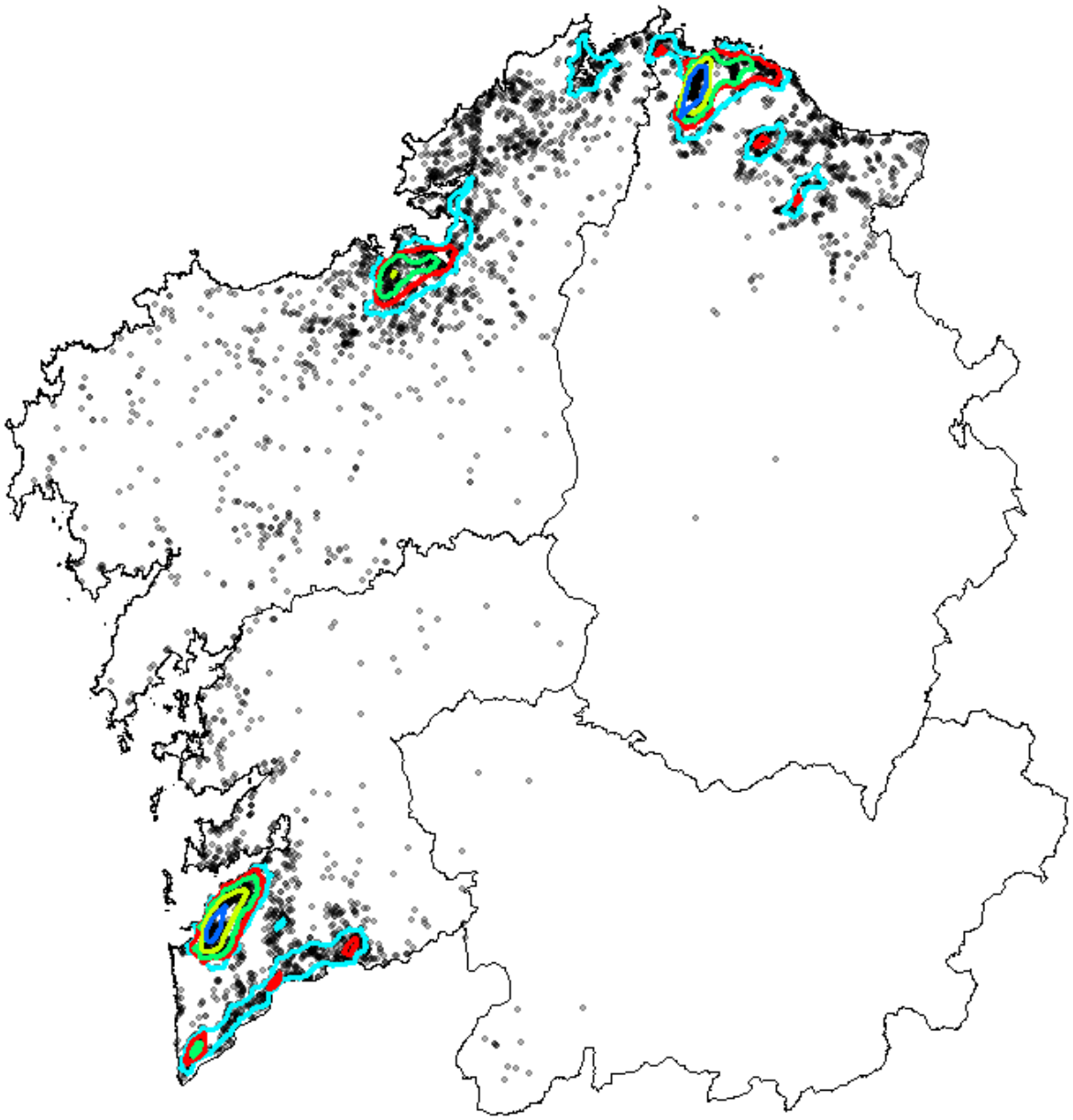


Figura 5.5: La muestra de nidos de velutina del año 2015 en Galicia en negro. Los conjuntos de nivel estimados para $\tau = 0.95, 0.9, 0.8, 0.7$ y 0.5 en color azul (oscuro), amarillo, verde, rojo y turquesa, respectivamente.

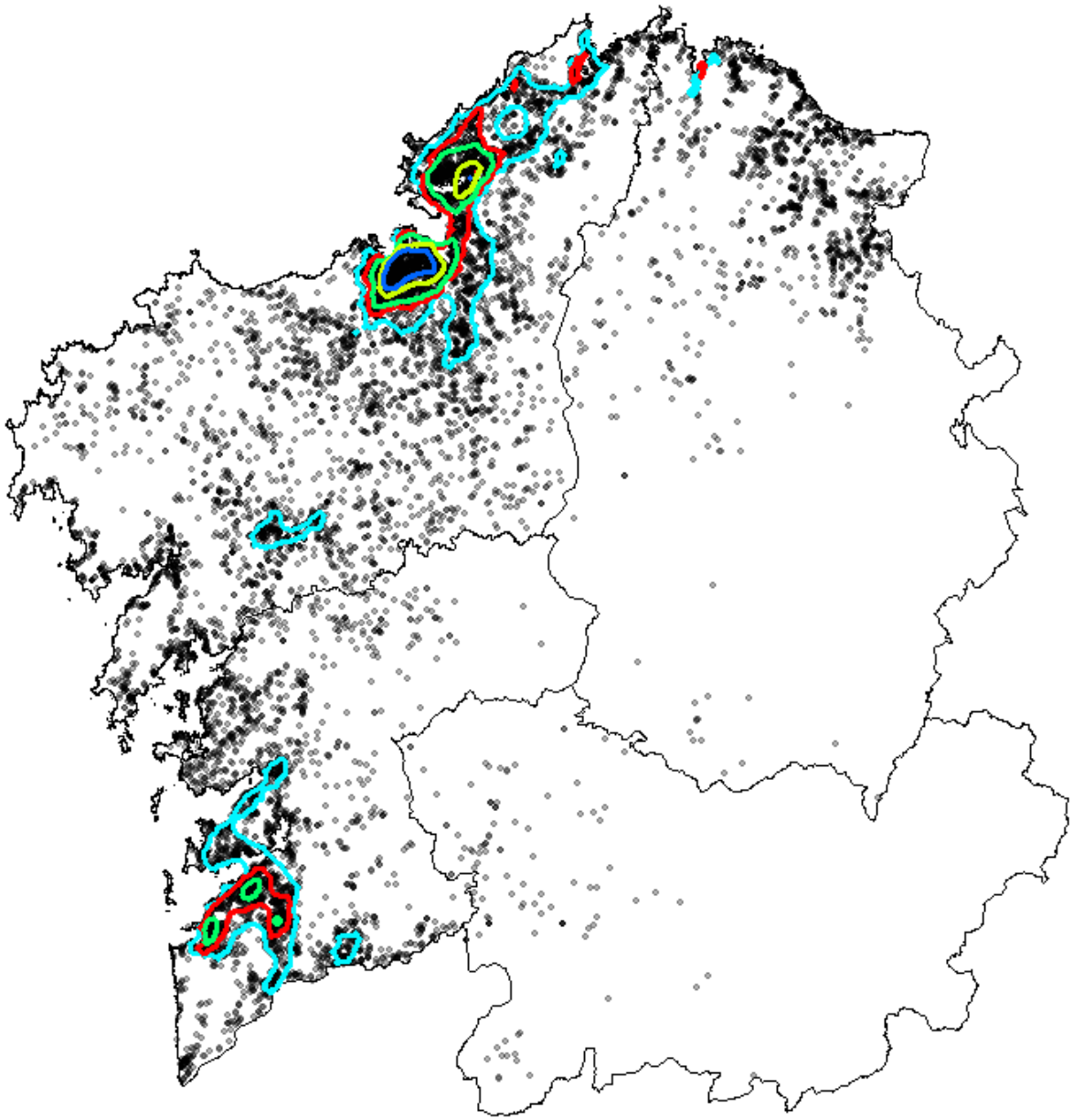


Figura 5.6: La muestra de nidos de velutina del año 2016 en Galicia en negro. Los conjuntos de nivel estimados para $\tau = 0.95, 0.9, 0.8, 0.7$ y 0.5 en color azul (oscuro), amarillo, verde, rojo y turquesa, respectivamente.

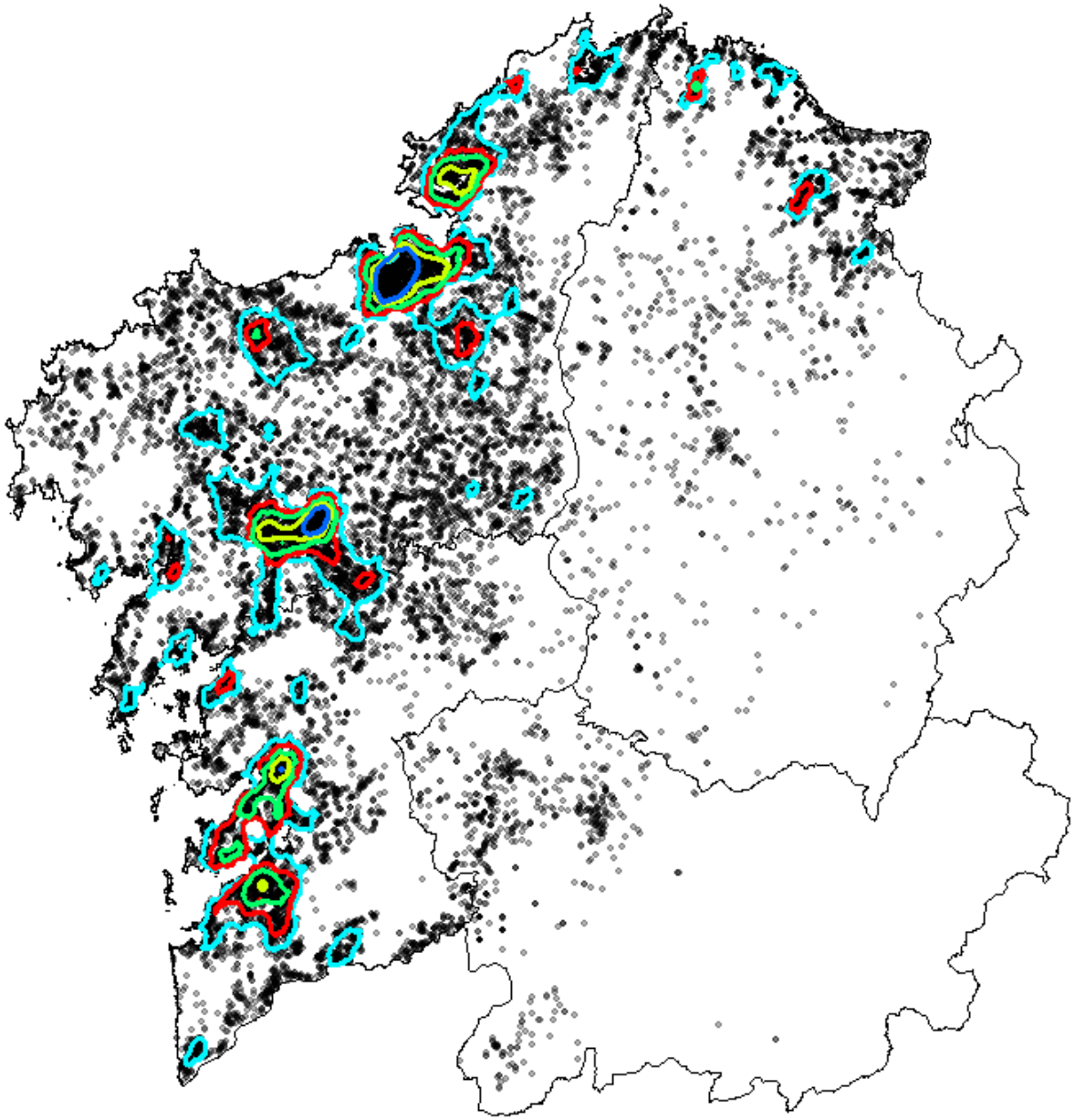


Figura 5.7: La muestra de nidos de velutina del año 2017 en Galicia en negro. Los conjuntos de nivel estimados para $\tau = 0.95, 0.9, 0.8, 0.7$ y 0.5 en color azul (oscuro), amarillo, verde, rojo y turquesa, respectivamente.

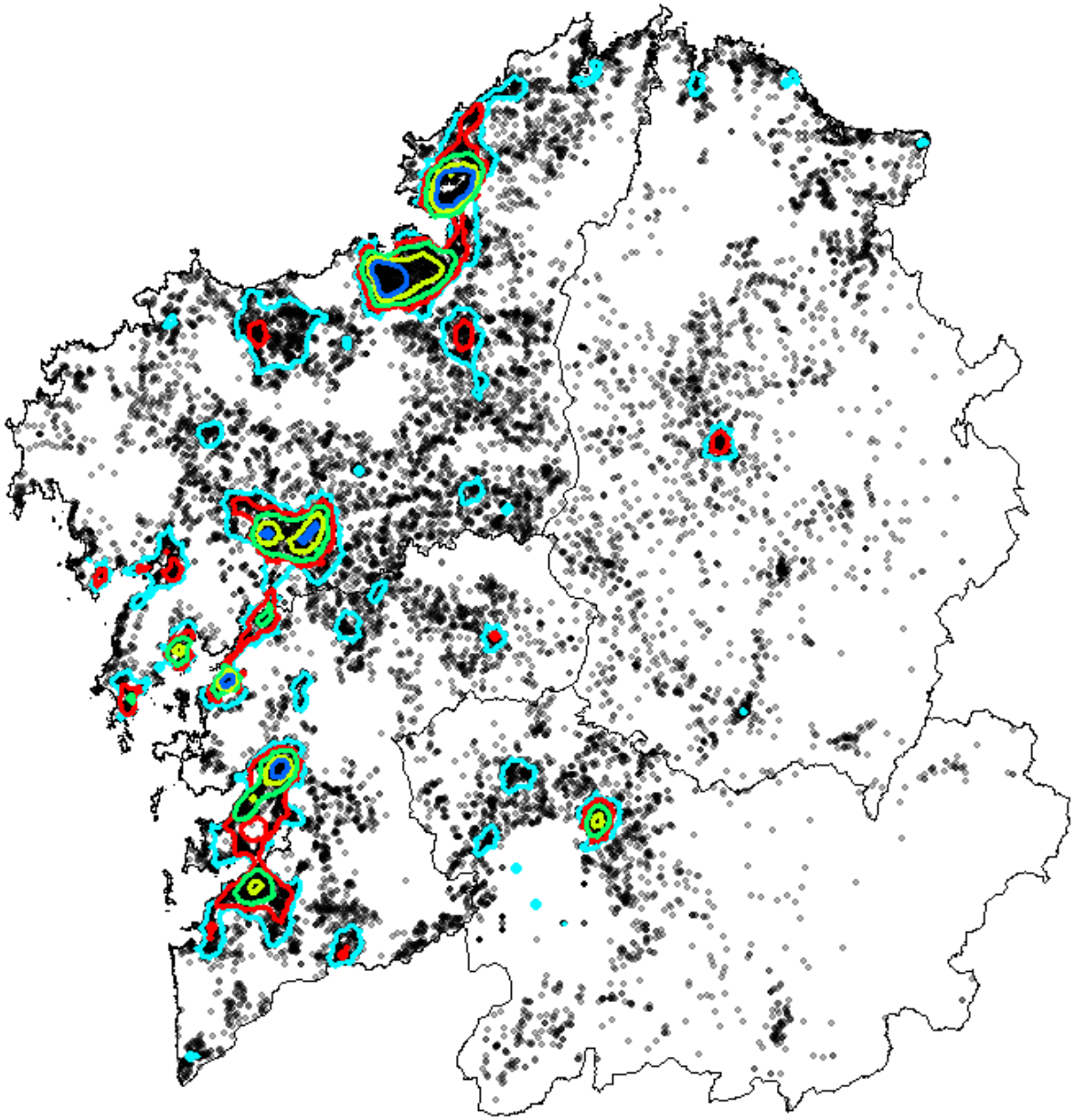


Figura 5.8: La muestra de nidos de velutina del año 2018 en Galicia en negro. Los conjuntos de nivel estimados para $\tau = 0.95, 0.9, 0.8, 0.7$ y 0.5 en color azul (oscuro), amarillo, verde, rojo y turquesa, respectivamente.

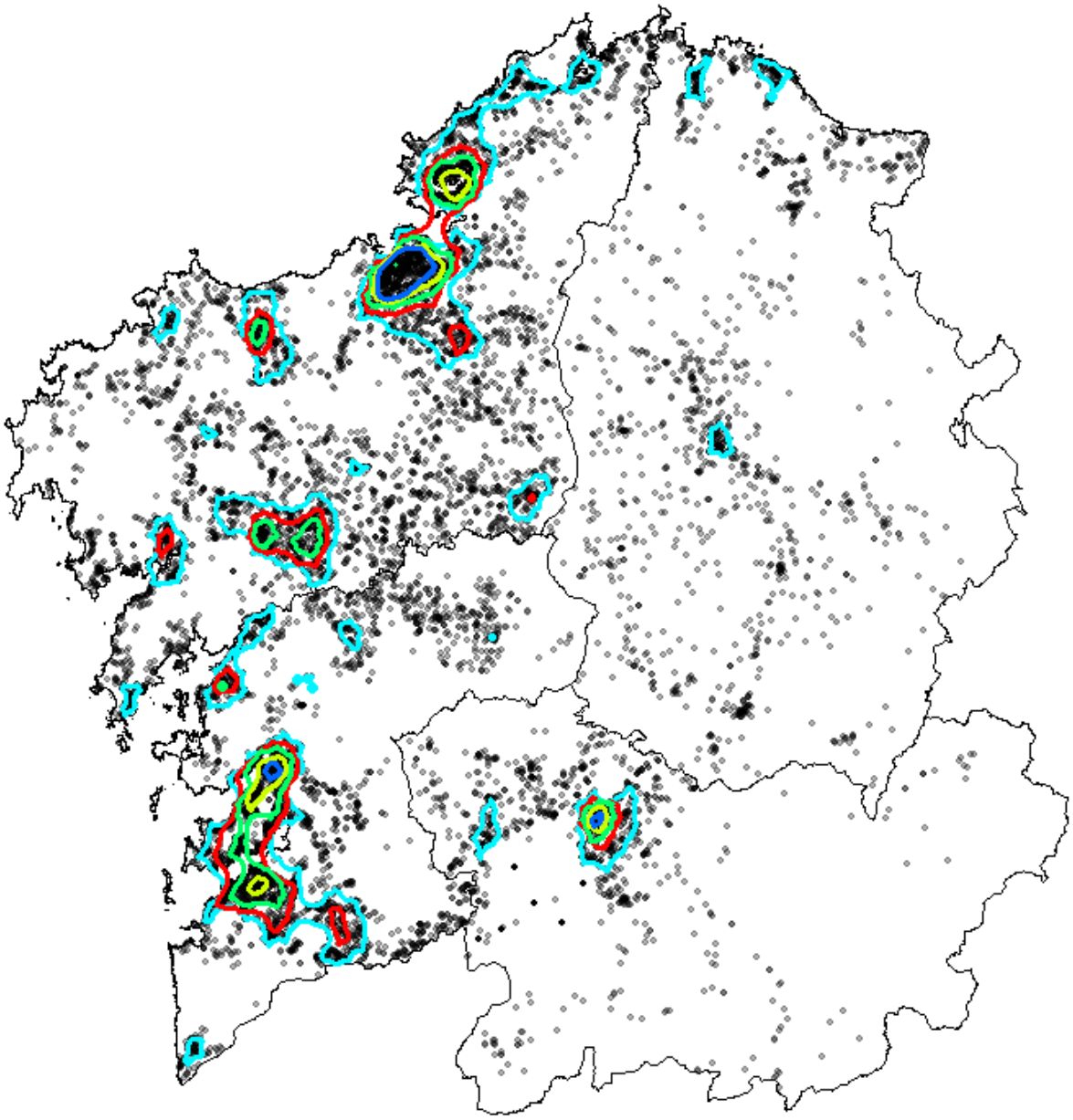


Figura 5.9: La muestra de nidos de velutina del año 2019 en Galicia en negro. Los conjuntos de nivel estimados para $\tau = 0.95, 0.9, 0.8, 0.7$ y 0.5 en color azul (oscuro), amarillo, verde, rojo y turquesa, respectivamente.

Capítulo 6

Conclusiones

En el inicio de este trabajo, en el Capítulo 1 hemos planteado el problema de la reconstrucción de determinados conjuntos de nivel a partir de una muestra aleatoria simple generada a partir de una distribución desconocida y hemos presentamos el conjunto de datos de la localización de los nidos de velutina que nos servirían de ejemplo. En el Capítulo 2 hemos formulado formalmente el problema de la estimación de conjuntos de nivel y presentado las tres metodologías principales para su realización junto a determinadas propiedades geométricas. En particular nos ha interesado aquellas metodologías en las que se asumen restricciones geométricas sobre los conjuntos de nivel, principalmente el método de la envoltura $r-$convexa que supone que el conjunto de nivel a estimar es $r-$convexo. El principal problema de este método es la dependencia sobre el parámetro r que es, por lo general, desconocido.

Hemos comenzado el Capítulo 3 abordando el problema de la estimación de este parámetro r que depende de una sucesión D_n. Primero hemos introducimos el concepto del parámetro r óptimo y hemos mostrado su consistencia uniforme. Posteriormente hemos descrito el método propuesto por Saavedra (2014) y estudiado sus órdenes de convergencia. Por último hemos presentando el algoritmo para la reconstrucción de conjuntos de nivel a partir de una muestra aleatoria simple mediante el método anterior. Para estudiar el buen funcionamiento de este algoritmo, en el Capítulo 4 lo hemos implementado en el software R y hemos realizado un estudio de simulación para comprobar el funcionamiento del algoritmo frente a la variación de los parámetros de entrada y del tamaño de la muestra sobre la que aplicamos el algoritmo. De esta forma hemos comprobado su buen funcionamiento general a medida que aumentamos el tamaño muestral. También hemos estudiado la buena o mala estimación del número de componente conexas del conjunto de nivel estimado, observando que suele ser correcta salvo ante componentes muy cercanas, debido a que el método propuesto tiende a unificarlas, y ante conjuntos de nivel que están cerca de contener una nueva componente conexa.

Para finalizar este trabajo, hemos mostrado una aplicación sobre los datos de la velutina, estudiando a lo largo de los años la focalización de los nidos de la avispa detectados a lo largo del territorio gallego. Hemos mostrado cómo rápidamente se han concentrado los nidos en las zonas más pobladas, principalmente las grandes ciudades y, con un ejemplo en Santiago de Compostela, mostrando cómo la localización dentro de estas ciudades es céntrica, principalmente en las zonas más peatonales.

Una extensión de este trabajo para realizar en el futuro podría ser la consideración de métodos más sofisticados para la estimación de la sucesión D_n. También proponer un test no paramétrico para comparar varias poblaciones en cualquier dimensión midiendo, por ejemplo, la distancia entre las fronteras de los conjuntos de nivel estimados.

Bibliografía

Baíllo, A. and Cuevas, A. (2006). Parametric versus nonparametric tolerance re- gions in detection problems. Comput. Statist. 21 523-536.

Castro L. y Pagola-Carte S. (2010) Vespa velutina Lepeletier, 1836 (Hymenoptera: Vespidae), recolectada en la Península Ibérica. Heteropterus Rev. Entomol 10(2):193-196.

Cuevas, A., Fraiman, R. and Pateiro-López, B. (2012). On statistical properties of sets fulfilling rolling-type conditions. Adv. Appl. Probab. 44 311-329.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multi- variate kernel density estimator. Ann. I. H. Poincaré. 6 907-921.

Grübel, R. (1988). The length of the shorth. Ann. Stat. 16 619-628.

Hartigan, J. (1987). Estimation of a convex density contour in two dimensions.J. Amer. Statist. Assoc. 82 267-270.

Hyndman, R.J. (1996). Computing and graphing highest density regions. Am. Stat. 50 120-126.

INPN (Inventaire National du Patrimoine Naturel) (2020) Le Frelon asiatique Vespa velutina. http://frelonasiatique.mnhn.fr/home/. Accedido 25 de agosto de 2020.

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., and Saavedra-Nieves, P. (2019). Minimax Hausdorff estimation of density level sets. arXiv preprint arXiv:1905.02897.

Rojas-Nossa, S.V., Novoa, N., Serrano, A. et al. (2018). Performance of baited traps used as control tools for the invasive hornet Vespa velutina and their impact on non-target insects. Apidologie 49, 872-885 .

Rortais, A.; Villemant , C.; Gargomin, O.; Rome, Q.; Haxaire, J.; Papachristoforou, A.;, Arnold, G. (2010). A new enemy of honeybees in Europe: the Asian hornet Vespa velutina. Settele J (ed) Atlas of biodiversity risks?from Europe to the globe, from stories to maps. Pensoft, Sofia, p. 11.

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

Samworth, R. and Wand, M. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. Ann. Statist. 38 1767-1792.

Singh, A., Scott, C. and Nowak, R. (2009). Adaptive Hausdorff estimation of density level sets. Ann. Statist. 37 2760-2782.

Viejo, José Luis. (2020). Datos ambientales preliminares del avispón asiático (Vespa velutina Lepeletier, 1836) (Hymenoptera, Vespidae) en Asturias, España. Boletín de la Real Sociedad Española de Historia Natural. Sección biológica. 114. 10.29077/bol/114/ce02_rolea.

Villemant, C.; Haxaire, J. & Streito, J.C. (2006). Premier bilan de l?invasion de Vespa velutina Lepeletier en France (Hymenoptera, Vespidae). Bulletin de la société entomologique de France, 111(4), 535.

Walther, G. (1997). Granulometric Smoothing. Ann. Stat. 25 2273-2299.

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.

Xunta de Galicia. (2016). Protocolo de vixilancia e control fronte á avespa asiática (Vespa velutina). https://d2fyhpf0e1eajp.cloudfront.net/contido/subidas/2017/03/Protocolo_vixilancia_e_control_vespa_velutina_Galicia_Rev_2016.pdf Accedido 25 de agosto del 2020.