USC
UNIVERSIDADE
DE SANTIAGO
DE COMPOSTELA

FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Inferencia estadística en procesos puntuales sobre grafos lineales

Ignacio González Pérez

Julio, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRAO DE MATEMÁTICAS

Traballo Fin de Grao

Inferencia estadística en procesos puntuales sobre grafos lineales

Ignacio González Pérez

Julio, 2022

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Trabajo propuesto

Área de Coñecemento: Estatística e Investigación Operativa
Título: Inferencia Estadística en procesos puntuales sobre grafos lineales
Breve descrición do contido
Este trabajo tiene por objetivo general realizar tareas de Inferencia Estadística en un contexto particular en el que los elementos de interés son eventos o sucesos que ocurren sobre un grafo lineal. La motivación del estudio de este tipo de elementos surge del hecho de que muchos eventos geolocalizados, como pueden ser los accidentes de tráficos, colisiones con animales salvajes, eventos delictivos... tienen lugar sobre un soporte que puede ser caracterizado como un grafo lineal. Trabajar con este soporte lineal embebido en el espacio euclídeo bidimensional supone nuevos retos, asociados especialmente al uso de la métrica del camino más corto que sustituye a la habitual distancia euclídea. Este trabajo se estructurará como sigue: 1) Revisión de herramientas de Inferencia Estadística para espacios bidimensionales. 2) Introducción a los procesos puntuales sobre el plano euclídeo. 3) Introducción a la teoría de grafos lineales. 4) Estimación de la función de densidad en procesos puntuales sobre grafos lineales. 5) Ilustración con simulaciones o bases de datos reales.
Recomendacións
Outras observacións

Índice

Resumen

Este trabajo constituye un recorrido que, desde el nivel del alumnado del Grado en Matemáticas, permite llegar a abordar problemas de inferencia no paramétrica sobre la función de intensidad de procesos puntuales definidos sobre grafos lineales. Para ello, hemos establecido una estructura en capítulos que ha de ser entendida como una consecución de peldaños de complejidad ascendente, en los que, de manera gradual, se van presentando los conceptos y herramientas necesarias hasta llegar al desarrollo del problema final (y objetivo de este trabajo).

Comenzamos introduciendo una serie de conceptos esenciales de la Estadística y la Teoría de la Probabilidad que se necesitan y manejan a lo largo de todo el trabajo. En el Capítulo 1 nos centramos en un problema clásico y bien conocido que es la estimación de la función de densidad, focalizándonos en las técnicas no paramétricas, y en particular en los métodos núcleo. Este capítulo no ha de entenderse una mera introducción, ya que es de esencial importancia en el estudio de los procesos puntuales por la íntima relación existente entre las funciones de densidad y las funciones de intensidad.

En el Capítulo 2, presentamos los procesos puntuales en el plano euclídeo y sus modelos esenciales, como los procesos de Poisson. Una vez que se manejan estos conceptos sobre el plano euclídeo, en el Capítulo 3 se incrementa un grado el nivel de complejidad presentándolos sobre grafos lineales, en donde ya no disfrutamos de ventajas como la existencia de la métrica euclídea o del concepto de gradiente. De nuevo, focalizamos nuestro estudio en los procesos de Poisson, estudiando en profundidad diversas técnicas de estimación no paramétrica de la función de intensidad, así como la cuestión clave de los selectores del parámetro ventana.

Uno de los problemas más interesantes y estudiados en la literatura estadística es el de comparación de dos (o más) poblaciones. El Capítulo 4 se dedica íntegramente a la presentación de este problema en el marco de los procesos puntuales en grafos lineales. Además, se aportan soluciones innovadoras que consisten en tres test estadísticos que permiten concluir si, dados dos patrones de puntos sobre un mismo grafo lineal, provienen de procesos con funciones de intensidad proporcionales, es decir, con una densidad espacial común que se traduce en una misma estructura espacial (en términos de propiedades de primer orden).

En el Capítulo 5 se lleva a cabo un exhaustivo estudio de simulación con varios escenarios en los que se comprueba la calidad y buen comportamiento de los métodos de contraste propuestos en el capítulo anterior. Para ello se presentan tanto resultados sobre el ajuste de cada contraste en distintos tipos de grafos lineales, así como de la potencia de los mismos.

Para finalizar el trabajo, se presenta una aplicación de parte de los métodos descritos para el contraste de dos poblaciones sobre un conjunto de datos reales de accidentes de tráfico en la ciudad de Río de Janeiro (Brasil) entre 2019 y 2022. Además, en el apéndice de este trabajo puede encontrarse el código empleado, con garantías de reproducibilidad. Cabe decir que no se incluye la base de datos reales por cuestiones de confidencialidad.

Abstract

This work constitutes a journey which, starting at an undergraduate level, will allow the addressing of non-parametric problems dealing with point processes on linear network's intensity function. To do so, a chapter structure has been established, which must be understood as a series of complexity-ascending steps. Gradually, the concepts and techniques required to fully understand the final problem (and aim of this work) will be presented.

Firstly, some Statistics and Probability Theory core concepts will be introduced, as they will be used all through our discussion. Chapter 1 is centered on the well-known density function's estimation problem, focusing on non-parametric techniques, particularly kernel methods. This chapter must not be understood as a mere introduction, as it is of utmost importance when studying point processes due to the intimate relation exsiting between density and intensity functions.

In Chapter 2 point processes on the euclidean plane, and their essential models such as Poisson processes, are presented. Once familiarised with these concepts on the euclidean plane, a step forward is taken in Chapter 3, introducing them on linear networks, where the euclidean distance or concepts such as gradient are no longer available. Once again, our study focuses on Poisson processes, studying in great length diverse intensity function's non-parametric estimators, plus the key issue of bandwidth selection.

One of the most interesting, and studied, problems in statistical literature is that of comparing two (or more) populations. Chapter 4 is dedicated integrally to this problem's discussion in the point processes on linear network's framework. Furthermore, innovative solutions consisting of three statistical test are presented. Given two point patterns on the same linear network these

test allow to determine whether those patterns are realizations of two point processes with proportional intensity functions; meaning they share density function, and therefore have the same spatial structure (in terms of first-order properties).

In Chapter 5 an exhaustive simulation study is performed, analysing in various scenarios the quality and well-behaving of the contrast methods proposed in the previous chapter. In order to do so, results of both level and power of those test are presented in different linear networks.

Finally, some of the proposed methods have been applied so as to contrast two populations, based on a real-life dataset of traffic accidents in Río de Janeiro (Brasil) from 2019 up to 2022. Moreover, in this work's appendix the code which has been used, with guaranteed reproducibility, can be found. It is worth mentioning that the database has not been included due to confidentiality issues.

Introducción

A la hora de estudiar cualquier proceso aleatorio, el concepto capital es el de variable aleatoria. Para definir de forma precisa este concepto, partimos del conjunto de posibles resultados del proceso aleatorio que estemos estudiando, el denominado espacio muestral Ω. Ahora bien, no cualquier subconjunto del espacio muestral es susceptible al estudio de probabilidades. Por ello, debemos escoger un subconjunto de $\mathrm{P}(\Omega) = \{\tau \colon \tau \subset \Omega\}$ suficientemente manejable. Se introduce así el concepto de σ-álgebra:

Definición 0.1. Diremos que $\mathscr{A} \subset \mathrm{P}(\Omega)$ es una **$\boldsymbol{\sigma}$-álgebra** de Ω si verifica que: (i) $\emptyset \in \mathscr{A}$, (ii) $\Omega \backslash A \in \mathscr{A}$ $\forall A \in \mathscr{A}$, y (iii) $\{A_i\}_{i\in\mathbb{N}} \subset \mathscr{A} \Rightarrow \bigcup_{i\in\mathbb{N}} A_i \in \mathscr{A}$

Cuando $\Omega = \mathbb{R}^m$, resulta usual la elección de la σ-álgebra de Borel, β_m, que es la menor σ-álgebra sobre $\mathbb{R}^m$ que contiene a todos los abiertos. Al par $(\Omega, \mathscr{A})$, con $\mathscr{A}$ una σ-álgebra sobre Ω, se le denomina espacio de medida. Este concepto permite definir ya el de probabilidad:

Definición 0.2. Dado un espacio de medida $(\Omega, \mathscr{A})$, se define una **probabilidad** sobre este espacio como una aplicación $\mathbb{P} : \mathscr{A} \to [0,1]$ tal que $\mathbb{P}(\Omega) = 1$ y tal que si $\{A_i\}_{i\in\mathbb{N}} \subset \mathscr{A}$ es tal que $A_i \cap A_j = \emptyset$ $\forall i \neq j$ entonces $\mathbb{P}\left(\bigcup_{i\in\mathbb{N}} A_i\right) = \sum_{i\in\mathbb{N}} \mathbb{P}(A_i)$

A la terna $(\Omega, \mathscr{A}, \mathbb{P})$, donde $(\Omega, \mathscr{A})$ es un espacio de medida, y $\mathbb{P}$ es una probabilidad sobre este, se le denomina espacio de probabilidad. En este punto ya podemos definir el concepto central como es el de vector aleatorio:

Definición 0.3. Dado un espacio de probabilidad $(\Omega, \mathscr{A}, \mathbb{P})$, un **vector aleatorio** m-dimensional $\boldsymbol{X}$ es una aplicación $\boldsymbol{X} : (\Omega, \mathscr{A}, \mathbb{P}) \to (\mathbb{R}^m, \beta_m, \mathbb{P}^*)$, siendo $(\mathbb{R}^m, \beta_m, \mathbb{P}^*)$ el espacio de probabilidad inducido por $\boldsymbol{X}$, estando $\mathbb{P}^*$ definida como $\mathbb{P}^*(B) = \mathbb{P}[\boldsymbol{X}^{-1}(B)]$ $\forall B \in \beta_m$. Para garantizar que $\mathbb{P}^*$ esté bien definida, $\boldsymbol{X}$ ha de ser una aplicación medible; es decir, tal que $\boldsymbol{X}^{-1}(B) \in \mathscr{A}$ $\forall B \in \beta_m$.

La naturaleza de los vectores aleatorios puede ser muy diversa. Indistintamente, nosotros nos centraremos en los vectores aleatorios absolutamente continuos:

Definición 0.4. Sea $\boldsymbol{X} = (X_1, \ldots, X_m)'$ un vector aleatorio m-dimensional. Este se dice **absolutamente continuo** si existe una función $f : \mathbb{R}^m \to \mathbb{R}$, que llamaremos su función de densidad de probabilidad, tal que:

$$\mathbb{P}(\boldsymbol{X} \in A) = \int_A f(\boldsymbol{x}) d\boldsymbol{x},$$

para cualquier $A \subset \mathbb{R}^m$ perteneciente a la σ-álgebra relativa al espacio de probabilidad sobre el que está definido $\boldsymbol{X}$.

De las propiedades de la probabilidad se deduce que necesariamente f es una función no negativa y que $\int_{\mathbb{R}^m} f(\boldsymbol{x}) d\boldsymbol{x} = 1$. Además, cualquier función que verifique estas dos propiedades determina de forma unívoca una distribución de probabilidad en $\mathbb{R}^m$; es decir, las dos condiciones anteriores caracterizan las funciones de densidad de los vectores aleatorios absolutamente continuos. En este mismo contexto podemos definir la función de distribución de $\boldsymbol{X}$ y relacionarla con su función de densidad, como:

Definición 0.5. Sea $\boldsymbol{X}$ un vector aleatorio m-dimensional. Definimos su **función de distribución** (conjunta) como la aplicación $F : \mathbb{R}^m \to \mathbb{R}$ tal que:

$$F(\boldsymbol{x}) = F(x_1, \ldots, x_m) = \mathbb{P}(X_1 \leq x_1, \ldots, X_m \leq x_m) = \int_{-\infty}^{x_m} \cdots \int_{-\infty}^{x_1} f(y_1, \ldots, y_m) dy_1 \cdots dy_m,$$

$\forall \boldsymbol{x} = (x_1, \ldots, x_m)' \in \mathbb{R}^m$, lo que denotaremos usualmente por $F(\boldsymbol{x}) = \mathbb{P}(\boldsymbol{X} \leq \boldsymbol{x})$.

Capítulo 1

Estimación de la función de densidad en espacios euclídeos

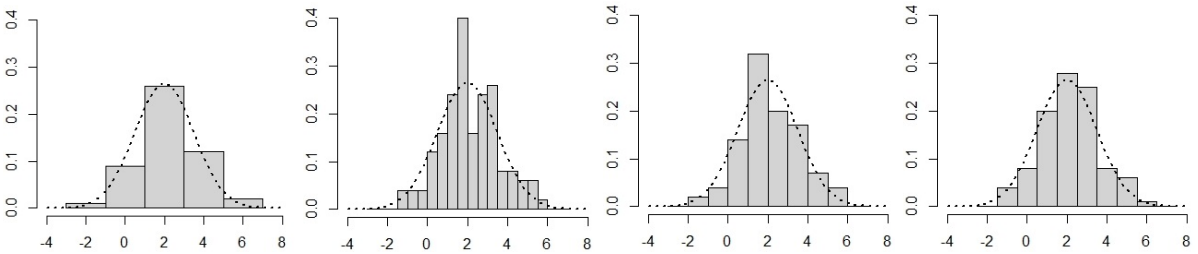
A la hora de estimar la función de densidad de un vector aleatorio $\boldsymbol{X}$, el escenario usual suele ser contar con una muestra aleatoria simple de este vector; es decir, un conjunto de vectores aleatorios $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ independientes entre sí e idénticamente distribuidos a $\boldsymbol{X}$ cuya realización nos proporcionará una muestra. En aquellos casos en los que se pueda suponer que la distribución de $\boldsymbol{X}$ pertenece a una familia paramétrica conocida, los parámetros desconocidos pueden estimarse a partir de la muestra empleando técnicas de momentos, máxima verosimilitud, etc. Este es el contexto de la estimación paramétrica de la función de densidad.

Si carecemos de información acerca de la distribución de $\boldsymbol{X}$, lo más natural es buscar un método que permita estimar la función de densidad "dejando a los datos hablar por sí mismos"; es decir, sin imponer ningún tipo de restricción a la forma de f. Entramos así en el contexto de la estimación no paramétrica de la función de densidad.

1.1. Estimación no paramétrica de la densidad unidimensional

Supongamos inicialmente que nos encontramos en un caso unidimensional. La estimación más elemental de la densidad es la frecuencia relativa, que toma la forma de un histograma. Dividiendo la recta real en intervalos de igual longitud h[1], que denominamos ventanas o bandas, la estimación de la función de densidad dada por el histograma en un punto $x \in \mathbb{R}$ es:

$$\widehat{f}(x;h) = \frac{\text{nº de observaciones en la ventana que contiene a } x}{nh}.$$



(a) $h = 2,\ x_0 = -3.$ (b) $h = 0.5,\ x_0 = -3.$ (c) $h = 1,\ x_0 = -3.$ (d) $h = 1,\ x_0 = -2.5.$

Figura 1.1: distintos histogramas para una muestra de tamaño 100 de la distribución $\mathcal{N}(2, 2.25)$, donde h es el ancho de banda y x_0 el extremo izquierdo de la primera ventana. La densidad real se representa de forma punteada.

A la hora de construir un estimador como este, han de tomarse dos decisiones: el ancho de banda, h, y el punto inicial x_0. Ambas decisiones son capitales de cara a la forma final del estimador. En la Figura 1.1 vemos como diferentes elecciones del ancho de banda o del punto inicial generan estimadores considerablemente distintos para la misma muestra. Este problema se soluciona, en primera instancia, introduciendo el concepto de histograma móvil. Fijado un ancho de banda h, para cada $i \in \{1, \ldots, n\}$ se considera el histograma que tiene como única observación X_i y toma una ventana centrada en este punto. El estimador $\widehat{f}$ se obtiene promediando todos estos histogramas. De esta forma, podemos escribir:

$$\widehat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} \mathbf{1}_{[-1/2,1/2]} \left(\frac{x - X_i}{h} \right).$$

Si notamos que $\mathbf{1}_{[-1/2,1/2]}$ es la función de densidad de una distribución uniforme en $[-1/2, 1/2]$, podemos plantearnos sustituirla por otras funciones de densidad k que varíen más suavemente con la distancia, *suavizando* nuestro estimador. Surgen así los estimadores tipo núcleo de la función de densidad:

$$\widehat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} k \left(\frac{x - X_i}{h} \right) = \frac{1}{n} \sum_{i=1}^{n} k_h(x - X_i), \tag{1.1}$$

donde hemos introducido la notación $k_h(\cdot) = h^{-1}k(\cdot/h)$. La función núcleo k ha de ser una densidad unimodal simétrica respecto al cero. Estas condiciones aseguran que el estimador $\widehat{f}$ sigue siendo una función de densidad.

Uno de los criterios de referencia para determinar la calidad de un estimador es su error cuadrático medio, denotado por sus siglas en inglés, *Mean Square Error*:

[1] En un contexto más general, estos intervalos no tienen que tener la misma longitud.

Definición 1.1. Sea θ una cantidad desconocida y $\hat{\theta}$ un estimador de θ. Se define el **error cuadrático medio** de este estimador como: $\text{MSE}\left(\hat{\theta}\right) = \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathbb{V}\text{ar}\left(\hat{\theta}\right) + \left[\mathbb{E}\left(\hat{\theta}\right)-\theta\right]^2$.

Si f es la verdadera densidad de X, tenemos que, bajo muestreo aleatorio simple:

$$\mathbb{E}\left[\widehat{f}(x;h)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[k_h(x-X_i)\right] = \mathbb{E}\left[k_h(x-X)\right] = \int k_h(x-y)f(y)dy = \int k(z)f(x-hz)dz.$$

Antes de entrar en más detalle debemos explicitar las hipótesis con las que vamos a trabajar. Bajo estas podremos hacer cálculos asintóticos del MSE del estimador tipo núcleo, así como asegurar el buen comportamiento de este. Las hipótesis son:

(U_i) El ancho de banda $h \equiv h_n$ se representa por una sucesión no aleatoria de valores que verifican, omitiendo el subíndice n, que $\lim_{n\to\infty} h = 0$ pero $\lim_{n\to\infty} nh = \infty$.

(U_{ii}) El núcleo k es una densidad de probabilidad acotada, simétrica entorno al origen, y con momento $\mu_l(k) = \int x^l k(x)dx$ finito para $l = 0, 1, 2$.

(U_{iii}) La densidad f es tal que f'' es continua y de cuadrado integrable.

Antes de realizar estos cálculos asintóticos debemos introducir el concepto de "o pequeña". Dadas $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ dos sucesiones de números reales, diremos que $(a_n)_{n\in\mathbb{N}}$ es una **o pequeña** de $(b_n)_{n\in\mathbb{N}}$, y escribiremos $a_n = o(b_n)$, si $\lim_{n\to\infty} |a_n/b_n| = 0$. Notemos que bajo la hipótesis (U_i) cualquier función de h puede entenderse como una sucesión de números reales.
La hipótesis (U_{iii}) permite hacer un desarrollo de Taylor de $f(x-hz)$ centrado en x hasta orden 2, como puede consultarse en [10]. De esta forma:

$$\begin{aligned}\mathbb{E}\left[\widehat{f}(x;h)\right] &= f(x)\int k(z)dz - hf'(x)\int zk(z)dz + \frac{1}{2}h^2 f''(x)\int z^2 k(z)dz + o(h^2)\\ &= f(x) + \frac{1}{2}h^2 f''(x)\mu_2(k) + o(h^2),\end{aligned} \tag{1.2}$$

en donde debemos notar que, bajo las hipótesis previas, $\mu_1(k) = 0$. Además, la hipótesis (U_i) garantiza la insesgadez asintótica del estimador; en efecto, ya que $\lim_{n\to\infty}\mathbb{E}\left[\widehat{f}(x;h)\right] = f(x)$. Analicemos ahora el término asociado a la varianza. Empleando la hipótesis de que tenemos muestreo aleatorio simple:

$$\begin{aligned}\mathbb{V}\text{ar}\left[\widehat{f}(x;h)\right] &= \mathbb{V}\text{ar}\left[\frac{1}{n}\sum_{i=1}^{n}k_h(x-X_i)\right] = \frac{1}{n}\mathbb{V}\text{ar}[k_h(x-X)]\\ &= \frac{1}{n}\int\left[k_h(x-y) - \mathbb{E}\left[k_h(x-X)\right]\right]^2 f(y)dy\\ &= \frac{1}{n}\left[\int k_h^2(x-y)f(y)dy - 2\mathbb{E}[k_h(x-X)]\int k_h(x-y)f(y)dy\right.\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}[k_h(x-X)]^2 \int f(y)dy\Big] \\
&= \frac{1}{n}\int k_h^2(x-y)f(y)dy - \frac{1}{n}\mathbb{E}[k_h(x-X)]^2 \\
&= \frac{1}{nh}\int k^2(z)f(x-hz)dz - \frac{1}{n}\mathbb{E}[\widehat{f}(x;h)]^2.
\end{aligned}
$$

Haciendo ahora un desarrollo de Taylor de f centrado en x, pero ahora a orden cero, y empleando el resultado de la ecuación (1.2), vemos que:

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{f}(x;h)\right] &= \frac{1}{nh}\int k^2(z)\left[f(x)+o(1)\right]dz - \frac{1}{n}[f(x)+o(h)]^2 \\
&= \frac{1}{nh}f(x)\int k^2(z)dz + o[(nh)^{-1}] - o(n^{-1}) \\
&= (nh)^{-1}R(k)f(x) + o[(nh)^{-1}],
\end{aligned}
\tag{1.3}
$$

donde hemos definido $R(g) = \int g^2(x)dx$. Debemos notar que la hipótesis (U_i) garantiza que la varianza del estimador converge a cero, lo que nos permite concluir que el estimador tipo núcleo es asintóticamente consistente. Teniendo entonces en cuenta los resultados obtenidos en las ecuaciones (1.2) y (1.3), concluimos que el error cuadrático medio del estimador tipo núcleo de f en x es:

$$
\mathrm{MSE}\left[\widehat{f}(x;h)\right] = (nh)^{-1}R(k)f(x) + \frac{h^4}{4}\mu_2(k)^2 f''(x)^2 + o\left[(nh)^{-1}+h^4\right]. \tag{1.4}
$$

El principal problema que presenta la expresión obtenida en la ecuación (1.4) es que se trata de un indicador *local* de la calidad de la estimación. Para obtener un indicador global, integramos esta expresión para obtener el denominado error cuadrático medio integrado MISE:

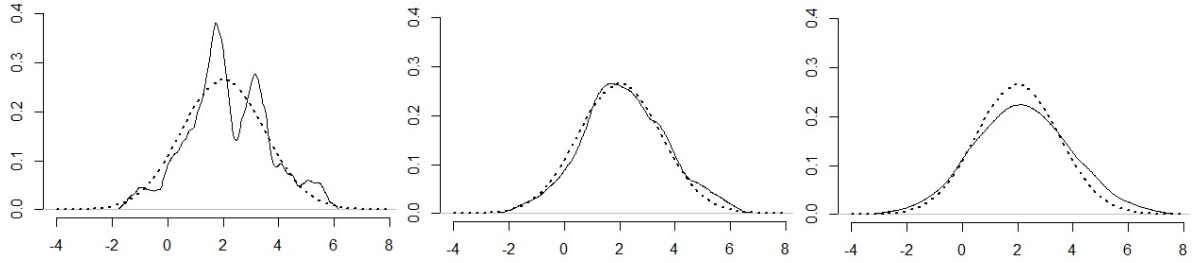
$$
\mathrm{MISE}\left[\widehat{f}(\cdot;h)\right] = \int \mathrm{MSE}\left[\widehat{f}(x;h)\right]dx = (nh)^{-1}R(k) + \frac{h^4}{4}\mu_2(k)^2 R(f'') + o\left[(nh)^{-1}+h^4\right].
$$

De quedarnos con los términos dominantes, obtenemos una estimación asintótica global de la calidad de nuestro estimador no paramétrico, el denominado error cuadrático medio integrado asintótico AMISE:

$$
\mathrm{AMISE}\left[\widehat{f}(\cdot;h)\right] = (nh)^{-1}R(k) + \frac{h^4}{4}\mu_2(k)^2 R(f''), \tag{1.5}
$$

donde el primer término es relativo a la varianza, mientras que el segundo es el relativo al sesgo cuadrático. Notemos que, dada la muestra, según $h \to 0$ el sesgo cuadrático decrece mientras que el término asociado a la varianza aumenta; es decir, según tomamos anchos de banda más pequeños vamos obteniendo estimadores con menor sesgo pero con mayor varianza. Esto es conoce como compensación varianza-sesgo, o en la literatura anglosajona como *variance-bias trade-off*.

Este comportamiento se ejemplifica en la Figura 1.2; en (a) tenemos un ancho de banda demasiado pequeño que genera una estimación con excesiva varianza pero poco sesgo, lo que nos



(a) $h = 0.2$. (b) $h = 0.5$. (c) $h = 1$.

Figura 1.2: estimaciones tipo núcleo para una muestra empleada en la Figura 1.1. Se ha empleado en todas en núcleo de Epanechnikov con distintos anchos de banda. La densidad real se representa de forma punteada.

lleva a decir que nuestro estimador está infrasuavizado. Según aumentamos h, vemos como el estimador se suaviza buscando un equilibrio sesgo-varianza tal como ocurre en (b). Finalmente, anchos de banda demasiado grandes generan estimadores con excesivo sesgo, denominados sobresuavizados, como el que vemos en (c). Esta compensación del error de estimación entre la varianza y el sesgo cuadrático motiva buscar el ancho de banda óptimo. Una opción es tomar h tal que se minimice el AMISE calculado en la ecuación (1.5). Derivando respecto a h, se concluye que el ancho de banda óptimo para este criterio es:

$$h_{\text{AMISE}} = \left[\frac{R(k)}{n\mu_2(k)^2 R(f'')}\right]^{1/5}, \tag{1.6}$$

que si lo sustituimos de vuelta en la ecuación (1.5), llegamos a que:

$$\inf_{h>0} \text{AMISE}\left[\widehat{f}(\cdot;h)\right] = \frac{5}{4}\left[\mu_2(k)^2 R(k)^4 R(f'')\right]^{1/5} n^{-4/5}. \tag{1.7}$$

De la ecuación (1.7) deducimos que el orden de consistencia (asintótico) de nuestro estimador no paramétrico de la función de densidad es $n^{-4/5}$, que es menor que la tasa usual de convergencia paramétrica de n^{-1}. Esto nos indica que, aunque mucho más versátiles, los estimadores tipo núcleo no convergen con tanta rapidez como los paramétricos.

A la hora de elegir el ancho de banda con el que construir un estimador tipo núcleo, lo lógico sería emplear el obtenido en la ecuación (1.6). El problema reside en que este depende de cantidades teóricas que son desconocidas, pues hacen referencia a la densidad de X. Por ello, debemos determinar métodos que permitan seleccionar un h que proporcione una estimación de calidad.

Una primera idea, propuesta en [27], es asumir la hipótesis de que nuestra muestra procede de una población normal. En este caso podemos calcular $R(f'') = 3/(8\sqrt{\pi}\sigma^5)$, lo que permite

estimar el ancho de banda óptimo para el AMISE como:

$$h_{\text{Norm}} = \left[\frac{8\sqrt{\pi}R(k)}{3n\mu_2(k)^2}\right]^{1/5}\sigma.$$

Ahora bien, en muchos de los casos ni siquiera se tienen indicios de normalidad; por ello, se han diseñado métodos más generales que permitan obtener un ancho de banda efectivo en ausencia de más hipótesis sobre la población que las que tenemos hasta el momento. Primeramente, notamos que el error cuadrático medio integrado puede descomponerse como:

$$\text{MISE}\left[\widehat{f}(\cdot;h))\right] = \mathbb{E}\left[\int \widehat{f}(x;h)^2dx\right] - 2\mathbb{E}\left[\int \widehat{f}(x;h)f(x)dx\right] + \int f(x)^2dx,$$

y como $\int f^2(x)dx$ no depende de h, minimizar el MISE equivale a minimizar:

$$\text{MISE}\left[\widehat{f}(\cdot;h)\right] - \int f(x)^2dx = \mathbb{E}\left[\int \widehat{f}(x;h)^2dx - 2\int \widehat{f}(x;h)f(x)dx\right]. \tag{1.8}$$

A priori puede parecer que tenemos el mismo problema que con el ancho de banda óptimo en la ecuación (1.6), ya que el segundo sumando de la función objetivo anterior depende de f, que es desconocida. Ahora bien, si notamos que $\int \widehat{f}(x;h)f(x)dx = \mathbb{E}\left[\widehat{f}(X,h)\right]$, podemos tratar de estimar esta cantidad mediante $n^{-1}\sum_{i=1}^n \widehat{f}(x_i,h)$. El problema con este estimador, como puede consultarse en [27], es que presenta un fuerte sesgo negativo al evaluar $\widehat{f}$ en puntos de la muestra. Para solucionar este problema, Bowman propone en [7] sustituir $\widehat{f}(x_i,h)$ por la estimación no paramétrica de la densidad en x_i empleando todos los datos de la muestra a excepción del propio x_i, que denotaremos por $\widehat{f}_{-i}(x_i,h)$. Surge así la función del ancho de banda:

$$\text{LSCV}(h) = \int \widehat{f}(x;h)^2dx - \frac{2}{n}\sum_{i=1}^n \widehat{f}_{-i}(X_i;h),$$

que es un estimador insesgado de la función objetivo dada en la ecuación (1.8). En efecto, tenemos primeramente que:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \widehat{f}_{-i}(X_i;h)\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[\widehat{f}_{-i}(X_i;h)\right] = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{\substack{j=1\\j\neq i}}^n \mathbb{E}\left[k_h(X_i - X_j)\right].$$

Dadas X y Z variables aleatorias independientes e idénticamente distribuidas con densidad f, la función de densidad de $Y = X - Z$ es $g(y) = \int f(x-y)f(x)dx$. Así, bajo muestreo aleatorio simple:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \widehat{f}_{-i}(X_i;h)\right] = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{\substack{j=1\\j\neq i}}^n \int k_h(y)g(y)dy = \iint k_h(y)f(x)f(x-y)dxdy.$$

Veamos que la esperanza del segundo sumando de la ecuación (1.8) coincide con este resultado. En efecto:

$$\begin{aligned}\mathbb{E}\int \widehat{f}(x;h)f(x)dx &= \int f(x)\mathbb{E}\left[\widehat{f}(x;h)\right]dx = \int f(x)\left[\int k_h(x-y)f(y)dy\right]dx \\ &= \iint k_h(x-y)f(x)f(y)dxdy = \iint k_h(y)f(x)f(x-y)dxdy,\end{aligned}$$

tal y como queríamos ver. Así, LSCV es un estimador insesgado de la función objetivo dada en la ecuación (1.8). Por ello, eligiendo el ancho de banda h que minimice LSCV(h) (que notemos es una función que podemos computar para cada h) obtenemos un ancho de banda que, presumiblemente, conducirá a una estimación de f con un MISE cercano a su valor mínimo. El empleo de este tipo de estimadores, en los que se estiman cantidades evaluadas en una observación empleando todas las demás, recibe el nombre de validación cruzada (en inglés cross-validation). Es por ello que esta técnica para calcular un h óptimo se conozca como *Least Squares Cross-Validation*, ya que se emplean técnicas de validación cruzada para minimizar el error cuadrático.

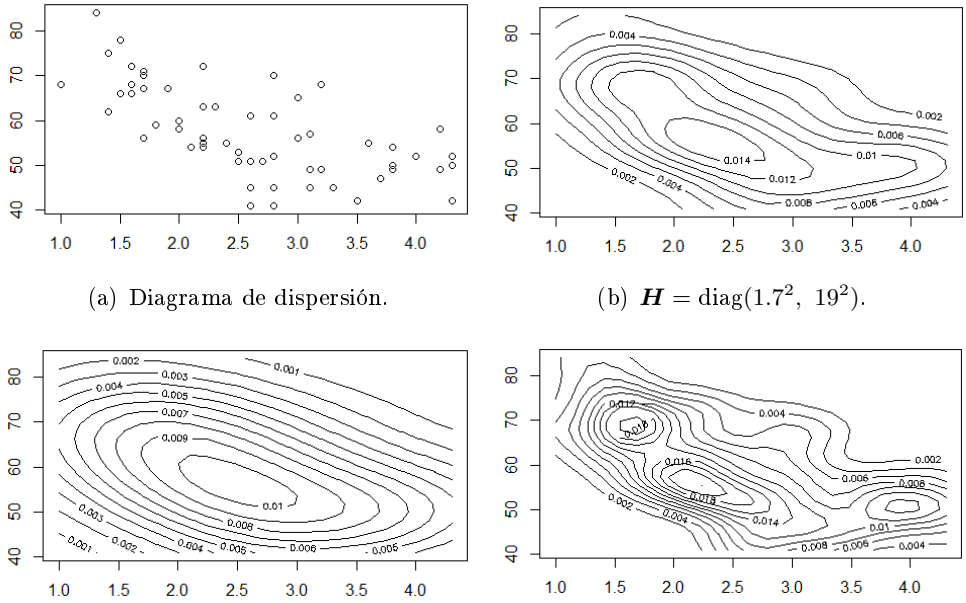
Técnicas más elaboradas de selección del ancho de banda construyen procesos en varias etapas, tras cada una de las cuales se obtiene un ancho de banda que proporciona una estimación de mejor calidad. Otros métodos consisten también en elegir otros funcionales que estimen el MISE, y tratar de minimizarlos. Estas técnicas de selección de h, junto con muchas otras, pueden consultarse en [27].

1.2. Estimación no paramétrica de la densidad multidimensional

En un escenario multivariante, $\boldsymbol{X}$ es ahora un vector aleatorio d-dimensional del cual poseemos una m.a.s. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. A la hora de extender nuestro estimador no paramétrico de la función de densidad al caso multivariante, la principal diferencia se encuentra en que el ancho de banda h pasa a ser una matriz de ancho de banda $\boldsymbol{H}$. Tomaremos $\boldsymbol{H} \in \mathcal{F} = \{A \in \mathcal{M}_{d\times d}(\mathbb{R})\colon A \text{ simétrica y definida positiva}\}$. Esto es natural, ya que si recordamos que en el caso univariante h medía la dispersión asociada a los núcleos promediados, en el caso multivariante jugará el papel de una matriz de covarianzas. De esta forma, el estimador tipo núcleo d-dimensional de la función de densidad, como puede consultarse en [27], es:

$$\widehat{f}_{\mathcal{F}}(\boldsymbol{x};\boldsymbol{H}) = \frac{1}{n}\sum_{i=1}^{n} K_{\boldsymbol{H}}(\boldsymbol{x}-\boldsymbol{X}_i), \text{ donde } K_{\boldsymbol{H}}(\boldsymbol{x}) = |\boldsymbol{H}|^{-1/2}K(\boldsymbol{H}^{-1/2}\boldsymbol{x}),$$

siendo la función núcleo K una densidad de probabilidad en $\mathbb{R}^d$. Al ser $\boldsymbol{H}$ una matriz $d \times d$ simétrica, posee $d(d+1)/2$ entradas independientes. Con el objetivo de simplificar el problema de selección de esta matriz (en el cual profundizaremos más adelante) es costumbre imponer restricciones sobre la forma de $\boldsymbol{H}$, como por ejemplo $\boldsymbol{H} \in \mathcal{D} = \{\text{diag}(h_1^2, \ldots, h_d^2) \in \mathcal{M}_{d\times d}(\mathbb{R})\colon \boldsymbol{h} \in \mathbb{R}^d\}$,



(a) Diagrama de dispersión.

(b) $\boldsymbol{H} = \mathrm{diag}(1.7^2,\ 19^2)$.

(c) $\boldsymbol{H} = \mathrm{diag}(3^2,\ 30^2)$.

(d) $\boldsymbol{H} = \mathrm{diag}(1,\ 14^2)$.

Figura 1.3: diagrama de dispersión (a) y curvas de nivel de estimaciones de la densidad bivariante. Se ha empleado un núcleo normal bivariante alineado con los ejes y varias matrices $\boldsymbol{H} \in \mathcal{D}$. Datos extraídos de [26].

en cuyo caso el estimador pasa a tener la forma:

$$\widehat{f}_{\mathcal{D}}(\boldsymbol{x};\boldsymbol{h}) = \frac{1}{n}\left(\prod_{j=1}^{d} h_j\right)^{-1} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}, \ldots, \frac{x_d - X_{id}}{h_d}\right).$$

Un ejemplo de estimación en el caso bivariante empleando una matriz $\boldsymbol{H}$ de esta clase puede verse en la Figura 1.3, donde podemos apreciar nuevamente el compensación varianza-sesgo según varían los valores de h_j. Una mayor simplificación es tomar $\boldsymbol{H} \in \mathcal{S} = \{h^2\boldsymbol{I}_d \colon h > 0\}$. Este último caso presenta especial interés tanto teórico como práctico debido a la simplificación del estimador, el cual pasa a ser:

$$\widehat{f}_{\mathcal{S}}(\boldsymbol{x};h) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h}\right). \tag{1.9}$$

Para estudiar las propiedades de estos estimadores, debemos calcular el AMISE. Por simplicidad, nos restringiremos a estimadores de la forma dados en la ecuación (1.9), debido a que presentan la mejor relación coste-beneficio tanto conceptual como computacional. Para poder obtener estimaciones asintóticas del error cuadrático de los estimadores anteriores, emplearemos la siguiente versión débil del teorema de Taylor en varias variables:

Teorema 1.2. *Sea $g \in \mathcal{C}^2(\mathbb{R}^m, \mathbb{R})$, y $(\boldsymbol{\alpha}_n)_{n\in\mathbb{N}}$ una sucesión de vectores de $\mathbb{R}^m$ convergente a $\mathbf{0}$. Denotemos por $\nabla g(\boldsymbol{x})$ al vector gradiente de g en $\boldsymbol{x}$, y por $Hg(\boldsymbol{x})$ a la matriz Hessiana de g en $\boldsymbol{x}$; es decir, la matriz $m \times m$ tal que $(Hg(\boldsymbol{x}))_{ij} = D_{ij}g(\boldsymbol{x})$. Además, dada $A \in \mathcal{M}_{m\times m}(\mathbb{R})$ definimos su traza como $tr(A) = \sum_{i=1}^m A_{ii}$. Tenemos entonces que:*

$$g(\boldsymbol{x}+\boldsymbol{\alpha}_n) = g(\boldsymbol{x}) + \boldsymbol{\alpha}_n'\nabla g(\boldsymbol{x}) + \frac{1}{2}tr\left[Hg(\boldsymbol{x})\boldsymbol{\alpha}_n\boldsymbol{\alpha}_n'\right] + o(\boldsymbol{\alpha}_n'\boldsymbol{\alpha}_n).$$

Versiones más generales del teorema de Taylor en varias variables pueden consultarse en [24]. Además del resultado anterior, para obtener dichas estimaciones, así como para poder asegurar el buen comportamiento del estimador dado en la ecuación (1.9), asumiremos las siguientes hipótesis establecidas en [27]:

(M_i) Las entradas de la matriz hessiana de f, Hf son continuas y de cuadrado integrable.

(M_{ii}) El ancho de banda h se representa por una sucesión no aleatoria de valores que verifican, omitiendo el subíndice n, que $\lim_{n\to\infty} h = 0$ pero $\lim_{n\to\infty} nh^d = \infty$.

(M_{iii}) La función núcleo K posee un soporte compacto, está acotada, y es tal que: $\int K(\boldsymbol{z})d\boldsymbol{z} = 1$, $\int \boldsymbol{z}K(\boldsymbol{z})d\boldsymbol{z} = \mathbf{0}$, y $\int \boldsymbol{z}\boldsymbol{z}'K(\boldsymbol{z})d\boldsymbol{z} = \mu_2(K)\boldsymbol{I}_d$; siendo $\mu_2(K) = \int z_i^2 K(\boldsymbol{z})d\boldsymbol{z}$ invariante con i.

Para la esperanza tenemos entonces que:

$$\begin{aligned}
\mathbb{E}\left[\widehat{f}_S(\boldsymbol{x};h)\right] &= \frac{1}{h^d}\mathbb{E}\left[K\left(\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right)\right] = \frac{1}{h^d}\int K\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{h}\right)f(\boldsymbol{y})d\boldsymbol{y} = \int K(\boldsymbol{z})f(\boldsymbol{x}-h\boldsymbol{z})d\boldsymbol{z} \\
&= \int K(\boldsymbol{z})\left[f(\boldsymbol{x}) - h\boldsymbol{z}'\nabla f(\boldsymbol{x}) + \frac{h^2}{2}tr\left[Hf(\boldsymbol{x})\boldsymbol{z}\boldsymbol{z}'\right] + o(h^2)\right]d\boldsymbol{z} \\
&= f(\boldsymbol{x}) - h\nabla f(\boldsymbol{x})'\int \boldsymbol{z}K(\boldsymbol{z})d\boldsymbol{z} + \frac{h^2}{2}\int K(\boldsymbol{z})tr\left[Hf(\boldsymbol{x})\boldsymbol{z}\boldsymbol{z}'\right]d\boldsymbol{z} + o(h^2) \\
&= f(\boldsymbol{x}) + \frac{h^2}{2}tr\left[Hf(\boldsymbol{x})\int \boldsymbol{z}\boldsymbol{z}'K(\boldsymbol{z})d\boldsymbol{z}\right] + o(h^2) \\
&= f(\boldsymbol{x}) + \frac{h^2}{2}\mu_2(K)tr\left[Hf(\boldsymbol{x})\right] + o(h^2).
\end{aligned}$$

El resultado anterior bajo la hipótesis (M_{ii}) nos permite concluir, al igual que ocurría univariantemente, que el estimador dado en la ecuación (1.9) es asintóticamente insesgado. La varianza de este estimador es:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{f}_S(\boldsymbol{x};h)\right] &= \frac{1}{nh^{2d}}\mathbb{V}\mathrm{ar}\left[K\left(\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right)\right] = \frac{1}{nh^{2d}}\left[\mathbb{E}\left[K\left(\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right)^2\right] - \mathbb{E}\left[K\left(\frac{\boldsymbol{x}-\boldsymbol{X}}{h}\right)\right]^2\right] \\
&= \frac{1}{nh^{2d}}\int K\left(\frac{\boldsymbol{x}-\boldsymbol{y}}{h}\right)^2 f(\boldsymbol{y})d\boldsymbol{y} - \frac{1}{n}\mathbb{E}\left[\widehat{f}(\boldsymbol{x};h)\right]^2 \\
&= \frac{1}{nh^d}\int K(\boldsymbol{z})^2 f(\boldsymbol{x}-h\boldsymbol{z})d\boldsymbol{z} - \frac{1}{n}\left[f(\boldsymbol{x}) + o(h)\right]^2
\end{aligned}$$

$$= \frac{1}{nh^d} \int K(\boldsymbol{z})^2 \left[f(\boldsymbol{x}) + o(1) \right] d\boldsymbol{z} - o(n^{-1}) = \frac{1}{nh^d} f(\boldsymbol{x}) R(K) + o(n^{-1}h^{-d}).$$

Nuevamente, la hipótesis (M_{ii}) garantiza la consistencia asintótica del estimador multivariante definido en la ecuación (1.9). En virtud de lo expuesto hasta el momento, concluimos que su error cuadrático medio integrado asintótico es:

$$\text{AMISE}\left[\widehat{f}(\cdot;h)\right] = \frac{R(K)}{nh^d} + \frac{h^4}{4}\mu_2(K)^2 \left[\int tr\left[Hf(\boldsymbol{x})\right]^2 d\boldsymbol{x}\right]^2, \tag{1.10}$$

que podemos ver como una extensión casi directa del resultado univariante, dado en la ecuación (1.5), al caso d-dimensional. La ecuación (1.10) pone de manifiesto que la compensación varianza-sesgo está también presente en el caso multivariante, como se ejemplifica en la Figura 1.3. Ahora bien, notemos que, fijado n, el término del AMISE asociado a la varianza crece como h^{-d}. Esto nos dice que en altas dimensiones necesitaremos tamaños muestrales considerablemente mayores para poder controlar la varianza de nuestro estimador.

Al igual que en el caso univariante, una vez muestreada nuestra variable aleatoria, debemos escoger la matriz de ancho de banda con la que obtener la mejor estimación posible de la función de densidad. Ahora bien, a diferencia del caso unidimensional, surge un problema adicional: debemos escoger también la forma de la matriz $\boldsymbol{H}$ a emplear. Ya hemos comentado diversas opciones, como son restringir $\boldsymbol{H}$ a su diagonal o tomarla proporcional a la identidad. En caso de que queramos una matriz de ancho de banda completa, una opción es tomar $\boldsymbol{H}$ como h^2 veces la matriz de covarianzas muestral. Esto resulta equivalente, como puede verse en [27], a estandarizar multivariantemente los datos, estimar la densidad mediante el estimador definido en la ecuación (1.9), y luego aplicar la tranformación inversa para recuperar la estimación de la densidad original.

En caso de que optemos por un ancho de banda de la forma $\boldsymbol{H} = h^2\boldsymbol{I}_d$, podríamos vernos tentados a tomar h tal que minimice el AMISE calculado en (1.10), pero volvemos a encontrarnos con el problema de que este dependerá de $\int tr[Hf(\boldsymbol{x})]^2 d\boldsymbol{x}$, cantidad que desconocemos. Por ello, al igual que en el caso univariante, proponemos un método de selección del ancho de banda empleando técnicas de validación cruzada que traten de minimizar una estimación del error cuadrático. El funcional a minimizar puede definirse para una matriz $\boldsymbol{H}$ simétrica y definida positiva como:

$$\text{LSCV}(\boldsymbol{H}) = \int \widehat{f}(\boldsymbol{x};\boldsymbol{H})^2 d\boldsymbol{x} - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{-i}(\boldsymbol{x}_i;\boldsymbol{H}),$$

donde nuevamente $\widehat{f}_{-i}(\cdot;\boldsymbol{H})$ representa la estimación tipo núcleo obtenida sin emplear la observación $\boldsymbol{x}_i$. La minimización de LSCV proporcionaría una matriz $\boldsymbol{H}$ que daría lugar a un estimador cuyo MISE se encontraría presumiblemente cerca del valor mínimo de este.

Capítulo 2

Procesos puntuales en el plano euclídeo

2.1. Procesos puntuales

En múltiples ocasiones nos encontramos ante conjuntos de datos que representan las localizaciones en el plano de una serie de observaciones o eventos: epicentros de terremotos, incendios forestales, localización de una determinada especie de árbol en un bosque, etc. Esto es lo que se conoce como un patrón de puntos. A la hora de analizar conjuntos de datos de esta naturaleza, pueden surgir preguntas como: ¿se distribuyen los puntos uniformemente sobre la región de muestreo?, ¿depende la "densidad" de puntos de alguna variable explicativa?, ¿hay evidencias de clusterización?. Para dar una respuesta rigurosa a estas preguntas debemos darnos cuenta de que estas no se refieren al patrón de puntos en sí, sino al proceso que subyace en la generación de dicho patrón. Una conclusión tal como "los puntos están uniformemente distribuidos" carece de sentido. Los puntos son observaciones fijas, no aleatorias. La conclusión apropiada es que el patrón en estudio es una realización de un mecanismo que distribuye los puntos uniformemente. Necesitamos entonces de un objeto estadístico que nos permita estudiar aquellos datos que se nos presenten como patrones de puntos en el plano. Surge así el concepto de proceso puntual:

Definición 2.1. Un **proceso puntual en el plano euclídeo** es un mecanismo aleatorio que genera localizaciones distribuidas en una región $W \subset \mathbb{R}^2$, y cuya realización es un patrón de puntos.

Denotaremos los procesos puntuales por letras mayúsculas: $\mathbf{X}$; mientras que los patrones de puntos por letras minúsculas: $\mathbf{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\} \subset W \subset \mathbb{R}^2$, en ambos casos en negrita por tratarse de elementos bidimensionales. Usualmente restringiremos nuestro estudio a procesos puntuales finitos, que son aquellos tales que cualquier realización de $\mathbf{X}$ es un patrón de puntos $\mathbf{x}$ finito; y tales que el número de puntos observado en cualquier subconjunto B de la región de observación, que denotamos por $N(\mathbf{X} \cap B)$, es una variable aleatoria bien definida, usualmente

denominada medida de contar. Ahora bien, para evitar problemas en razonamientos que haremos más adelante, siempre que hablemos de un subconjunto o subregión $B \subset W$ supondremos que B es conexo y compacto en la topología usual; además, cuando digamos que $B_1, B_2 \subset W$ son disjuntos, admitiremos que a lo sumo se intersequen en un conjunto de medida nula.

Antes de profundizar en el estudio de los procesos puntuales, debemos enfatizar que no cualquier patrón de puntos $\mathbf{x}$ puede entenderse como una realización de un proceso puntual. Aquellos patrones de puntos que se correspondan con la observación de si un determinado evento ocurre o no en una serie de puntos del plano previamente fijados no pueden ser resultado de un proceso puntual. Tampoco podrá ser resultado de un proceso puntual un patrón de puntos cuyo cardinal sea conocido de antemano, o uno tal que los puntos posean cierta ordenación temporal. Esto es consecuencia de la doble aleatoriedad de los procesos puntuales: tanto el número de puntos observados como su ubicación en la región de observación son aleatorios.

2.1.1. Procesos puntuales de Poisson homogéneos

El primer tipo de procesos puntuales que vamos a estudiar son aquellos en los que existe el mayor grado de aleatoriedad posible. Es por ello que este tipo de procesos suelen denominarse de aleatoriedad espacial completa, o por sus siglas en inglés: CSR (*complete spatial randomness*). En [3] se caracterizan estos procesos en base a dos principios fundamentales:

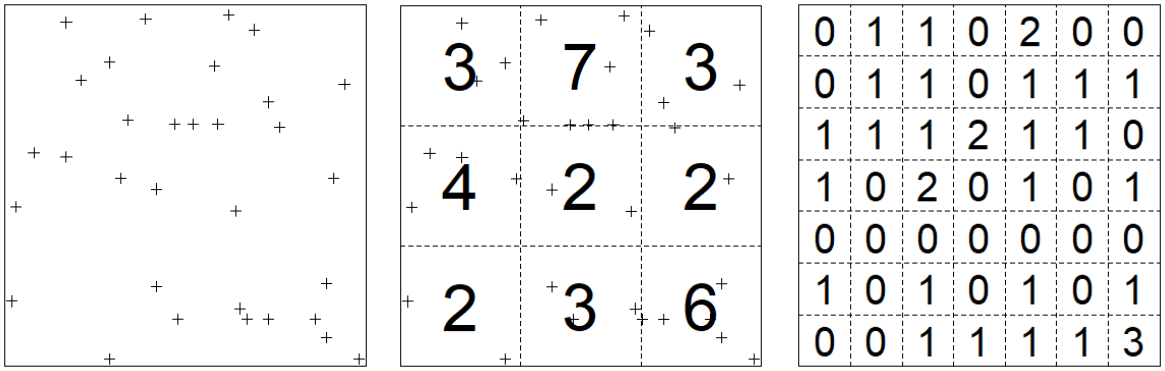
Homogeneidad: los puntos no tienen preferencia por ninguna subregión de W.

Independencia: dadas dos regiones disjuntas, las observaciones en una de ellas carecen de influencia sobre las de la otra.

La hipótesis de homogeneidad implica que el número esperado de puntos en una región $B \subset W$, ha de ser proporcional al área de B, que denotamos por $|B|$; formalmente:

$$\exists \lambda \in \mathbb{R}^+ \text{ tal que } \mathbb{E}\left[N(\mathbf{X} \cap B)\right] = \lambda |B|, \ \forall B \subset W.$$

La constante λ, denominada intensidad del proceso puntual, se corresponde con el número esperado de puntos por unidad de área. La hipótesis de independencia nos dice que para $A, B \subset W$ regiones disjuntas, las variables aleatorias $N(\mathbf{X} \cap A)$ y $N(\mathbf{X} \cap B)$ son independientes. En particular, si dividimos nuestra región de observación W en subregiones, usualmente denominadas cuadrantes, el número de puntos observados en cada uno es independiente de los demás, para cualquier tamaño de los cuadrantes. Según los cuadrantes se hacen cada vez más pequeños, la gran mayoría no contendrá ningún punto y algunos contendrán exactamente uno, como ejemplificamos en la Figura 2.1. Para evitar que pudiera haber dos puntos en un cuadrante para cualquier división de W, en [3] se propone añadir la hipótesis de orden:



(a) Patrón de puntos. (b) 9 cuadrantes. (c) 49 cuadrantes.

Figura 2.1: localización de pinos negros Japoneses en una región de muestreo de un bosque (a), junto con dos subdivisiones de la región de observación y los correspondientes conteos (b), (c). Datos extraídos del paquete [4].

Orden: hay una probabilidad despreciable de que una región suficientemente pequeña contenga más de un punto. De forma más precisa:

$$\lim_{|B|\to 0} \frac{1}{|B|}\mathbb{P}\left[N(\mathbf{X}\cap B)\geq 2\right]=0, \quad \forall B\subset W.$$

Vemos como la hipótesis de orden se materializa en la Figura 2.1; en la subfigura (b) el bajo número de subregiones da lugar a que en todas ellas se observe un número elevado de puntos; ahora bien, cuando en (c) reducimos el área de los cuadrantes vemos como en la mayoría de ellos ya no observamos ningún pino, y en casi todos los restantes observamos uno. La hipótesis de orden nos garantiza que, de seguir haciendo los cuadrantes más pequeños, no tendríamos ninguno conteniendo más de una observación. De esta forma eliminamos la posibilidad de observar dos puntos en la misma localización espacial.

Así, como las observaciones en diferentes cuadrantes son independientes y la probabilidad de que algún cuadrante contenga más de un punto (para una división suficientemente fina) es despreciable, $N(\mathbf{X}\cap B)$ es el número de cuadrantes de B que contienen un punto; es decir, $N(\mathbf{X}\cap B)$ representa el número de éxitos (encontrar un punto en un cuadrante) en un número elevado de intentos independientes con baja probabilidad de éxito. Por ello, como puede consultarse en [11], las tres hipótesis adoptadas garantizan que $N(\mathbf{X}\cap B)$ sigue una distribución de Poisson. Este es el motivo por el que a este tipo de procesos puntuales se les denomina procesos de Poisson homogéneos. El "apellido" de homogeneidad obedece a que la intensidad λ no varía a lo largo de la región de observación.

Recordemos que la distribución de Poisson queda descrita conociendo su esperanza μ. Así, bajo CSR, $N(\mathbf{X}\cap B)$ sigue una distribución de Poisson de media $\mu=\lambda|B|$. Notemos que $|B|$

depende de la región que consideremos, mientras que λ es intrínseca al proceso puntual. La importancia de este tipo de procesos puntuales reside en que, además de modelar múltiples fenómenos físicos, servirán como hipótesis nula en múltiples contrastes de interés.

2.1.2. Procesos puntuales de Poisson inhomogéneos

La primera modificación a los procesos puntuales CSR surge de prescindir de la hipótesis de homogeneidad, dando lugar a los procesos de Poisson *inhomogéneos*. Este tipo de procesos se caracterizan porque ahora la intensidad λ es una función de la posición dentro de la región de observación $\lambda(\boldsymbol{x})$. Dada una región $B \subset W$, supongamos que la dividimos en cuadrantes arbitrariamente pequeños. En cada cuadrante de posición $\boldsymbol{x}$ y área $\Delta\boldsymbol{x}$ (que denotamos de esta forma por tratarse de un área infinitesimal), el número esperado de puntos será $\lambda(\boldsymbol{x})\Delta\boldsymbol{x}$. Por tanto, el número esperado de puntos en B será la suma de estos valores en los cuadrantes de B. En el límite en el que estos cuadrantes se vuelven infinitesimales, la suma discreta converge a una integral, de tal forma que:

$$\mathbb{E}\left[N(\mathbf{X} \cap B)\right] = \int_B \lambda(\boldsymbol{y}) d\boldsymbol{y}, \tag{2.1}$$

que en el caso de que $\lambda(\boldsymbol{y})$ sea constante recuperamos los procesos homogéneos. Teniendo en cuenta que las hipótesis de independencia y orden siguen aplicando a este tipo de procesos, $N(\mathbf{X} \cap B)$ sigue una distribución de Poisson para todo $B \subset W$ como consecuencia de la reproductividad de la distribución de Poisson. La diferencia está en que ahora la media de esta distribución es $\mu = \int_B \lambda(\boldsymbol{y}) d\boldsymbol{y}$. Al igual que en los procesos homogéneos, vemos como la función de intensidad no depende de la región $B \subset W$ considerada, sino que es una propiedad intrínseca del proceso puntual.

2.2. Función de intensidad

Como hemos visto hasta el momento, la intensidad es una propiedad intrínseca de los procesos puntuales que los caracteriza a primer orden; en efecto, está relacionada con la esperanza o primer momento. Por ello, para estudiar procesos puntuales en el plano debemos primeramente profundizar en el estudio de su función de intensidad. Podemos definir la **función de intensidad** de un proceso puntual $\mathbf{X}$, en ausencia de ninguna hipótesis, como aquella función de la posición que verifica la relación dada en la ecuación (2.1) para todo $B \subset W$. A la vista de esta definición, la función de intensidad de un proceso puntual se interpreta como el número esperado de puntos por unidad de área.

Ahora bien, no debemos confundir el concepto de intensidad con el de función de densidad de probabilidad. Recordamos del Capítulo 1 que una función $f : \mathbb{R}^2 \to \mathbb{R}$ es función de densidad

de un vector aleatorio bidimensional si y solo si $f(\boldsymbol{x}) \geq 0$ para todo $\boldsymbol{x} \in \mathbb{R}^2$ y $\int_{\mathbb{R}^2} f(\boldsymbol{x})d\boldsymbol{x} = 1$. Por su parte, $\lambda : \mathbb{R}^2 \to \mathbb{R}$ es función de intensidad de un proceso puntual en el plano si y solo si $\lambda(\boldsymbol{y}) \geq 0$ para todo $\boldsymbol{y} \in \mathbb{R}^2$; es decir, no es necesario que esté normalizada para tener integral unidad. Otra diferencia notable es que la integral de la función de densidad de un vector aleatorio bivariante en una subregión $B \subset \mathbb{R}^2$ representa la probabilidad de observar dicho vector en B, mientras que la integral en esa misma región B de la función de intensidad de un proceso puntual en el plano representa el número esperado de puntos observados en B.

Indistintamente, las funciones de intensidad y densidad no son totalmente ajenas. Dada λ una función de intensidad en el plano, podemos definir su densidad asociada como: $f_\lambda(\boldsymbol{x}) = \lambda(\boldsymbol{x})/\int_{\mathbb{R}^2} \lambda(\boldsymbol{y})d\boldsymbol{y}$. Al ser esta una función de densidad de probabilidad, podemos estimarla empleando las técnicas estudiadas en el Capítulo 1, lo que permitirá construir estimadores de la función de intensidad multiplicado el estimador de la densidad asociada por un estimador de $\int_{\mathbb{R}^2} \lambda(\boldsymbol{y})d\boldsymbol{y}$, que usualmente será el número observado de puntos.

2.2.1. Estimación de la función de intensidad

Sea entonces $\mathbf{x}$ una realización de un proceso puntual de Poisson en el plano $\mathbf{X}$. Si tenemos evidencia a favor de que $\mathbf{X}$ es un proceso de Poisson homogéneo, podemos estimar su intensidad como el número observado de puntos por unidad de área, como se propone en [3]:

$$\widehat{\lambda} = \frac{N(\mathbf{x})}{|W|}, \tag{2.2}$$

que es un estimador insesgado de la intensidad del proceso puntual, ya que:

$$\mathbb{E}\left(\widehat{\lambda}\right) = \frac{\mathbb{E}\left[N(\mathbf{x})\right]}{|W|} = \frac{\mathbb{E}\left[N(\mathbf{X} \cap W)\right]}{|W|} = \frac{\lambda|W|}{W} = \lambda,$$

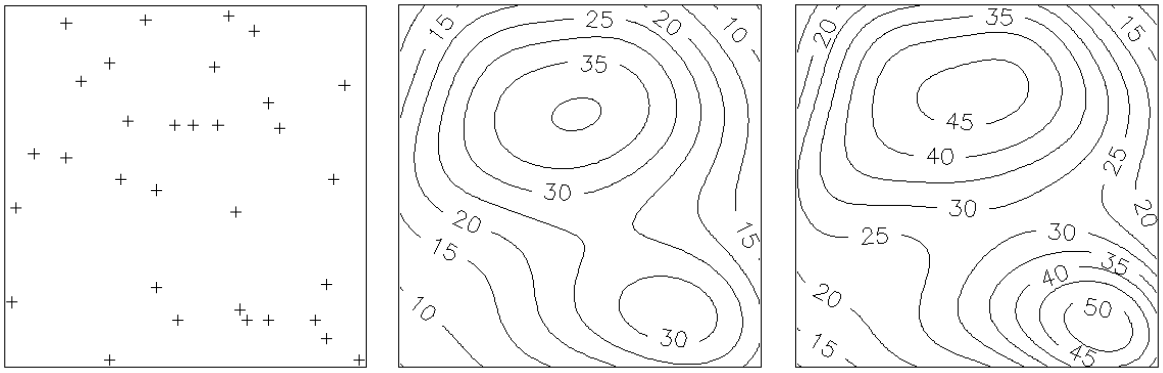
y además su varianza es:

$$\mathbb{V}\mathrm{ar}\left(\widehat{\lambda}\right) = \mathbb{V}\mathrm{ar}\left[\frac{N(\mathbf{x})}{|W|}\right] = \frac{\mathbb{V}\mathrm{ar}\left[N(\mathbf{X} \cap W)\right]}{|W|^2} = \frac{\lambda|W|}{|W|^2} = \frac{\lambda}{|W|},$$

lo que puede verse en analogía con el hecho de la varianza de la media muestral de una muestra aleatoria simple de tamaño n de una variable aleatoria con varianza σ^2 es precisamente σ^2/n. La diferencia es que ahora el papel del tamaño muestral lo juega el área de la región de observación.

En caso de no tener evidencias de homogeneidad, debemos admitir que $\mathbf{X}$ es un proceso de Poisson inhomogéneo: su intensidad varía con la posición. De no poder afirmar que nuestro proceso puntual siga algún modelo (como ocurre en la mayoría de los casos), es natural estimar $\lambda(\boldsymbol{y})$ no paramétricamente. Si $N(\mathbf{x}) = N$, entonces un primer estimador propuesto en [3] es:

$$\widehat{\lambda}_0(\boldsymbol{y}) = \frac{1}{h^2}\sum_{i=1}^{N} L\left(\frac{\boldsymbol{y} - \boldsymbol{x}_i}{h}\right). \tag{2.3}$$



(a) Patrón de puntos. (b) Estimación mediante $\widehat{\lambda}_0$. (c) Estimación mediante $\widehat{\lambda}_D$.

Figura 2.2: localización de pinos negros Japoneses en una región de muestreo de un bosque (a), junto con dos estimaciones no paramétricas de su intensidad (b), (c). En ambas se ha empleado un núcleo Gaussiano estándar y el ancho de banda se ha tomado empleando la regla de Scott [25] isótropa. Datos extraídos del paquete [4].

Del estimador tipo núcleo dado en la ecuación (2.3) debemos destacar que se ha tomado como matriz de ancho de banda $\boldsymbol{H} = h^2\boldsymbol{I}_2$, y que el núcleo L se entiende una función de densidad de probabilidad bivariante, isótropa[1], unimodal y centrada en $\mathbf{0}$. Notemos además que en el estimador $\widehat{\lambda}_0$ no dividimos entre N, a diferencia de en los estimadores estudiados en el Capítulo 1, ya que se está estimando una función de intensidad a través de su densidad asociada.

El principal problema que presenta el estimador dado en la ecuación (2.3) es que decrece fuertemente cerca de la frontera de W. Si, por ejemplo, nuestro proceso puntual es la ubicación de árboles en una región de muestreo en un bosque, como en el ejemplo de la Figura 2.1, los árboles que se encuentran cerca de los límites de W, pero fuera de esta región, no se están considerando a la hora de estimar la intensidad. Esto se conoce como un efecto frontera. Este tipo de efectos causan que $\widehat{\lambda}_0$ presente un fuerte sesgo negativo cerca de la frontera de W. Tratando de corregir estos problemas se propone la corrección de Diggle [12]:

$$\widehat{\lambda}_D(\boldsymbol{y}) = \frac{1}{h^2}\sum_{i=1}^{N}\frac{1}{e(\boldsymbol{x}_i)}L\left(\frac{\boldsymbol{y}-\boldsymbol{x}_i}{h}\right), \text{ donde } e(\boldsymbol{z}) = \int_W L(\boldsymbol{z}-\boldsymbol{w})d\boldsymbol{w}. \tag{2.4}$$

Esta corrección trata de compensar los efectos de borde al ponderar por $e(\boldsymbol{x}_i)^{-1}$, que es una medida de cómo de cerca está $\boldsymbol{x}_i$ de la frontera de W. Esto puede verse en la Figura 2.2, donde presentamos el patrón de puntos dado en la Figura 2.1 con dos estimaciones de la función de intensidad; en (b) se ha empleado el estimador dado en (2.3), mientras que en (c) se ha empleado

[1] Tal que la distribución que representa posee por matriz de covarianzas $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_2$.

la corrección de Diggle. Podemos ver en estas figuras como $\widehat{\lambda}_0$ toma valores mucho menores que $\widehat{\lambda}_D$ cerca de la frontera de W, consecuencia de los efectos frontera.

En determinados contextos en los que se estudia un proceso puntual se conocen los valores de una magnitud a lo largo de la región de observación. Estos observables se denominan covariables del proceso puntual, y acostumbran a describirse por una función $Z : W \to \mathbb{R}$. Idealmente, los valores de una covariable se conocen de forma precisa a lo largo de todo W; en la práctica, suele conocerse la imagen de Z a lo largo de una malla de W suficientemente fina.

Si estudiando un proceso puntual en presencia de una covariable tenemos evidencias de inhomogeneidad, resulta natural pensar que la covariable pueda tener un efecto sobre la intensidad de dicho proceso puntual. Así, de forma alternativa a la estimación tipo núcleo, podemos tratar de buscar una función ρ tal que:

$$\lambda(\boldsymbol{y}) = \rho\left[Z(\boldsymbol{y})\right].$$

Como generalmente carecemos de ninguna información a cerca de la forma funcional de ρ, optamos por una estimación no paramétrica como la propuesta en [3]. La idea fundamental consiste en determinar la distancia entre dos puntos del plano no empleando la distancia euclidiana, sino la función que describe la covariable Z; concretamente, empleando el área de la región comprendida entre dos curvas de nivel de esta función. Para ello, comenzamos definiendo la función de acumulación de Z como:

$$C(z) = \frac{1}{|W|}\int_W \mathbf{1}_{\{Z(\boldsymbol{w})<z\}} d\boldsymbol{w}.$$

Usualmente nos veremos en la necesidad de estimar C, al no conocer la forma precisa de Z a lo largo de todo W. Para su estimación, dividimos la región de observación en una fina malla a lo largo de la cual conocemos Z, y estimamos la función de acumulación como:

$$\widehat{C}(z) = \frac{\#\{\text{puntos } \boldsymbol{y} \text{ de la malla en los que } Z(\boldsymbol{y}) \leq z\}}{\#\{\text{puntos de la malla}\}},$$

donde $\#A$ denota el cardinal de un conjunto A. Una vez conocida (o estimada) C, estimamos ρ a través de un estimador tipo núcleo transformado:

$$\widehat{\rho}(z) = \frac{1}{|W|}\sum_{i=1}^{N} k_h[C(z) - C[Z(\boldsymbol{x}_i)]],$$

lo que proporciona la estimación de la intensidad:

$$\widehat{\lambda}(\boldsymbol{y}) = \frac{1}{|W|}\sum_{i=1}^{N} k_h[C[Z(\boldsymbol{y})] - C[Z(\boldsymbol{x}_i)]], \tag{2.5}$$

donde k en este caso ha de ser una función núcleo apropiada para la estimación no paramétrica en el caso unidimensional, y $k_h(\cdot) = h^{-1}k(\cdot/h)$. Notemos que:

$$C[Z(\boldsymbol{y})] - C[Z(\boldsymbol{x}_i)] \propto \int_W \mathbf{1}_{\{Z(\boldsymbol{w}) \text{ entre } Z(\boldsymbol{y}) \text{ y } Z(\boldsymbol{x}_i)\}} d\boldsymbol{w},$$

que es el área encerrada entre las curvas de nivel de Z que pasan por $\boldsymbol{y}$ y $\boldsymbol{x}_i$. De esta forma, vemos como el estimador de la función de intensidad dado en la ecuación (2.5) no es más que un estimador tipo núcleo dónde la distancia entre dos puntos se mide a través del área de la región delimitada por las curvas de nivel de Z que pasan por dichos puntos, normalizada por el área de la región de observación.

2.2.2. Marcas

Los puntos de una realización de un proceso puntual pueden llevar asociados múltiples atributos. Cualquier información asociada a cada punto de dicha realización se denomina marca. Esto nos lleva a hablar de patrones de puntos marcados: $\mathbf{y} = \{(\boldsymbol{x}_1, \mathfrak{m}_1), \ldots, (\boldsymbol{x}_N, \mathfrak{m}_N)\}$, donde $\mathbf{x} = \{\boldsymbol{x}_i\}_{i=1}^N$ es un patrón de puntos en el plano y $\{\mathfrak{m}_i\}_{i=1}^N$ es el conjunto de marcas de $\mathbf{x}$. Esto lleva a definir un **proceso puntual marcado** como aquel mecanismo aleatorio cuya realización es un patrón de puntos marcado. Las marcas de un patrón de puntos pueden ser de múltiples tipos: desde valores categóricos hasta observaciones multivariantes.

La gran diferencia entre una marca y una covariable de un proceso puntual es que las marcas son intrínsecas a este y sus valores se obtienen como parte la realización del proceso puntual, mientras que las covariables son extrínsecas; es decir, sus valores a lo largo de la región de observación no son aleatorios y no dependen de la realización del proceso puntual.

Capítulo 3

Procesos puntuales en grafos lineales

En el capítulo anterior hemos estudiado los procesos puntuales en el plano euclídeo. Las realizaciones de estos procesos puntuales eran patrones de puntos en una región de observación $W \subset \mathbb{R}^2$, que considerábamos conexa y compacta en la topología usual. Ahora bien, en determinadas ocasiones nos encontramos ante patrones de puntos que no se disponen a lo largo de un subconjunto 2-dimensional del plano, sino a lo largo de un conjunto 1-dimensional embebido en este: accidentes de tráfico en una red de carreteras, defectos en una instalación eléctrica, escapes a lo largo de un sistema de cañerías, avistamientos de aves a lo largo de la línea de costa... Este tipo de eventos están constreñidos a ocurrir en espacios de una dimensión, lo que supone una diferencia sustancial con los patrones de puntos que hemos estudiado hasta el momento. Debemos entonces definir con precisión los espacios en los que vamos a estudiar este tipo de patrones de puntos. Introducimos para ello el concepto de grafo lineal [8]:

Definición 3.1. Un **grafo lineal** $\mathcal{G}$ es una dupla $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ donde $\mathcal{V} = \{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{n_v}\} \subset \mathbb{R}^2$ es el conjunto de nodos o vértices del grafo, y $\mathcal{A} = \{l_1, \dots, l_{n_l}\}$ es el conjunto de aristas del grafo. Estas aristas son segmentos que empiezan y acaban en nodos del grafo:

$$\forall i \in \{1, \dots, n_l\} \; \exists \; \boldsymbol{v}_{i1}, \boldsymbol{v}_{i2} \in \mathcal{V}, \text{ tales que } \boldsymbol{v}_{i1} \neq \boldsymbol{v}_{i2} \text{ y } l_i = \{(1-t)\boldsymbol{v}_{i1} + t\boldsymbol{v}_{i2} \in \mathbb{R}^2 \colon t \in [0,1]\}.$$

Asumiremos que dos aristas únicamente pueden intersecarse en nodos del grafo. Definimos además el subconjunto del plano representado por este grafo como la unión de sus aristas: $\mathcal{L}_\mathcal{G} = \bigcup_{i=1}^{n_l} l_i \subset \mathbb{R}^2$, que supondremos conexo.

Debemos comentar que existe cierta dualidad en la representación de un grafo lineal. Primeramente, es frecuente abusar del lenguaje y denominar grafo lineal, o simplemente grafo, tanto a la dupla $\mathcal{G}$ como al subconjunto del plano $\mathcal{L}_\mathcal{G}$. Intuitivamente, podría parecer que la descripción apropiada de un grafo sería $\mathcal{L}_\mathcal{G}$, pues es lo que representamos gráficamente y sobre la que se puede

efectuar cálculo diferencial e integral. Indistintamente, la representación en dupla nodos-aristas de un grafo lineal es apropiada tanto desde un punto de vista algebraico como computacional.

Empleando grafos lineales pueden representarse una amplia variedad de espacios. Retomando los ejemplos previos: en un mapa las calles o carreteras se representan por aristas y los cruces por nodos; en una instalación eléctrica los cables se identifican con las aristas y las clemas con los nodos; el litoral de una determinada región puede aproximarse por segmentos consecutivos, unidos pos nodos. Esta versatilidad hace de los grafos lineales la herramienta apropiada para describir los espacios 1-dimensionales contenidos en el plano sobre los que se observan los patrones de puntos que venimos comentando.

Ahora bien, los grafos lineales presentan diferencias sustanciales con los subconjuntos del plano empleados en el capítulo anterior. Primeramente, no todos los puntos de $\mathcal{L}_{\mathcal{G}}$ presentan las mismas propiedades locales. Por ejemplo, el número de aristas conectadas en cada nodo de $\mathcal{G}$ puede depender del nodo en cuestión. La segunda gran diferencia es la métrica empleada en el grafo. En el plano hemos empleado la distancia euclídea, es decir, la distancia entre dos puntos es la longitud del segmento que los une. El problema surge en que la gran mayoría de los grafos no son convexos, careciendo de sentido emplear la métrica euclídea para medir distancias sobre ellos. La forma apropiada de medir distancias en un grafo lineal es empleando la métrica del camino más corto. Para definir este concepto de forma precisa necesitamos el concepto de camino [8]:

Definición 3.2. Sea $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ un grafo lineal. Dados $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{V}$, un **camino** en $\mathcal{G}$ entre ellos es un conjunto de nodos $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s\} \subset \mathcal{V}$ tal que $\boldsymbol{v}_1 = \boldsymbol{v}, \boldsymbol{v}_s = \boldsymbol{w}$ y para todo $i \in \{1, \ldots, s-1\}$ $\exists\, l_i \in \mathcal{A}$ conectando $\boldsymbol{v}_i$ y $\boldsymbol{v}_{i+1}$. Definimos la **longitud** de este camino como $\sum_{i=1}^{s-1} \|\boldsymbol{v}_{i+1} - \boldsymbol{v}_i\|$. Un **ciclo** en $\mathcal{G}$ será un camino en $\mathcal{G}$ que empieza y termina en el mismo nodo.

Al igual que con el concepto de grafo lineal, existe una dualidad en la forma de representar un camino en un grafo: bien como un conjunto ordenado de nodos o como el conjunto de puntos del plano que forman una poligonal entre los nodos inicial y final. Podemos definir entonces la métrica del camino más corto en un grafo lineal como:

Definición 3.3. Sea $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ un grafo lineal. Dados $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{V}$, definimos la **distancia del camino más corto** (o simplemente la distancia) entre $\boldsymbol{v}$ y $\boldsymbol{w}$ como la menor de las longitudes de los caminos en $\mathcal{G}$ entre $\boldsymbol{v}$ y $\boldsymbol{w}$. La denotaremos por $d_{\mathcal{G}}(\boldsymbol{v}, \boldsymbol{w})$.

Notemos que $d_{\mathcal{G}}$ está bien definida, ya que el conjunto de caminos entre dos nodos es finito al serlo $\mathcal{V}$. El cálculo del camino más corto entre dos nodos de un grafo lineal es un problema de programación. El algoritmo más extendido para el cálculo de la distancia del camino más corto entre dos nodos de un grafo lineal es el denominado algoritmo de Dijkstra, cuya formulación puede consultarse en [8]. Una demostración de que este algoritmo en efecto encuentra el camino

más corto entre dos nodos de un grafo lineal en un tiempo polinomial puede encontrarse en [9]. Ahora bien, en múltiples ocasiones vamos a necesitar calcular la distancia (más corta) entre dos puntos $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{L}_{\mathcal{G}}$ que no sean necesariamente nodos de $\mathcal{G}$. Para ello, lo que haremos será añadir $\boldsymbol{p}$ y $\boldsymbol{q}$ al conjunto de nodos de $\mathcal{G}$, modificando también su conjunto de aristas, obteniendo un nuevo grafo $\widetilde{\mathcal{G}} = \left(\widetilde{\mathcal{V}}, \widetilde{\mathcal{A}}\right)$ tal que $\widetilde{\mathcal{V}} = \mathcal{V} \bigcup \{\boldsymbol{p}, \boldsymbol{q}\}$ y $\mathcal{L}_{\mathcal{G}} = \mathcal{L}_{\widetilde{\mathcal{G}}}$. Hecho esto, calculamos la distancia entre $\boldsymbol{p}$ y $\boldsymbol{q}$ como $d_{\widetilde{\mathcal{G}}}(\boldsymbol{p}, \boldsymbol{q})$ empleando el algoritmo de Dijkstra. El procedimiento para añadir un nuevo nodo a un grafo lineal puede verse en detalle en [22].

3.1. Procesos puntuales en grafos lineales

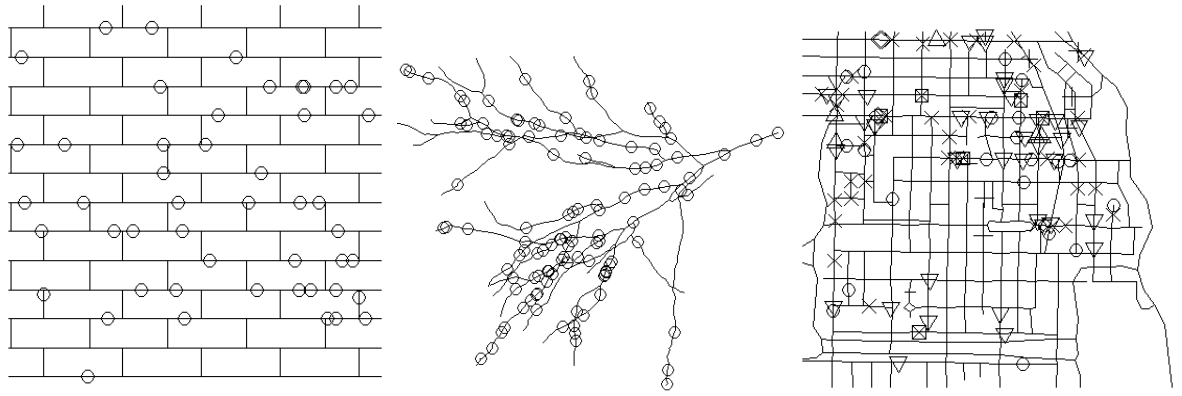
Al examinar patrones de puntos en un grafo lineal, surgen las mismas preguntas que cuando el patrón de puntos se disponía en una región del plano: ¿se distribuyen los puntos uniformemente en el grafo?, ¿depende la "densidad" de puntos de alguna variable explicativa?, ¿depende la "densidad" de puntos cerca de un nodo del número de aristas que se conectan en él?. Ya sabemos que estas preguntas no aluden al patrón de puntos en sí, sino al proceso responsable de su generación. Debemos por ello introducir el objeto estadístico que nos permita estudiar patrones de puntos en un grafo. Estos mecanismos serán los procesos puntuales en grafos lineales:

Definición 3.4. Un **proceso puntual en un grafo lineal** $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ es un mecanismo aleatorio que genera localizaciones distribuidas en $\mathcal{L}_{\mathcal{G}} \subset \mathbb{R}^2$, y cuya realización es un patrón de puntos.

Denotaremos los procesos puntuales en un grafo $\mathcal{G}$ por letras mayúsculas $\mathbf{P}$; mientras que denotaremos los patrones de puntos en este grafo por $\mathbf{p} = \{\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N\} \subset \mathcal{L}_{\mathcal{G}} \subset \mathbb{R}^2$. Empleamos por defecto la letra $\boldsymbol{p}$ para denotar los elementos de un patrón de puntos en un grafo $\mathbf{p}$ para enfatizar que, aunque sigan siendo puntos del plano, están constreñidos a encontrarse en un subconjunto 1-dimensional del plano como es $\mathcal{L}_{\mathcal{G}}$. Es por este mismo motivo que los procesos puntuales en un grafo los denotaremos por defecto por $\mathbf{P}$ en vez de por $\mathbf{X}$.

En la Figura 3.1 encontramos ejemplos de patrones de puntos en grafos. En (a) se muestra la ubicación de telas de araña a lo largo de las juntas de una pared de ladrillo. El material del que están compuestos los ladrillos y el mortero hacen que las telas de araña solo puedan encontrarse en las juntas, por lo que este conjunto de datos ha de entenderse como un patrón de puntos en un grafo lineal. En (b) encontramos espinas dendríticas en un fragmento de la red dendrítica de una neurona. Este es un ejemplo donde se emplea un grafo lineal para modelar un espacio unidimensional, similar al ejemplo de avistamiento de aves en el litoral.

Al igual que en los procesos puntuales en el plano, trabajaremos con procesos puntuales en



(a) (b) (c)

Figura 3.1: ejemplos de patrones puntuales en grafos lineales extraídos del paquete [4].

grafos lineales finitos, que son aquellos tales que cualquier realización $\mathbf{p}$ de $\mathbf{P}$ es un patrón de puntos finito en el grafo; y tales que el número de puntos observado en cualquier subconjunto $\mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}$, lo que denotamos por $N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}})$, es una variable aleatoria bien definida. Para que los razonamientos que haremos en secciones posteriores sean correctos, todas las subregiones[1] $\mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}$ que consideremos serán conexas y compactas como subconjuntos de $\mathbb{R}^2$, de tal forma que pueden identificarse con un grafo lineal $\mathcal{G}'$ tal que $\mathcal{L}'_{\mathcal{G}} = \mathcal{L}_{\mathcal{G}'}$. Además, diremos que dos subregiones de $\mathcal{L}_{\mathcal{G}}$ son disjuntas si a lo sumo se intersecan en un conjunto de medida nula.

Al igual que cuando estudiamos procesos puntuales en el plano, debemos empezar definiendo los procesos puntuales en grafos lineales en los que existe el mayor grado de aleatoriedad. Al igual que en los procesos puntuales en el plano, estos procesos se denominan de aleatoriedad espacial completa (CSR), y se construyen en base a las hipótesis siguientes:

Homogeneidad: los puntos no tienen preferencia por ninguna subregión de $\mathcal{L}_{\mathcal{G}}$.

Independencia: dadas dos regiones disjuntas, las observaciones en una de ellas carecen de influencia sobre las de la otra.

Orden: hay una probabilidad despreciable de que una región suficientemente pequeña contenga más de un punto. De forma más precisa:

$$\lim_{|\mathcal{L}'_{\mathcal{G}}| \to 0} \frac{1}{|\mathcal{L}'_{\mathcal{G}}|} \mathbb{P}\left[N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}}) \geq 2\right] = 0, \quad \forall \mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}},$$

donde $|\mathcal{L}'_{\mathcal{G}}|$ denota la longitud total del grafo $\mathcal{L}'_{\mathcal{G}}$ (la suma de las longitudes de sus aristas). Al igual que en procesos puntuales en el plano, la hipótesis de homogeneidad implica que el número esperado de puntos observados en una subregión $\mathcal{L}_{\mathcal{G}}$ ha de ser proporcional a su longitud:

$$\exists \lambda \in \mathbb{R}^+ \text{ tal que } \mathbb{E}\left[N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}})\right] = \lambda |\mathcal{L}'_{\mathcal{G}}|, \ \forall \mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}.$$

[1] Que inicialmente no tienen que ser subgrafos en el sentido estricto de la teoría de grafos lineales.

Como ya sabemos, λ recibe el nombre de intensidad del proceso puntual. Por su parte, la hipótesis de independencia garantiza que dadas dos subregiones $\mathcal{L}'_{\mathcal{G}}, \mathcal{L}''_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}$ disjuntas, las variables aleatorias $N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}})$ y $N(\mathbf{P} \cap \mathcal{L}''_{\mathcal{G}})$ son independientes. Un razonamiento análogo al realizado en el capítulo anterior permite demostrar que si $\mathbf{P}$ es un proceso puntual CSR en un grafo $\mathcal{G}$ con intensidad λ, entonces la variable aleatoria $N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}})$ sigue una distribución de Poisson de parámetro $\mu = \lambda|\mathcal{L}'_{\mathcal{G}}|$, para todo $\mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}$. Por ello, a los procesos puntuales CSR en un grafo se les denomina también procesos puntales de Poisson homogéneos en un grafo lineal.

La primera desviación de los procesos puntuales CSR en un grafo la encontramos si prescindimos de la hipótesis de homogeneidad. Surgen así los procesos puntuales de Poisson *inhomogéneos* en un grafo. Estos se caracterizan porque la intensidad es ahora una función de la posición en el grafo $\lambda(\boldsymbol{p})$. Dada una subregión $\mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}$, podemos realizar un razonamiento análogo al de la Sección 2.1.2, dividiéndola en regiones de longitud arbitrariamente pequeña, lo que permite concluir que si $\mathbf{P}$ es un proceso puntual de Poisson inhomogéneo en un grafo $\mathcal{G}$, entonces:

$$\mathbb{E}\left[N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}})\right] = \int_{\mathcal{L}'_{\mathcal{G}}} \lambda(\boldsymbol{p}) d\boldsymbol{p}, \quad \forall \mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}, \tag{3.1}$$

donde por $d\boldsymbol{p}$ denotamos el diferencial de línea en $\mathcal{L}'_{\mathcal{G}}$. Recordando que las hipótesis de independencia y orden siguen vigentes en este tipo de procesos puntuales, y la reproductividad de la distribución de Poisson, tenemos que $N(\mathbf{P} \cap \mathcal{L}'_{\mathcal{G}})$ sigue una distribución de Poisson de parámetro $\mu = \int_{\mathcal{L}'_{\mathcal{G}}} \lambda(\boldsymbol{p}) d\boldsymbol{p}$, de acuerdo con la ecuación (3.1), para todo $\mathcal{L}'_{\mathcal{G}} \subset \mathcal{L}_{\mathcal{G}}$.

Al igual que ocurría en los procesos puntuales en el plano, la realización de un proceso puntual en un grafo lineal puede no consistir únicamente en las ubicaciones espaciales de los puntos en el grafo, sino de más información acompañando a estas. Como ya sabemos, esta información adicional vinculada a los eventos se conoce como una marca del proceso puntual. Un ejemplo de una realización de un proceso puntual marcado en un grafo lineal lo encontramos en la Figura 3.1, (c), donde presentamos eventos criminales en el mapa de calles de un barrio de Chicago. El tipo de crimen (asalto, robo, allanamiento, robo de vehículo...) constituye una marca al proceso puntual. Debemos recordar que las marcas son intrínsecas al proceso puntual, y por ello sus valores se obtienen como parte de la realización de este. Por su parte, aquellas magnitudes cuyos valores se conocen a lo largo del grafo, pero que son extrínsecas al proceso puntual, se conocen como covariables del proceso puntual en el grafo. Al igual que ocurría en el plano, una covariable a un proceso puntual $\mathbf{P}$ en un grafo lineal $\mathcal{G}$ se representa mediante una función $Z : \mathcal{L}_{\mathcal{G}} \to \mathbb{R}$.

3.2. Función de intensidad

Como ya sabemos, todo proceso puntual queda caracterizado en términos de primer orden por su función de intensidad, y los procesos puntuales en grafos lineales no son una excepción. Dado

$\mathbf{P}$ un proceso puntual en un grafo lineal $\mathcal{G}$, se define su función de intensidad $\lambda : \mathcal{L}_{\mathcal{G}} \to [0\,,\infty)$ como una función de la posición en el grafo verificando la ecuación (3.1). A la vista de esta ecuación, podemos interpretar la función de intensidad de un proceso puntual en un grafo lineal como el número esperado de puntos por unidad de longitud.

Supongamos que poseemos un patrón de puntos $\mathbf{p}$ en un grafo lineal $\mathcal{G}$, y que deseamos emplear la información que este nos brinda para estimar la intensidad del proceso puntual $\mathbf{P}$ en $\mathcal{G}$ que lo genera. Si tenemos evidencias de que $\mathbf{P}$ es un proceso homogéneo, su intensidad no varía a lo largo de $\mathcal{L}_{\mathcal{G}}$. Recordando que la intensidad de un proceso puntual en un grafo lineal se interpreta como el número esperado de puntos por unidad de longitud, en [3] se propone el estimador:

$$\widehat{\lambda} = \frac{N(\mathbf{p})}{|\mathcal{L}_{\mathcal{G}}|}, \tag{3.2}$$

que, al igual que ocurría en el plano, es un estimador insesgado de la intensidad de $\mathbf{P}$.

En el caso de que no tengamos evidencias significativas a cerca de la homogeneidad de $\mathbf{P}$, debemos asumir que su intensidad varía a lo largo de $\mathcal{L}_{\mathcal{G}}$. Al carecer de ninguna información a cerca de la forma funcional de λ, la estrategia usual empleada será la estimación no paramétrica. Una primera idea podría ser no tener en cuenta el hecho de que nuestro patrón de puntos solo puede observarse sobre $\mathcal{L}_{\mathcal{G}}$. Al estar este conjunto embebido en el plano euclídeo, podríamos tomar las coordenadas espaciales de los puntos de $\mathbf{p}$, entender este patrón de puntos como un patrón de puntos en una región $W \subset \mathbb{R}^2$ que contenga a $\mathcal{L}_{\mathcal{G}}$, y emplear las herramientas de estimación discutidas en el Capítulo 2. Este tipo de técnicas lleva a resultados incorrectos: si por ejemplo $\mathbf{P}$ posee intensidad constante, el procedimiento de estimación que hemos descrito tenderá a sobreestimar la intensidad en aquellas zonas donde el grafo sea más denso, y a subestimarla donde la concentración de aristas sea menor.

Para ejemplificar este problema, podemos comparar las estimaciones de la intensidad (asumiendo homogeneidad) empleando el patrón de puntos dado en la Figura 3.1, (b). Empleando el estimador dado en la ecuación (3.2) obtenemos $\widehat{\lambda} = 0.06\ \mu\text{m}^{-1}$; mientras que si obviamos la estructura de grafo lineal sobre la que se encuentran nuestros puntos y estimamos la intensidad empleando el estimador dado en la ecuación (2.2) concluímos que $\widehat{\lambda} = 0.00255\ \mu\text{m}^{-2}$, que vemos es una subestimación considerable.

Este problema no afecta solo a la función de intensidad. Si empleamos únicamente las coordenadas espaciales de un patrón de puntos en un grafo para estudiar la correlación mediante las distancias entre pares de puntos, comúnmente se concluirá que existe un fenómeno de clusterización a cortas distancias (debido a que los puntos se disponen sobre las aristas del grafo) y homogeneidad a largas distancias (debido a la distancia entre aristas del grafo). Estos fenómenos se han observado en datos reales, como puede verse [28] en más detalle.

Otra estrategia más elaborada consiste en dar cabida a la restricción espacial impuesta por el grafo sustituyendo la distancia euclídea por la distancia del camino más corto en los estimadores tipo núcleo de la función de intensidad. Para ello, podríamos plantearnos sustituir las funciones núcleo $L(\boldsymbol{y}-\boldsymbol{x}_i)$ empleadas en las estimaciones dadas en las ecuaciones (2.3) y (2.4), por una función núcleo apropiada para la estimación en una dimensión con argumento la distancia del camino más corto en el grafo: $k[d_{\mathcal{G}}(\boldsymbol{p},\boldsymbol{p}_i)]$, con $\boldsymbol{p}\in\mathcal{L}_{\mathcal{G}}$ y $\boldsymbol{p}_i\in\mathbf{p}$. Indistintamente, esta tampoco es una estrategia fructuosa. Como puede consultarse en [22], la función núcleo inducida en el grafo $k[d_{\mathcal{G}}(\cdot,\boldsymbol{p}_i)]$ no es siquiera una función de densidad en $\mathcal{L}_{\mathcal{G}}$, al no estar normalizada. La razón fundamental de ello es que, a diferencia de lo que ocurre en $\mathbb{R}$, en un grafo el número de puntos a una distancia de uno dado no tiene que ser siempre dos, debido a la no homogeneidad de los grafos que ya hemos comentado.

3.2.1. Estimadores tipo núcleo equitativos

Una posible solución al problema de la función núcleo es emplear estimadores de la intensidad que no solamente empleen la distancia del camino más corto, sino que tengan en cuenta la estructura de $\mathcal{L}_{\mathcal{G}}$ como subconjunto 1-dimensional del plano. Dado un punto $\boldsymbol{p}\in\mathcal{L}_{\mathcal{G}}$ y un ancho de banda h definimos $\mathcal{L}_{\boldsymbol{p}}=\{\boldsymbol{q}\in\mathcal{L}_{\mathcal{G}}\colon d_{\mathcal{G}}(\boldsymbol{p},\boldsymbol{q})\leq h\}$. En [22] se propone considerar como estimador de la función de intensidad:

$$\widehat{\lambda}(\boldsymbol{p})=\sum_{i=1}^{N}G_{\boldsymbol{p}_i,h}(\boldsymbol{p}), \tag{3.3}$$

donde, dado $\boldsymbol{q}\in\mathcal{L}_{\mathcal{G}}$, se define la función núcleo en $\mathcal{L}_{\mathcal{G}}$ centrada en $\boldsymbol{q}$ para el ancho de banda h como una función $G_{\boldsymbol{q},h}:\mathcal{L}_{\mathcal{G}}\to\mathbb{R}$ verificando que:

$$G_{\boldsymbol{q},h}(\boldsymbol{p})\geq 0\ \forall\boldsymbol{p}\in\mathcal{L}_{\mathcal{G}},\ G_{\boldsymbol{q},h}(\boldsymbol{p})=0\ \forall\boldsymbol{p}\notin\mathcal{L}_{\boldsymbol{q}}\ \text{ y }\int_{\mathcal{L}_{\boldsymbol{q}}}G_{\boldsymbol{q},h}(\boldsymbol{p})d\boldsymbol{p}=1. \tag{3.4}$$

Para que el estimador propuesto en la ecuación (3.3) sea un estimador insesgado de la función de intensidad, en [22] se propone construir las funciones núcleo $G_{\boldsymbol{q},h}$ en términos de la denominada **función núcleo básica** w. Estas son funciones de la distancia más corta desde $\boldsymbol{q}$ a través de $\mathcal{L}_{\mathcal{G}}$, y sobre las que asumiremos las siguientes hipótesis:

(R_i) $\int_{\mathcal{L}_{\mathcal{G}}}w[d_{\mathcal{G}}(\boldsymbol{q},\boldsymbol{p})]d\boldsymbol{q}=1$ para todo $\boldsymbol{p}\in\mathcal{L}_{\mathcal{G}}$.

(R_{ii}) w es una función no negativa, continua y no creciente respecto de la distancia del camino más corto entre dos puntos de $\mathcal{L}_{\mathcal{G}}$.

(R_{iii}) Fijado el ancho de banda h, $w[d_{\mathcal{G}}(\boldsymbol{q},\boldsymbol{p})]=0$ siempre que $d_{\mathcal{G}}(\boldsymbol{q},\boldsymbol{p})\geq h$.

Notemos que w no tiene que ser necesariamente una función de densidad en $\mathbb{R}$. De hecho, la hipótesis (R_i) soluciona el problema que aparece al emplear la distancia del camino más corto como argumento de funciones de densidad unidimensionales. Los distintos tipos de estimadores de la función de intensidad de un proceso puntual en un grafo propuestos en [22] surgen de cómo se construyen las funciones núcleo $\{G_{\boldsymbol{p}_i,h}\}_{i=1}^{N}$ a partir de la función básica w elegida. Ahora bien, antes de entrar en la descripción de estos, debemos introducir el concepto de grado de un nodo:

Definición 3.5. Dado $\boldsymbol{v} \in \mathcal{V}$ un nodo de un grafo $\mathcal{G}$, definimos el **grado** de $\boldsymbol{v}$, y lo denotamos por $deg(\boldsymbol{v})$, como el número de aristas de $\mathcal{G}$ que lo contienen. Por otra parte, dado $\boldsymbol{p} \in \mathcal{L}_{\mathcal{G}} \backslash \mathcal{V}$, diremos que posee grado dos, y lo denotaremos igualmente como $deg(\boldsymbol{p}) = 2$. Esto obedece a que este sería el grado que tendría $\boldsymbol{p}$ de introducirlo en $\mathcal{G}$ como nodo.

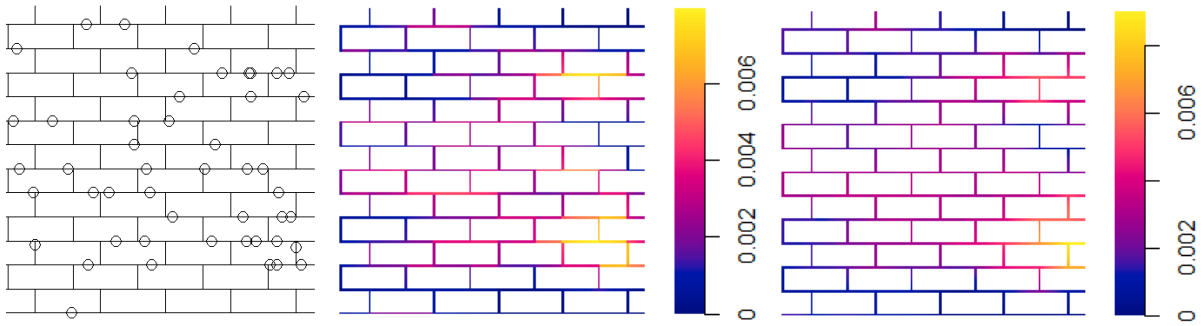
La primera clase de funciones tipo núcleo propuesta en [22] son los núcleos denominados equitativos discontinuos, o por sus siglas en inglés ESDK (*Equal-Split Discontinuous Kernel*). Para que este tipo de funciones estén bien definidas es necesario que no existan en el grafo ciclos de longitud menor que $2h$. La idea para construir estas funciones es partir de un punto del grafo y recorrer una distancia h, de tal forma que cuando llegamos a un nodo dividimos la densidad (notemos que $G_{\boldsymbol{q},h}$ es una densidad en el grafo de acuerdo a la ecuación (3.4)) equitativamente entre las distintas aristas por las que podríamos seguir recorriendo el grafo. El término discontinuo se debe a que estas funciones tipo núcleo no estarán definidas en los nodos del grafo, en los cuales presentarán una discontinuidad de salto finito. Definimos de forma precisa los ESDK como:

Definición 3.6. Dado $\mathcal{G} = (\mathcal{V}, \mathcal{A})$ un grafo lineal, tomamos $\boldsymbol{q} \in \mathcal{L}_{\mathcal{G}}$ y $\boldsymbol{p} \in \mathcal{L}_{\boldsymbol{q}} \backslash \mathcal{V}$. Sean $\boldsymbol{v}_1 \ldots, \boldsymbol{v}_m$ los nodos de $\mathcal{G}$ que encontramos al recorrer el camino más corto desde $\boldsymbol{q}$ hasta $\boldsymbol{p}$, numerados según los encontramos[2]. Supondremos además que $\boldsymbol{v}_1 \neq \boldsymbol{q}$ (ya que $\boldsymbol{q}$ podría ser un nodo). Se define entonces el **núcleo equitativo discontinuo** centrado en $\boldsymbol{q}$ con ancho de banda h como:

$$G_{\boldsymbol{q},h}(\boldsymbol{p}) = \frac{2w[d_{\mathcal{G}}(\boldsymbol{p}, \boldsymbol{q})]}{deg(\boldsymbol{q}) \prod_{i=1}^{m} [deg(\boldsymbol{v}_i) - 1]}.$$

Un estimador de la función de intensidad tal como el dado en la ecuación (3.3) empleando funciones tipo núcleo dadas en la Definición 3.6 se denomina un estimador tipo núcleo equitativo discontinuo de la intensidad de un proceso puntual en un grafo. Como puede consultarse en [22], este es un estimador insesgado de la intensidad de un proceso puntual CSR en un grafo lineal.
Un ejemplo de estimación de la función de intensidad empleando este tipo de funciones núcleo lo encontramos en la Figura 3.2. En (a) volvemos a presentar el patrón de puntos en un grafo de las telas de araña en las juntas del muro de ladrillo, y en (b) la estimación tipo núcleo de la función de intensidad empleando núcleos equitativos discontinuos. Vemos como este estimador es discontinuo en casi todos los nodos del grafo. Notemos como también es discontinuo en puntos intermedios de aristas que se encuentran a una distancia h de puntos de $\mathbf{p}$.

[2]De toda esta información se dispone tras calcular este camino más corto empleando el algoritmo de Dijkstra.



(a) Patrón de puntos. (b) Estimación equitativa discontinua. (c) Estimación equitativa continua.

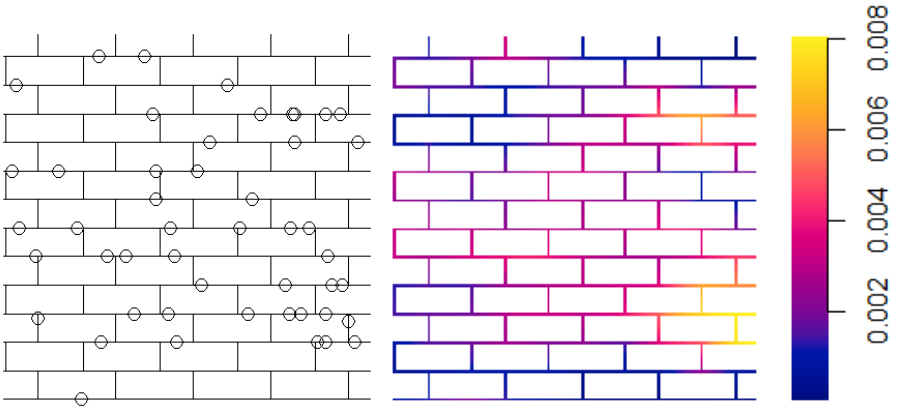
Figura 3.2: localización de telas de araña en las juntas de una pared de ladrillo (a), junto con dos estimaciones no paramétricas de su intensidad (b), (c), empleando como función núcleo básica el núcleo de Epanechnikov y con el ancho de banda dado por la regla de Scott modificada [23]. Datos extraídos de [4].

La principal desventaja que presenta la familia anterior de estimadores es, evidentemente, su discontinuidad. Uno desearía que la estimación de la función de intensidad fuese continua, y podría además tener interés en conocer una estimación de la intensidad en los nodos del grafo. Ateniendo a estos motivos, en [22] se propone una nueva clase de funciones núcleo, denominados núcleos equitativos continuos, o ESCK (*Equal-Slit Continuous Kernel*). Estos se siguen construyendo a partir de una función núcleo básica w, pero esta construcción es conceptualmente distinta. En el caso discontinuo recorremos el camino más corto entre dos puntos del grafo, y cuando llegamos a un nodo repartimos la densidad de probabilidad equitativamente entre las aristas por las que podemos seguir, lo que genera una discontinuidad. En el caso continuo recorremos el camino más corto entre dos puntos del grafo, y cuando llegamos a un nodo y "vemos" el número de aristas por el que podríamos seguir volvemos hacia atrás por el camino más corto que veníamos recorriendo y modificamos la función núcleo a partir de cierto punto, haciendo que decrezca más rápido, de tal forma que cuando llegamos al nodo en cuestión y repartimos equitativamente la densidad entre las aristas por la que podemos seguir, esto se hace con continuidad.

A diferencia de los ESDK, no existe una forma cerrada para los núcleos equitativos continuos en un grafo lineal. En [21] se propone un algoritmo iterativo para el cálculo de estas funciones. Un ejemplo de la estimación de la función de intensidad de un proceso puntual en un grafo lineal empleando ESCK's puede verse en la Figura 3.2, (c). Esta estimación es continua en los nodos del grafo, a diferencia de lo que ocurría en la Figura 3.2 (b) empleando núcleos discontinuos. Ahora bien, vemos que con el mismo ancho de banda y la misma función w, las diferencias entre estos dos estimadores son pequeñas. Además, la estimación empleando núcleos continuos es mucho más costosa computacionalmente, lo que se debe a la necesidad de recorrer los caminos más cortos entre puntos del grafo en ambos sentidos, propagación que no es necesaria en el caso discontinuo.

3.2.2. Estimación basada en la ecuación del calor

El principal problema que presentan los dos estimadores propuestos en la Sección 3.2.1 es su elevado coste computacional, incluso en el caso discontinuo. Una alternativa a este tipo de estimadores surge de darle una interpretación física a la estimación tipo núcleo. Como se describe en [2], dado un grafo $\mathcal{G}$ podemos pensar $\mathcal{L}_\mathcal{G}$ como una estructura compuesta de barras metálicas soldadas en los nodos. Tomado $\mathbf{P}$ un proceso puntual en $\mathcal{G}$ y $\mathbf{p}$ una realización de este, podemos entender los puntos observados $\mathbf{p} = \{\boldsymbol{p}_i\}_{i=1}^N$ como fuentes de calor en la estructura metálica proporcionando la misma cantidad de calor por unidad de tiempo. Pasado un tiempo $t = h^2$, la distribución de calor en el grafo es precisamente la estimación de la función de intensidad asociada a un ancho de banda h.



(a) Patrón de puntos. (b) Estimación basada en la ecuación del calor.

Figura 3.3: localización de telas de araña en las juntas de una pared de ladrillo (a), junto con la estimación de la función de intensidad basada en la ecuación del calor (b). El ancho de banda se ha tomado empleando modificación a la regla de Scott propuesta en [23]. Datos extraídos del paquete [4].

Teniendo en cuenta que el flujo de calor en el grafo vendrá dado por la ecuación del calor reducida:

$$\partial_t \lambda = \frac{1}{2}\partial_x^2 \lambda, \tag{3.5}$$

que se verifica en todos los puntos de $\mathcal{L}_\mathcal{G}$ a excepción de los nodos (en los que $\partial_x^2\lambda$ podría no estar bien definida), donde se verifica la condición de conservación de flujo de calor. Así, el estimador de la función de intensidad de $\mathbf{P}$ con ancho de banda h basado en la ecuación del calor puede calcularse resolviendo numéricamente la ecuación (3.5) a lo largo del grafo, tomando como fuentes los puntos de $\mathbf{p}$, en el instante $t = h^2$. Este estimador es computacionalmente mucho menos costoso que los estimadores equitativos propuestos en la Sección 3.2.1. Además,

bajo ciertas condiciones, que pueden consultarse en [20], este estimador y el estimador equitativo continuo de la función de intensidad son equivalentes. Un ejemplo de la estimación de la función de intensidad de un proceso puntual en un grafo puede encontrarse en la Figura 3.3.

3.2.3. Selección del ancho de banda

Recordemos que a la hora de computar cualquiera de los tres estimadores anteriores de la función de intensidad de un proceso puntual en un grafo debemos escoger el ancho de banda h. Esta es una elección mucho más importante que la de la función núcleo (en el caso de que haya que emplear una). Debemos entonces desarrollar procedimientos que nos permitan obtener un valor del ancho de banda que proporcione una buena estimación de la función de intensidad.

Como vimos en el Capítulo 1, existen diversas técnicas para la selección del ancho de banda cuando el espacio en el que se encuentra nuestra muestra es un espacio euclídeo. Un procedimiento sencillo para calcular un estimador del ancho de banda óptimo para la estimación no paramétrica de la función de densidad de una variable aleatoria d-dimensional de la que se posee una muestra aleatoria de tamaño n es la denominada regla de Scott [25], que propone tomar $\boldsymbol{H} \in \mathcal{D}$ tal que:

$$h_i = s_i \cdot \left[\frac{1}{(d+2)n}\right]^{1/(d+4)}, \quad i = 1, \ldots, d,$$

donde s_i es la desviación típica muestral de la i-ésima componente de los puntos de la muestra. En [23] se adapta la regla de Scott unidimensional para la estimación no paramétrica de la función de intensidad de un proceso puntual en un grafo lineal, proponiendo como ancho de banda:

$$h = \left(\frac{1}{3N}\right)^{1/5} \bar{s}, \tag{3.6}$$

donde N es el número de puntos observados y $\bar{s} = \left(s_1^2 + s_2^2\right)^{1/2}$, calculando s_1 y s_2 obviando el grafo; es decir, como si los puntos de $\mathbf{p}$ pudieran disponerse en cualquier punto del plano. El ancho de banda dado en la ecuación (3.6) surge de tomar $d = 1$ en la regla de Scott (al encontrarnos en un grafo lineal) y tener en cuenta la dispersión de los puntos de $\mathbf{p}$ sobre $\mathcal{L}_\mathcal{G}$ a través de su dispersión como puntos del plano. Esta modificación de la regla de Scott ha sido la empleada en las estimaciones de la intensidad dadas en la Figura 3.2 y en la Figura 3.3.

Aunque el ancho de banda propuesto en la ecuación (3.6) tiene a su favor un bajo coste computacional, emplea una medida de la dispersión de los puntos en el grafo a partir únicamente de sus coordenadas espaciales. Como ya hemos comentado al principio de este capítulo, obviar la estructura del grafo lineal suele conducir a resultados falaces. Debemos entonces desarrollar técnicas adaptadas a la estructura de grafo lineal que proporcionen un ancho de banda apropiado para la estimación de la función de intensidad en el grafo que tengan en cuenta su estructura. En [23] se propone tomar el ancho de banda h que maximice una estimación del logaritmo de

la función de verosimilitud, asumiendo que estamos ante un proceso puntual de Poisson en un grafo. Se propone por ello tomar:

$$h = \arg\max_{h'>0} \mathrm{MLCV}(h'), \quad \text{con} \quad \mathrm{MLCV}(h) = \sum_{i=1}^{N} \ln\left[\widehat{\lambda}_{-i}(\boldsymbol{p}_i)\right] - \int_{\mathcal{L}_\mathcal{G}} \widehat{\lambda}(\boldsymbol{p}) d\boldsymbol{p}, \tag{3.7}$$

siendo $\widehat{\lambda}_{-i}$ la estimación de la función de intensidad en el grafo sin emplear la i-ésima observación de la realización. El nombre de esta función obedece a las siglas en inglés del nombre de este método: *Maximum Likelihood Cross-Validation*. Notemos que en aquellos estimadores de la función de intensidad en los que $\int_{\mathcal{L}_\mathcal{G}} \widehat{\lambda}(\boldsymbol{p}) d\boldsymbol{p} = N$ para cualquier valor de h (como ocurre en el estimador basado en la ecuación del calor) puede obviarse el segundo sumando en la ecuación (3.7) de cara a la maximización de MLCV.

Los dos métodos de selección del ancho de banda que se proponen presentan ventajas e inconvenientes. Aunque el basado en la maximización de la función dada en la ecuación (3.7) tiene más sentido desde un punto de vista teórico, es computacionalmente mucho más costoso que la regla propuesta en la ecuación (3.6). Una práctica habitual es calcular primero el ancho de banda dado en la ecuación (3.6), y luego maximizar MLCV sondeando valores de h cercanos al propuesto por la regla de Scott modificada, tomando así las ventajas de un método con fundamentación teórica pero reduciendo su alto coste computacional.

Capítulo 4

Comparación de funciones de intensidad

Uno de los problemas más estudiados en el ámbito de la estadística es la comparación de dos (o más) poblaciones. Este problema consiste en plantearse si, dadas dos (o más) muestras de poblaciones (siguiendo en principio diferentes distribuciones), en realidad dichas muestras están generadas por un mismo proceso (comparación de funciones de distribución, funciones de densidad...) o si al menos comparten alguna característica (comparación de medias, medianas, varianzas...).

Este tipo de comparaciones surgen también cuando los elementos de interés son procesos puntuales. No es extraño encontrarnos en escenarios en los que se observen dos patrones de puntos en una misma región: las ubicaciones de dos especies de flora en una región de un bosque, focos de incendios forestales naturales o provocados, colisiones coche-coche y coche-moto en una red de carreteras... Es de interés preguntarse entonces si estos dos procesos puntuales poseen propiedades estadísticas en común; en particular, si comparten su función de intensidad (que sabemos caracteriza los procesos en términos de primer orden).

El desarrollo de técnicas que permitan estudiar si dos patrones de puntos son realizaciones de procesos puntuales con funciones de intensidad proporcionales (equivalente a que las densidades asociadas sean iguales) es una cuestión que ya ha sido abordada en el plano euclídeo. En [30] se propone un estadístico tipo Kolmogorov-Smirnov y se establecen condiciones para su convergencia en distribución; en [14] se propone un test basado en la distancia L^2, así como un test no paramétrico basado en la función de riesgo relativo; y en [5] se desarrolla un estadístico para la comparación de funciones de intensidad empleando covariables.

Ahora bien, en el caso de que los procesos puntuales ante los que nos encontremos sean procesos puntuales definidos en un grafo lineal, aún no se han descrito este tipo de técnicas para comparación de funciones de intensidad. En este capítulo trataremos entonces de desarrollar procedimientos, innovadores en el campo, que permitan comparar las funciones de intensidad de dos procesos puntuales definidos en un mismo grafo lineal.

Sean $\mathbf{P}_1$ y $\mathbf{P}_2$ dos procesos puntuales en un grafo lineal $\mathcal{G}$ con funciones de intensidad λ_1 y λ_2, respectivamente. De estos procesos puntuales poseemos sendas realizaciones: $\mathbf{p}_1$ y $\mathbf{p}_2$, patrones de puntos en $\mathcal{L}_\mathcal{G}$. Denotamos $N_i = \#\mathbf{p}_i$, con $i = 1, 2$. Nuestro objetivo es entonces estudiar si estos dos procesos puntuales poseen la misma estructura espacial. Notemos que esto no se traduce en que tengan la misma función de intensidad, ya que estas podrían estar en distintas escalas; es decir, el número esperado de puntos podría ser diferente, aunque su distribución espacial fuera la misma. Lo que vamos a contrastar es entonces que estos procesos puntuales tengan la misma función de densidad relativa, ya que notemos que estas están normalizadas en $\mathcal{L}_\mathcal{G}$, lo que es equivalente a que sus funciones de intensidad sean proporcionales. Así, la hipótesis nula es:

$$\mathcal{H}_0 : \exists\ \eta > 0 \text{ tal que } \lambda_1(\boldsymbol{p}) = \eta\lambda_2(\boldsymbol{p})\ \forall \boldsymbol{p} \in \mathcal{L}_\mathcal{G}. \tag{4.1}$$

Antes de proceder con la construcción de los distintos estadísticos para contrastar la hipótesis nula dada en (4.1), debemos hacer un inciso. Estrictamente hablando, si queremos formular la hipótesis nula de igualdad de funciones de densidad para dos procesos puntuales en un grafo cualquiera, debemos hacerlo a través de la proporcionalidad de las funciones de intensidad condicionadas, $\lambda_c(\boldsymbol{p}|\boldsymbol{q})$, que representa la intensidad del proceso puntual en $\boldsymbol{p} \in \mathcal{L}_\mathcal{G}$ condicionada a observar un punto en $\boldsymbol{q} \in \mathcal{L}_\mathcal{G}$. Ahora bien, bajo la hipótesis de que nuestros procesos sean procesos puntuales de Poisson, la hipótesis de independencia nos dice que $\lambda_c(\boldsymbol{p}|\boldsymbol{q}) = \lambda(\boldsymbol{p})$ para todo $\boldsymbol{p}, \boldsymbol{q} \in \mathcal{L}_\mathcal{G}$. Por ello, si asumimos que trabajamos con procesos de Poisson, podemos formular la hipótesis nula que queremos contrastar tal y como se detalla en (4.1).

4.1. Test de Kolmogorov-Smirnov

La hipótesis nula que queremos contrastar establece una relación de proporcionalidad punto a punto entre funciones de intensidad. Traduciremos esta igualdad a una serie de subconjuntos $\mathcal{L}'_\mathcal{G} \subset \mathcal{L}_\mathcal{G}$, y estudiaremos en qué grado de verifica. Integrando la igualdad dada en (4.1) en $\mathcal{L}'_\mathcal{G}$, tenemos que:

$$\mathbb{E}\left[N(\mathbf{P}_1 \cap \mathcal{L}'_\mathcal{G})\right] = \eta\mathbb{E}\left[N(\mathbf{P}_2 \cap \mathcal{L}'_\mathcal{G})\right], \quad \forall \mathcal{L}'_\mathcal{G} \subset \mathcal{L}_\mathcal{G}, \tag{4.2}$$

donde únicamente permitimos subconjuntos de $\mathcal{L}_\mathcal{G}$ que pertenezcan a su σ-álgebra de Borel (la σ-álgebra en $\mathcal{L}_\mathcal{G}$ generada por todos los abiertos en la topología inducida por la métrica del camino más corto). La ecuación (4.2) motiva una estimación global de η como $\widehat{\eta} = N_1/N_2$. Si

denotamos $N_i(\mathcal{L}'_\mathcal{G}) = \#(\mathbf{p}_i \cap \mathcal{L}'_\mathcal{G})$ para todo $\mathcal{L}'_\mathcal{G} \subset \mathcal{L}_\mathcal{G}$ e $i = 1, 2$, inspirándonos en [30] podemos definir una medida de discrepancia como:

$$D(\mathcal{L}'_\mathcal{G}) = \left|N_1(\mathcal{L}'_\mathcal{G}) - \widehat{\eta} N_2(\mathcal{L}'_\mathcal{G})\right| = N_1 \left|\frac{N_1(\mathcal{L}'_\mathcal{G})}{N_1} - \frac{N_2(\mathcal{L}'_\mathcal{G})}{N_2}\right|, \quad \forall \mathcal{L}'_\mathcal{G} \subset \mathcal{L}_\mathcal{G}. \tag{4.3}$$

Una propuesta de estadístico para determinar la discrepancia de la hipótesis nula sería el supremo de todas las posibles discrepancias dadas en la ecuación (4.3), con $\mathcal{L}'_\mathcal{G} \subset \mathcal{L}_\mathcal{G}$. El problema es que la σ-álgebra de Borel de $\mathcal{L}_\mathcal{G}$ es una familia de subconjuntos demasiado grande. Por ello, para procesos puntuales en el plano euclídeo, en [30] se encuentra suficiente tomar el supremo de las discrepancias dadas en (4.3) entre un π-sistema de W que genere la σ-álgebra de Borel.

Definición 4.1. Sea Φ un conjunto, diremos que $\mathcal{P} \subset \mathrm{P}(\Phi)$ es un **π-sistema** de Φ si $\mathcal{P} \neq \emptyset$ y $A \cap B \in \mathcal{P}$ para todo $A, B \in \mathcal{P}$. Definimos también la **σ-álgebra generada** por $\mathcal{P}$ como la menor σ-álgebra de Φ que lo contiene. Finalmente, diremos que un π-sistema en Φ **genera una σ-álgebra** en Φ si esta está contenida en la σ-álgebra generada por el π-sistema.

Así, inspirándonos en el estadístico dado en [30], tomamos $\mathcal{P}$ un π-sistema en $\mathcal{L}_\mathcal{G}$ que genere su σ-álgebra de Borel, y proponemos como estadístico para contrastar la hipótesis nula (4.1):

$$T_{KS} = \frac{1}{\xi}\sqrt{\frac{N_1 N_2}{N_1 + N_2}} \sup_{\mathcal{L}'_\mathcal{G} \in \mathcal{P}} \left|\frac{N_1(\mathcal{L}'_\mathcal{G})}{N_1} - \frac{N_2(\mathcal{L}'_\mathcal{G})}{N_2}\right|, \tag{4.4}$$

donde ξ es una constante normalizadora. En [30], la constante análoga a ξ garantiza la convergencia del estadístico a un puente Browniano cuando se estudian procesos puntuales en el plano. Sin tener ningún resultado teórico acerca de la convergencia del estadístico propuesto en la ecuación (4.4), optamos por generalizar la estimación de ξ dada en [30] como sigue: sea $\{\mathcal{L}_\gamma\}_{\gamma=1}^{\Gamma}$ una partición de $\mathcal{L}_\mathcal{G}$ tal que en todo elemento de la partición se observa al menos un punto de alguno de los dos patrones observados, y definimos, para todo $\gamma \in \{1, \ldots, \Gamma\}$ e $i = 1, 2$:

$$\widehat{N_i}(\mathcal{L}_\gamma) = N_i \cdot \frac{N_1(\mathcal{L}_\gamma) + N_2(\mathcal{L}_\gamma)}{N_1 + N_2},$$

que no es más que una estimación del número de puntos observados en cada elemento de la partición para cada patrón de puntos. Usando estas cantidades estimamos entonces ξ como:

$$\widehat{\xi} = \left[\frac{1}{\Gamma - 1}\sum_{\gamma=1}^{\Gamma}\left[\frac{[N_1(\mathcal{L}_\gamma) - \widehat{N_1}(\mathcal{L}_\gamma)]^2}{\widehat{N_1}(\mathcal{L}_\gamma)} + \frac{[N_2(\mathcal{L}_\gamma) - \widehat{N_2}(\mathcal{L}_\gamma)]^2}{\widehat{N_2}(\mathcal{L}_\gamma)}\right]\right]^{1/2}, \tag{4.5}$$

que representa una medida de discrepancia entre el número de puntos observados en cada patrón para cada elemento de la partición, y los valores estimados empleando los dos patrones de forma conjunta. En efecto, vemos como cada uno de los sumandos de (4.5) es de la forma de un estadístico χ^2. Notemos que (4.5) está bien definido ya que $\widehat{N_i}(\mathcal{L}_\gamma) > 0$ para todo $i = 1, 2$ y

$\gamma \in \{1, \ldots, \Gamma\}$, ya que por construcción hemos garantizado que en todos los elementos de la partición se observa al menos un punto de alguno de los dos patrones observados.

A la hora de calcular el estadístico dado en la ecuación (4.4), han de elegirse tanto el π-sistema, $\mathcal{P}$, como la partición del grafo. Una partición natural del grafo es su conjunto de aristas $\mathcal{A}$. Ahora bien, este podría no ser una partición válida, ya que podría darse que en alguna arista no se observara ningún punto en ninguno de los dos patrones. Esto no es un gran problema, ya que podemos tomar cada una de las aristas en las que no se observe ningún punto y unirlas, como subconjunto de $\mathcal{L}_\mathcal{G}$, a una arista en la que si se observe algún punto, obteniendo así una partición de $\mathcal{L}_\mathcal{G}$ válida. Notemos que esto es equivalente a restringir la suma dada en la ecuación (4.5) a aquellas aristas en las que se observe algún punto.

La elección del π-sistema, $\mathcal{P}$, no es una cuestión tan directa. La opción que proponemos tiene como principal objetivo simplificar la computación del estadístico propuesto en la ecuación (4.4). Proponemos tomar $\mathcal{P}$ como el conjunto de bolas en $\mathcal{L}_\mathcal{G}$, centradas en uno de sus puntos, respecto a la métrica del camino más corto en $\mathcal{L}_\mathcal{G}$, es decir:

$$\mathcal{P} = \{B_{\mathcal{L}_\mathcal{G}}(\boldsymbol{q}, r) \colon r > 0\} \quad \text{donde} \quad B_{\mathcal{L}_\mathcal{G}}(\boldsymbol{q}, r) = \{\boldsymbol{p} \in \mathcal{L}_\mathcal{G} \colon d_\mathcal{G}(\boldsymbol{p}, \boldsymbol{q}) < r\}.$$

Para definir $\mathcal{P}$ es necesario elegir el punto $\boldsymbol{q}$ a partir del cual se construye el π-sistema empleado en el estadístico. La elección de este punto depende del tipo de grafo ante el que nos encontremos. Como por defecto trabajaremos con grafos conexos, podemos distinguir si estos grafos no son o son árboles, es decir, si poseen o no ciclos. Un ejemplo de grafo con ciclos es el dado en la Figura 3.1 (a). En este tipo de grafos una elección razonable es un punto centrado en el grafo. En caso de que nuestro grafo sea un árbol, como por ejemplo el que vemos en la Figura 3.1 (b), resulta natural tomar como punto base para la construcción de $\mathcal{P}$ el nodo raíz.

4.2. Test de Cramer von Mises

Para el procedimiento que describiremos en esta sección, debemos recordar que la hipótesis nula formulada en (4.1) es equivalente a la igualdad de las funciones de densidad relativas. Por ello, bajo la hipótesis nula, la distancia (en un determinado espacio de funciones) entre estas funciones de densidad ha de ser cero. Así, una medida de discrepancia respecto de la hipótesis nula puede ser la distancia en L^2 entre las estimaciones de las funciones de densidad. Recordemos que si $\widehat{\lambda}_i(\boldsymbol{p})$ es una estimación de la función de intensidad de $\mathbf{P}_i$ como las estudiadas en el Capítulo 3, con $i = 1, 2$, entonces $\widehat{\lambda}_i(\boldsymbol{p})/N_i$ es una estimación de su función de densidad.

Proponemos entonces como estadístico de contraste para la hipótesis nula dada en (4.1):

$$T_{CvM} = \int_{\mathcal{L}_\mathcal{G}} \left[\frac{\widehat{\lambda}_1(\boldsymbol{p})}{N_1} - \frac{\widehat{\lambda}_2(\boldsymbol{p})}{N_2} \right]^2 d\boldsymbol{p}. \tag{4.6}$$

Recordemos como, bajo la hipótesis nula, las funciones de densidad asociadas a ambos procesos puntuales son iguales. Por ello, mayores valores del estadístico reflejan mayor discrepancia con la hipótesis nula.

A la hora de computar el estadístico propuesto en la ecuación (4.6), surge la cuestión a cerca de cómo de similares han de ser las estimaciones de la intensidad de cada uno de los procesos puntuales. Idealmente, nos gustaría que estas se construyeran de la forma más parecida posible, con el objetivo de que las discrepancias detectadas por el estadístico se deban a diferencias en la estructura espacial de los procesos puntuales estudiados, y no a diferencias causadas por los distintos métodos de estimación. Por ejemplo, si empleamos estimaciones no paramétricas como las estudiadas en el Capítulo 3, parece conveniente emplear en ambos casos el mismo tipo de núcleos equitativos, o en ambos casos la estimación basada en la ecuación del calor.

Una vez decididos los estimadores a emplear, el cómputo de este estadístico T_{CvM} no necesita de la elección de otros elementos, como sí ocurría en el test de Kolmogorov-Smirnov presentado en la Sección 4.1. Si bien es cierto que esta ventaja del aligeramiento en cuanto al proceso se ve difuminada por un mayor coste computacional, tal y como explicaremos en el Capítulo 5.

4.3. Test no paramétrico basado en la función de riesgo relativo

Otra forma alternativa de reescribir la hipótesis nula dada en (4.1) es emplear la función de riesgo relativo (cociente entre intensidades):

$$\mathcal{H}_0 : r(\boldsymbol{p}) = \frac{\lambda_1(\boldsymbol{p})}{\lambda_2(\boldsymbol{p})} \text{ es constante a lo largo de } \mathcal{L}_\mathcal{G}. \tag{4.7}$$

Así, trataremos de contrastar si el cociente de intensidades varía a lo largo del grafo. Ahora bien, notemos que debido al posible distinto número esperado de puntos observados en la realización de cada uno de los procesos puntuales, no conocemos el valor de la constante a la que iguala la función de riesgo relativo. Por ello, resulta más apropiado contrastar que el cociente de las funciones de densidad de cada proceso puntual en el grafo sea constate e igual a 1, o equivalentemente que su logaritmo sea nulo en todo el grafo. Por ello, estimamos el logaritmo de la función de riesgo relativo como:

$$\widehat{\rho}(\boldsymbol{p}) = \ln \left[\frac{N_2}{N_1} \frac{\widehat{\lambda}_1(\boldsymbol{p})}{\widehat{\lambda}_2(\boldsymbol{p})} \right].$$

Así, podemos formular un test de efecto de $\boldsymbol{p}$ sobre $\widehat{\rho}(\boldsymbol{p})$: si reindexamos los puntos observados como $\mathbf{p}_1 \bigcup \mathbf{p}_2 = \{\boldsymbol{p}_j\}_{j=1}^N$, con $N = N_1 + N_2$, podemos entender los valores de $\{\widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$ como una muestra de una variable respuesta frente a la variable explicativa dada por la posición en el grafo, con valores asociados $\{\boldsymbol{p}_j\}_{j=1}^N$. Tenemos entonces un problema de regresión en el grafo, que nos permite reformular el test con hipótesis nula dada en (4.7) como el test de efecto siguiente:

$$\mathcal{H}_0 : \mathbb{E}\left[\widehat{\rho}(\boldsymbol{p}_j)|\boldsymbol{p}_j\right] = \mu \quad \text{frente a} \quad \mathcal{H}_a : \mathbb{E}\left[\widehat{\rho}(\boldsymbol{p}_j)|\boldsymbol{p}_j\right] = m(\boldsymbol{p}_j), \tag{4.8}$$

siendo m una función a lo largo del grafo, usualmente denominada función de regresión. Necesitamos entonces herramientas de regresión en grafos lineales. Este es otro de los temas que no han sido abordados previamente en la literatura existente. Por ello, vamos a desarrollar un método no paramétrico de estimación de la función de regresión en un grafo lineal. Si en un contexto de regresión tenemos n valores de la variable respuesta $\{y_j\}_{j=1}^n$ y n valores de una variable explicativa $\{x_j\}_{j=1}^n$, una estimación no paramétrica de la función de regresión es la dada por el estimador de Nadaraya-Watson que, tal y como se detalla en [27], podemos escribir como:

$$\widehat{m}(x) = \frac{\sum_{j=1}^n S_h(x - x_j) y_j}{\sum_{j=1}^n S_h(x - x_j)}, \tag{4.9}$$

donde S es una función núcleo, h es el parámetro de ventana, y $S_h(\cdot) = h^{-1}S(\cdot/h)$. Las condiciones sobre S pueden consultarse en [27], aunque los núcleos habitualmente empleados en la estimación no paramétrica de la función de densidad siguen siendo válidos en este contexto.

Ahora bien, cuando nos encontramos ante un proceso puntual en un grafo lineal, la estimación de la función de densidad de este proceso ha de hacerse empleando las técnicas estudiadas en el Capítulo 3. Si empleamos núcleos equitativos, podemos escribir una estimación de la función de densidad de un proceso puntual en un grafo mediante una estimación de su intensidad como:

$$\frac{1}{N}\widehat{\lambda}(\boldsymbol{p}) = \frac{1}{N}\sum_{j=1}^N G_{\boldsymbol{p}_j,h}(\boldsymbol{p}), \tag{4.10}$$

siendo $\{\boldsymbol{p}_j\}_{j=1}^N$ un patrón de puntos en el grafo, y $G_{\boldsymbol{p}_j,h}$ el núcleo equitativo centrado en $\boldsymbol{p}_j$ con ancho de banda h. Comparando entonces los estimadores de la función de densidad en $\mathbb{R}$ y en un grafo lineal, e inspirándonos en el estimador de Nadaraya-Watson dado en la ecuación (4.9), proponemos un estimador no paramétrico de la función de regresión en un grafo lineal como:

$$\widehat{m}(\boldsymbol{p}) = \frac{\sum_{j=1}^N G_{\boldsymbol{p}_j,h}(\boldsymbol{p}) y_j}{\sum_{j=1}^N G_{\boldsymbol{p}_j,h}(\boldsymbol{p})}, \tag{4.11}$$

siendo $\{y_j = \widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$ el conjunto de observaciones de la variable respuesta, asociados a una serie de localizaciones en el grafo $\{\boldsymbol{p}_j\}_{j=1}^N \subset \mathcal{L}_\mathcal{G}$, y donde $G_{\boldsymbol{p}_j,h}$ sigue siendo un núcleo equitativo centrado en $\boldsymbol{p}_j$ con ancho de banda h.

En el contexto del problema de regresión ante el que nos encontramos, tenemos $N = N_1 + N_2$, los valores observados de la variable respuesta son $\{\widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$ asociados a las localizaciones en el grafo $\{\boldsymbol{p}_j\}_{j=1}^N = \mathbf{p}_1 \bigcup \mathbf{p}_2 \subset \mathcal{L}_{\mathcal{G}}$. De acuerdo con (4.8), el modelo bajo la hipótesis nula puede estimarse mediante la media muestral:

$$\widehat{\mu} = \frac{1}{N}\sum_{j=1}^{N}\widehat{\rho}(\boldsymbol{p}_j).$$

Por otra parte, la estimación de la función de regresión bajo la hipótesis alternativa emplea el estimador no paramétrico dado en (4.11):

$$\widehat{m}(\boldsymbol{p}_i) = \frac{\sum_{j=1}^{N} G_{\boldsymbol{p}_j,h}(\boldsymbol{p}_i)\widehat{\rho}(\boldsymbol{p}_j)}{\sum_{k=1}^{N} G_{\boldsymbol{p}_k,h}(\boldsymbol{p}_i)} = \sum_{j=1}^{N} \boldsymbol{S}_{i,j}\widehat{\rho}(\boldsymbol{p}_j),$$

donde hemos definido la matriz de suavizado $\boldsymbol{S} \in \mathcal{M}_{N\times N}(\mathbb{R})$ como:

$$\boldsymbol{S}_{i,j} = \frac{G_{\boldsymbol{p}_j,h}(\boldsymbol{p}_i)}{\sum_{k=1}^{N} G_{\boldsymbol{p}_k,h}(\boldsymbol{p}_i)}, \quad i,j \in \{1,\ldots,N\}.$$

Como el contraste de efecto propuesto en (4.8) puede entenderse como un contraste entre modelos anidados (dónde el modelo bajo la hipótesis nula es un caso particular del modelo bajo la hipótesis alternativa), emplearemos un test F. Definimos los residuos del modelo bajo la hipótesis nula como:

$$\mathrm{RSS}_0 = \sum_{j=1}^{N}\left[\widehat{\rho}(\boldsymbol{p}_j) - \widehat{\mu}\right]^2,$$

y bajo la hipótesis alternativa como:

$$\mathrm{RSS}_a = \sum_{j=1}^{N}\left[\widehat{\rho}(\boldsymbol{p}_j) - \widehat{m}(\boldsymbol{p}_j)\right]^2.$$

Como bajo la hipótesis nula estimamos un único parámetro, el número de grados de libertad de los residuos bajo la hipótesis nula es $df_0 = N - 1$. Para definir los grados de libertad de los residuos bajo la hipótesis alternativa seguimos la estrategia propuesta en [6], donde, en analogía con el modelo lineal general, se propone tomar $df_a = tr(\boldsymbol{I}_N - \boldsymbol{S})$. Entonces el estadístico de contraste para el test de efecto viene dado por:

$$T_{NP} = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_a)/(df_0 - df_a)}{\mathrm{RSS}_a/df_a}. \tag{4.12}$$

Al igual que ocurría en el test de Cramer von Mises presentado en la Sección 4.2, para computar el valor del estadístico propuesto en la ecuación (4.12) debemos elegir cómo vamos a llevar a cabo las distintas estimaciones no paramétricas requeridas. Reparamos primeramente en

el cálculo de $\{\widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$; debemos procurar de nuevo que las estimaciones de la intensidad de cada uno de los procesos puntuales sean lo más similares en lo que al proceso de estimación se refiere, por ello, es recomendable emplear el mismo tipo de núcleos equitativos, o la estimación basada en la ecuación del calor, en ambas estimaciones.

Además, para que los valores de $\{\widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$ estén bien definidos, necesitamos que los anchos de banda sean suficientemente grandes como para que la estimación de las funciones de intensidad de los procesos puntuales en estudio no se anulen en ninguno de los puntos de los dos patrones observados. Esto hace que no podamos tomar anchos de banda demasiado pequeños, lo que aumentará el coste computacional del cálculo de los valores $\{\widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$, ya el coste del cálculo de los núcleos equitativos aumenta con h.

Nótese que únicamente vamos a necesitar los valores de las estimaciones de la intensidad en los puntos de los patrones observados. Por ello, puede ser recomendable determinar estas cantidades empleando validación cruzada; es decir:

$$\widehat{\rho}(\boldsymbol{p}_j) = \begin{cases} \ln\left[\frac{N_2}{N_1-1}\frac{\widehat{\lambda}_{1,-j}(\boldsymbol{p}_j)}{\widehat{\lambda}_2(\boldsymbol{p}_j)}\right] & \text{si } \boldsymbol{p}_j \in \mathbf{p}_1 \\ \ln\left[\frac{N_2-1}{N_1}\frac{\widehat{\lambda}_1(\boldsymbol{p}_j)}{\widehat{\lambda}_{2,-j}(\boldsymbol{p}_j)}\right] & \text{si } \boldsymbol{p}_j \in \mathbf{p}_2 \end{cases},$$

donde $\widehat{\lambda}_{i,-j}$ es la estimación de la función de intensidad de $\mathbf{P}_i$ empleando como patrón de puntos $\mathbf{p}_i \backslash \{\boldsymbol{p}_j\}$, con $j \in \{1,\ldots,N\}$ e $i = 1, 2$.

A la hora de computar la estimación no paramétrica de la función de regresión en el grafo, debemos elegir nuevamente qué tipo de núcleos equitativos vamos a usar, así como qué ancho de banda, ya que este no tiene porqué coincidir con ninguno de los empleados anteriormente en el cálculo de los valores $\{\widehat{\rho}(\boldsymbol{p}_j)\}_{j=1}^N$.

4.4. Procedimiento de calibración

A lo largo de este capítulo hemos descrito tres procedimientos para construir estadísticos de contraste sobre la hipótesis nula dada en (4.1). Ahora bien, para afirmar si a un determinado nivel de significación tenemos evidencias significativas a favor o en contra de la hipótesis nula, el valor del estadístico de contraste no es suficiente, necesitamos conocer su nivel crítico [16]:

Definición 4.2. Sea el contraste de hipótesis $\mathcal{H}_0 : \theta \in \Theta_0$ frente a $\mathcal{H}_a : \theta \in \Theta_a$, para el cual se posee un estadístico conveniente $T(x_1,\ldots,x_s)$, siendo $(x_1,\ldots,x_s)$ una muestra. Se denomina **nivel crítico** o **p-valor** de una muestra $(y_1,\ldots,y_l)$ para el contraste en cuestión mediante el estadístico T a:

$$\mathfrak{p}(y_1,\ldots,y_l) = \sup_{\theta\in\Theta_0} \mathbb{P}\left[T(x_1,\ldots,x_s) \geq T(y_1,\ldots,y_l)|\theta\right],$$

donde $\mathbb{P}(A|B)$ denota la probabilidad del suceso A condicionada al suceso B.

Notemos que carecemos de resultados teóricos acerca de la distribución de nuestros estadísticos bajo la hipótesis nula que nos permitieran calcular p-valores de manera sencilla a partir de dicha distribución. Por ello, una posible solución es emplear el test de permutaciones adaptado al problema de las dos muestras. Considerados dos patrones de puntos que entendemos como realizaciones de sendos procesos puntuales en un grafo, se calcula uno de los tres estadísticos de contraste que hemos propuesto. Para estimar su nivel crítico, se consideran todas las posibles permutaciones de los puntos de los patrones observados, tomando dos subconjuntos de $\{\boldsymbol{p}_j\}_{j=1}^{N}$ de cardinales N_1 y N_2. Para cada una de estas permutaciones se calcula el valor del estadístico de contraste. Hecho esto, el p-valor se estima como la fracción de estos estadísticos que son mayores que el de partida (el asociado a los patrones de puntos sin permutar). En el caso de que ninguno de estos estadísticos sea mayor que el de partida (lo que conduciría a una estimación nula del nivel crítico) consideraremos como estimación del p-valor la mitad del inverso del número de permutaciones realizadas.

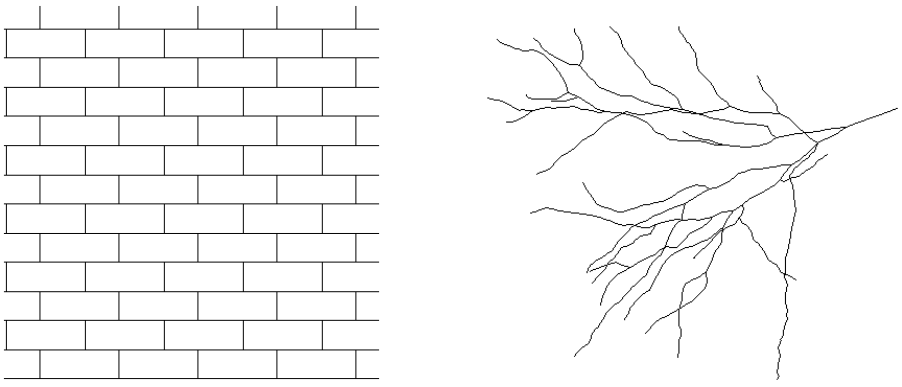
En la práctica este procedimiento presenta el problema de que el número total de permutaciones es computacionalmente inabordable a partir de tamaños muestrales no muy grandes. Por ello, una práctica estándar es considerar un subconjunto de n_p permutaciones, escogidas al azar, de entre todas las posibles. La elección de n_p no es cuestión menor, ya que debe escogerse suficientemente grande como para dar resultados fiables, al tiempo que ha de ser computacionalmente realizable. Como se discute en [19], tomar $n_p = 5000$ resulta suficiente en la mayoría de escenarios, llegando en casos excepcionales a tomar $n_p = 10000$ permutaciones. Por ello, de partida fijamos $n_p = 7500$.

Capítulo 5

Estudio de simulación

En el Capítulo 4 hemos diseñado procedimientos para contrastar si dos patrones de puntos, observados en un mismo grafo, están generados por procesos puntuales con funciones de intensidad proporcionales. En el presente capítulo presentamos el exhaustivo estudio de simulación llevado a cabo para analizar el comportamiento de nuestras propuestas. Al tratarse de procedimientos de contraste, el estudio de su comportamiento radica fundamentalmente en valorar dos cuestiones: el nivel (probabilidad de rechazar cuando la hipótesis nula es cierta) y la potencia (la probabilidad de rechazar cuando la hipótesis nula es falsa).

Para evaluar estas cantidades emplearemos técnicas de Monte Carlo: fijado un modelo, generaremos $M = 1000$ realizaciones (réplicas Monte Carlo) de cada uno de los dos procesos puntuales considerados, y determinaremos la conclusión de cada uno de los procedimientos de contraste empleando el test de permutaciones introducido en la Sección 4.4.



(a) *Spiders*. (b) *Dendrite*.

Figura 5.1: representación de los grafos *Spiders* (a) y *Dendrite* (b) que consideraremos en el estudio de simulación.

La primera elección que debemos hacer para construir nuestro estudio de simulación son los grafos en los que vamos a considerar los procesos puntuales. Recordemos que el cómputo del estadístico T_{KS} requiere de la elección de un punto distinguido en el grafo, el cual se elige de forma distinta en función de si dicho grafo es o no un árbol. Esto nos motiva a considerar dos soportes: como grafo con ciclos tomaremos el dado en la Figura 5.1 (a), que denominaremos *Spiders*, y como árbol usaremos el mostrado en la Figura 5.1 (b), al que nos referiremos como *Dendrite*, ambos obtenidos del paquete [4].

5.1. Modelos bajo la hipótesis nula

El primer paso de nuestro estudio de simulación consistirá en analizar el nivel de los contrastes propuestos en el Capítulo 4. Recordemos que fijado un nivel de significación $\alpha \in [0, 1]$, diremos que una muestra proporciona evidencias significativas en contra de la hipótesis nula en un contraste (empleando un determinado estadístico) cuando el p-valor asociado sea menor que α. Además, como estos niveles críticos se distribuyen uniformemente en el intervalo $[0, 1]$, ver [16], dada una muestra de niveles críticos, la fracción de estos menores que α debería ser precisamente α. Por ello, para estudiar el nivel, calcularemos en qué fracción de las M muestras replicadas se rechaza la hipótesis nula a un nivel de significación α, y compararemos este valor precisamente con α. Emplearemos los tres niveles críticos usuales: $\alpha \in \{0.1,\ 0.05,\ 0.01\}$, así como distintos valores del tamaño muestral esperado.

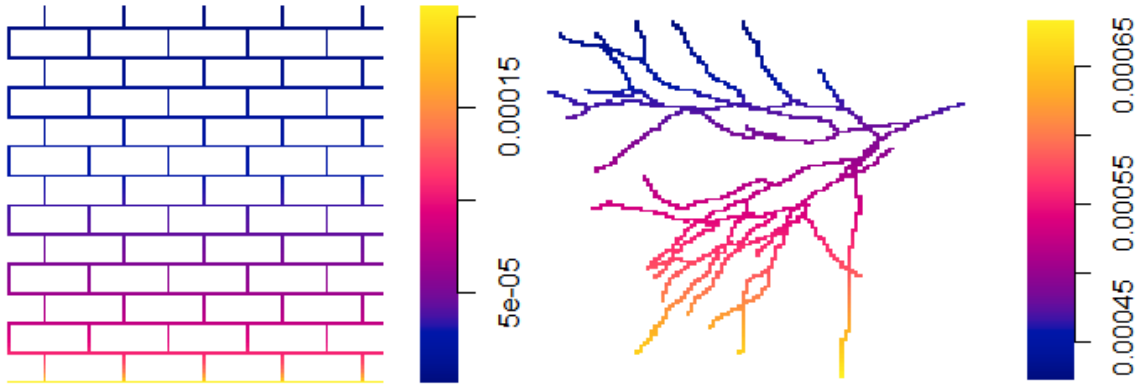
Para estudiar el nivel debemos considerar entonces modelos bajo la hipótesis nula, es decir, debemos considerar dos procesos puntuales generados con funciones de intensidad proporcionales; consideraremos en particular un escenario homogéneo y otro inhomogéneo. Si denotamos por m el número esperado de puntos, la función de intensidad homogénea que tomaremos será:

$$\lambda_{Hom}(\boldsymbol{p}) = \frac{m}{|\mathcal{L}_{\mathcal{G}}|}, \quad \forall \boldsymbol{p} \in \mathcal{L}_{\mathcal{G}}, \tag{5.1}$$

donde $\mathcal{L}_{\mathcal{G}}$ será el conjunto de puntos del plano del grafo considerado. Por su parte, la función de intensidad inhomogénea que consideraremos será:

$$\lambda_{Inh}(\boldsymbol{p}) = m \cdot \frac{1 - \sqrt[10]{p_2/1125}}{\int_{\mathcal{L}_{\mathcal{G}}} \left(1 - \sqrt[10]{q_2/1125}\right) d\boldsymbol{q}}, \quad \forall \boldsymbol{p} = (p_1, p_2) \in \mathcal{L}_{\mathcal{G}}. \tag{5.2}$$

En la Figura 5.2 se representa la función de densidad asociada a la función de intensidad inhomogénea dada en la ecuación (5.2), que se obtiene en el caso $m = 1$. Se ha decidido omitir la representación gráfica de los modelos homogéneos por tratarse de constantes sobre los grafos que no necesitan ayuda para su visualización.



(a) *Spiders*. (b) *Dendrite*.

Figura 5.2: representación de la función de densidad asociada a la intensidad inhomogénea, λ_{Inh}, definida en la ecuación (5.2), en los grafos *Spiders* (a) y *Dendrite* (b).

5.2. Modelos bajo la hipótesis alternativa

Una vez contrastado el buen ajuste del nivel de las propuestas, el siguiente paso es estudiar la potencia de las mismas. La potencia de un contraste es intuitivamente su capacidad de rechazar la hipótesis nula cuando la hipótesis alternativa es cierta. Para estudiar la potencia generaremos ahora pares de patrones de puntos procedentes de procesos puntuales en un mismo grafo con intensidades no proporcionales. Simularemos $M = 1000$ réplicas Monte Carlo y, aplicando cada una de las propuestas, obtendremos las correspondientes conclusiones de los contrastes a un nivel de significación $\alpha = 0.05$.

De nuevo analizaremos el escenario con intensidad homogénea e inhomogénea. En ambos casos, emplearemos para una de las poblaciones los modelos dados en (5.1) y (5.2), que denotaremos por λ_1. Como estamos bajo la hipótesis alternativa, necesitamos unos segundos modelos con intensidades no proporcionales a las anteriores. Para poder tener cierto grado de control sobre la discrepancia con la hipótesis nula, nos inspiramos en [14] y construimos λ_2 a partir de λ_1 añadiendo en media a puntos distribuidos de forma homogénea en un subgrafo de $\mathcal{L}_\mathcal{G}$:

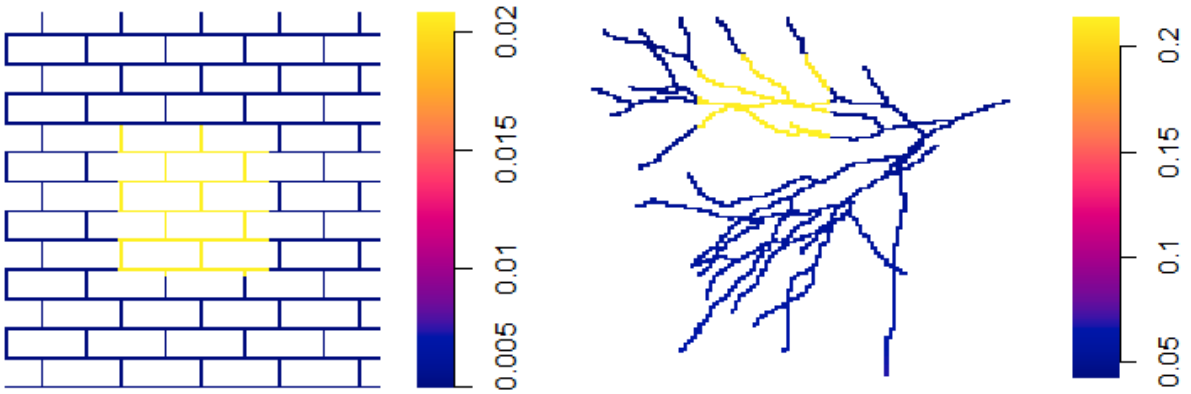
$$\lambda_2(\boldsymbol{p}) = \lambda_1(\boldsymbol{p}) + \frac{a}{|\mathcal{L}_\mathcal{G}^+|}\mathbf{1}_{\mathcal{L}_\mathcal{G}^+}(\boldsymbol{p}), \quad \forall \boldsymbol{p} \in \mathcal{L}_\mathcal{G}, \tag{5.3}$$

construida de tal forma que $\int_{\mathcal{L}_\mathcal{G}} \lambda_2(\boldsymbol{p})d\boldsymbol{p} = m + a$. Debemos especificar las regiones de los grafos donde se añaden estos a puntos de forma homogénea. Para *Spiders* optamos por:

$$\mathcal{L}^+_{Spiders} = \{\boldsymbol{p} = (p_1, p_2) \in \mathcal{L}_{Spiders} \colon \boldsymbol{p} \in [337.5,\ 787.5] \times [337.5,\ 787.5]\}, \tag{5.4}$$

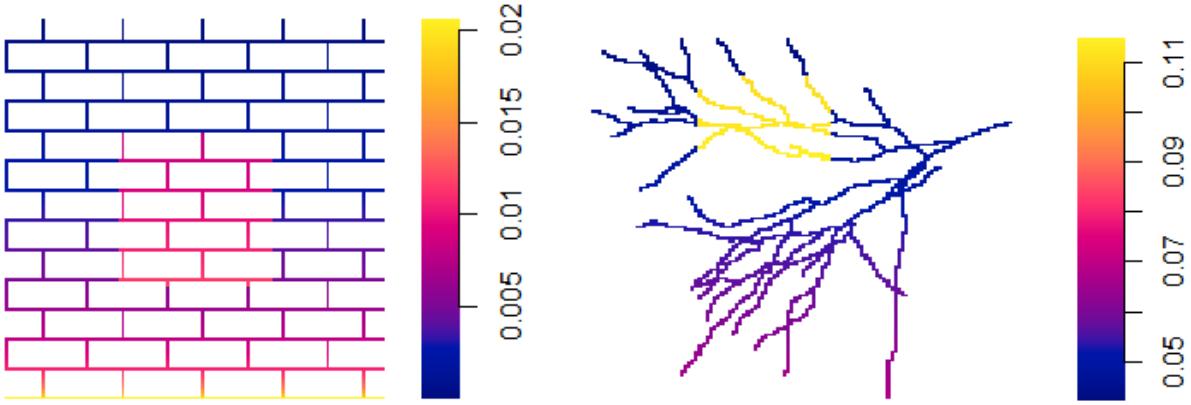
y para *Dendrite*:

$$\mathcal{L}^+_{Dendrite} = \{\boldsymbol{p} = (p_1, p_2) \in \mathcal{L}_{Dendrite} \colon \boldsymbol{p} \in [73.02,\ 139.955] \times [242.2,\ 298.3]\}. \tag{5.5}$$



(a) *Spiders*. (b) *Dendrite*.

Figura 5.3: representación de la función de intensidad homogénea perturbada de acuerdo a la ecuación (5.3) en los grafos *Spiders* (a) y *Dendrite* (b) para el caso $m = 100$ y $a = 50$.



(a) *Spiders*. (b) *Dendrite*.

Figura 5.4: representación de la función de intensidad inhomogénea perturbada de acuerdo a la ecuación (5.3) en los grafos *Spiders* (a) y *Dendrite* (b) para el caso $m = 100$ y $a = 20$.

En las ecuaciones (5.4) y (5.5) entendemos las coordenadas de los puntos del grafo como sus coordenadas en el sistema de referencia euclidiano en el que el grafo está embebido. En la Figura 5.3 representamos la función de intensidad homogénea λ_{Hom} modificada de acuerdo a la ecuación (5.3) en los dos grafos en estudio, y en la Figura 5.4 lo hacemos para la inhomogénea, λ_{Inh}.

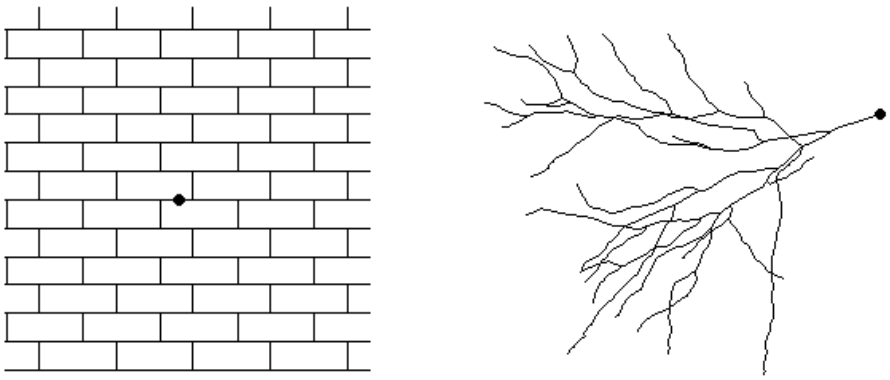
El parámetro a es el que determina la "distancia" que cada uno de los posibles escenarios presenta con la hipótesis nula de intensidades proporcionales, teniendo para $a = 0$ la hipótesis nula. Al igual que en el estudio del nivel, vamos a emplear varios tamaños muestrales esperados, m, y para cada uno de estos posibles valores queremos alejarnos de la hipótesis nula "al mismo ritmo". Esto motiva tomar una colección de valores de a que dependa de m. En este caso optamos por tomar múltiplos naturales de $m/4$.

5.3. Resultados

En esta sección presentamos los resultados del estudio de simulación, cuyos detalles hemos descrito en secciones anteriores, para los estadísticos de contraste propuestos en el Capítulo 4.

5.3.1. Test de Kolmogorov-Smirnov

Como se detalló en el Capítulo 4, en la definición de este estadístico está involucrado un π-sistema que tenemos que escoger, para lo que se necesita determinar el punto de la base del mismo. Para *Spiders* consideramos el punto del grafo más cercano (en distancia euclídea) al centroide de todos los nodos del grafo, mientras que en *Dendrite* consideramos el nodo raíz del árbol, que representamos en la Figura 5.5.



(a) *Spiders*. (b) *Dendrite*.

Figura 5.5: representación del punto base del π-sistema en *Spiders* (a) y *Dendrite* (b).

En la Tabla 5.1 mostramos la proporción de niveles críticos (obtenidos mediante el test de permutaciones) que nos llevan a rechazar la hipótesis nula a un nivel α, para distintos valores de m, para las funciones de intensidad propuestas en las ecuaciones (5.1) y (5.2) en *Spiders* y en la Tabla 5.2 los resultados análogos para *Dendrite*. Estos resultados muestran que el estadístico de Kolmogorov-Smirnov está bien calibrado a partir de tamaños muestrales del orden de $m = 100$; e incluso no se desvía excesivamente para tamaños muestrales pequeños como $m = 50$ y $m = 20$.

Habiendo obtenido resultados satisfactorios en el estudio del nivel del test de Kolmogorov-Smirnov, procedemos con su potencia. Consideraremos los dos modelos introducidos en la Sección 5.2. Para el caso homogéneo obtenemos los resultados que vemos en la Tabla 5.3, mientras que los del caso inhomogéneo se muestran en la Tabla 5.4. Lo primero que debemos comentar es que

	Modelo homogéneo			Modelo inhomogéneo		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$m = 20$	0.095	0.049	0.012	0.084	0.037	0.007
$m = 50$	0.093	0.054	0.014	0.099	0.056	0.009
$m = 100$	0.090	0.042	0.009	0.099	0.056	0.009
$m = 200$	0.092	0.040	0.007	0.105	0.050	0.015
$m = 500$	0.107	0.051	0.007	0.107	0.048	0.003

Tabla 5.1: proporción de rechazos bajo la hipótesis nula a distintos niveles de significación, $\alpha \in \{0.1,\ 0.05,\ 0.01\}$, para el test de Kolmogorov-Smirnov en el grafo *Spiders*, con modelos homogéneo e inhomogéneo y tamaños muestrales en media $m = 20,\ 50,\ 100,\ 200$ y 500.

	Modelo homogéneo			Modelo inhomogéneo		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$m = 20$	0.112	0.068	0.020	0.093	0.045	0.007
$m = 50$	0.112	0.067	0.015	0.096	0.046	0.010
$m = 100$	0.093	0.058	0.016	0.093	0.050	0.014
$m = 200$	0.094	0.059	0.009	0.097	0.046	0.004
$m = 500$	0.104	0.051	0.015	0.113	0.061	0.012

Tabla 5.2: proporción de rechazos bajo la hipótesis nula a distintos niveles de significación, $\alpha \in \{0.1,\ 0.05,\ 0.01\}$, para el test de Kolmogorov-Smirnov en el grafo *Dendrite*, con modelos homogéneo e inhomogéneo y tamaños muestrales en media $m = 20,\ 50,\ 100,\ 200$ y 500.

los valores de a escogidos no han sido los mismos para el estudio de la potencia en *Spiders* y en *Dendrite*. Esto obedece a que en *Spiders* se alcanzan valores de potencia muy altos para valores de a menores que en *Dendrite*, por lo que no es necesario "alejarnos" tanto de la hipótesis nula.

Tanto en la Tabla 5.3 como en la Tabla 5.4 vemos como para cada valor de m, según aumenta el valor de a, la potencia aumenta. Esto era de esperar, ya que nos alejamos de la hipótesis nula. Destacar también como, fijado a en función de m, según aumentamos también m aumenta la potencia, ya que tenemos más información para discernir si las funciones de intensidad de los procesos puntuales que generan los patrones observados son o no proporcionales. Estos comportamientos esperables se observan en los dos grafos que hemos estudiado. Por otra parte, la potencia crece más rápidamente en *Spiders* que en *Dendrite*, y alcanza antes (con menos discrepancia con la hipótesis nula) valores próximos a la unidad. En cualquier caso, se puede concluir que los resultados del estudio de la potencia para el test de Kolmogorov-Smirnov son satisfactorios.

	Spiders			*Dendrite*			
	$a = 3m/4$	$a = m/2$	$a = m/4$	a=$7m/4$	$a = 3m/2$	$a = 5m/4$	$a = m$
$m = 50$	0.915	0.639	0.279	0.328	0.319	0.264	0.214
$m = 100$	0.998	0.96	0.536	0.651	0.616	0.526	0.429
$m = 200$	1	1	0.856	0.964	0.937	0.898	0.799
$m = 500$	1	1	0.997	1	1	1	1

Tabla 5.3: proporción de rechazos bajo la hipótesis alternativa al nivel de significación $\alpha = 0.05$ para el test de Kolmogorov-Smirnov en los grafos *Spiders* y *Dendrite*, en el caso homogéneo, para diferentes tamaños muestrales esperados del primer patrón de puntos ($m = 50,\ 100,\ 200$ y 500) y diferentes valores del número de puntos añadidos de forma esperada homogéneamente en $\mathcal{L}_{\mathcal{G}}^{+}$ en el segundo patrón de puntos, a.

	Spiders			*Dendrite*			
	$a = 3m/4$	$a = m/2$	$a = m/4$	a=$7m/4$	$a = 3m/2$	$a = 5m/4$	$a = m$
$m = 50$	0.972	0.792	0.327	0.35	0.328	0.291	0.245
$m = 100$	0.999	0.988	0.618	0.675	0.608	0.535	0.445
$m = 200$	1	1	0.914	0.963	0.923	0.891	0.813
$m = 500$	1	1	1	1	1	1	0.999

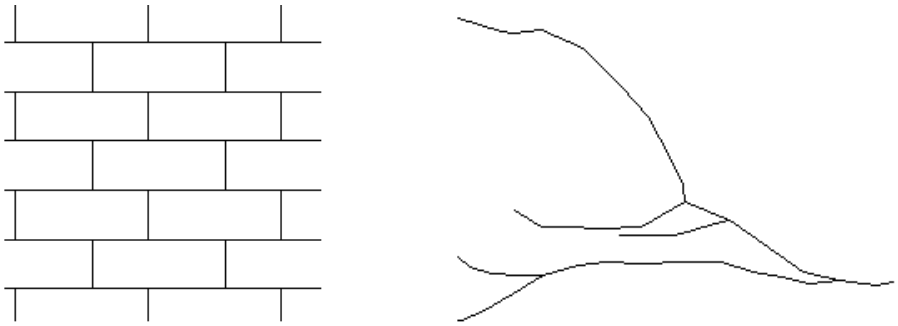
Tabla 5.4: proporción de rechazos bajo la hipótesis alternativa al nivel de significación $\alpha = 0.05$ para el test de Kolmogorov-Smirnov en los grafos *Spiders* y *Dendrite*, en el caso inhomogéneo, para diferentes tamaños muestrales esperados del primer patrón de puntos ($m = 50,\ 100,\ 200$ y 500) y diferentes valores del número de puntos añadidos de forma esperada homogéneamente en $\mathcal{L}_{\mathcal{G}}^{+}$ en el segundo patrón de puntos, a.

5.3.2. Test de Cramer von Mises

En primer lugar notemos la distinta naturaleza entre el estadístico de Cramer von Mises y el de Kolmogorov-Smirnov. El test de Cramer von Mises no calcula distancias entre pares de puntos, sino que mide distancias entre estimaciones de la función de intensidad.

En un primer paso, necesitamos decidir qué estimación vamos a emplear. Recordando las consideraciones hechas en la Sección 4.2, hemos optado por emplear en todos los casos la estimación basada en la ecuación del calor, introducida en la Sección 3.2.2, escogiendo en cada caso el ancho de banda mediante la regla de Scott modificada, definida en la ecuación (3.6). Estas elecciones tienen como principal objetivo reducir el coste computacional del cálculo de este estadístico.

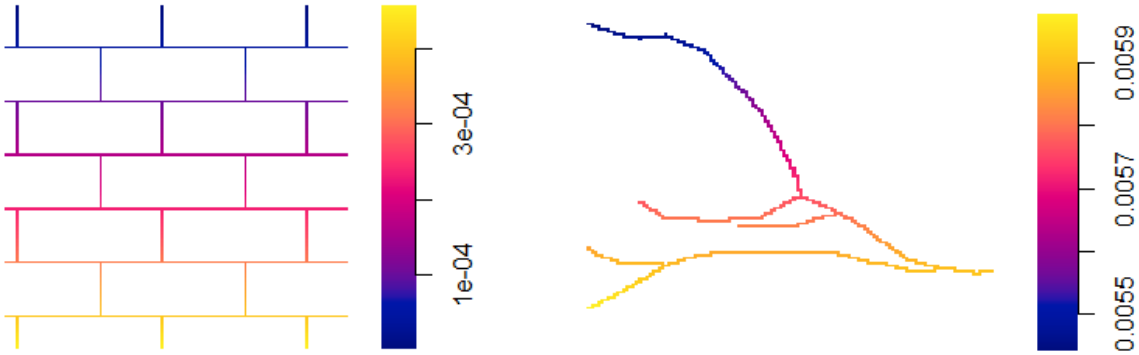
Aunque el contraste de Cramer von Mises no está condicionado a que el grafo sea o no un árbol, por consistencia con el estudio realizado para el test de Kolmogorov-Smirnov estudiaremos el nivel y la potencia del mismo tanto en *Spiders* como en *Dendrite*. A pesar de las elecciones hechas en lo referente al método de estimación y selección de los correspondientes parámetros, el número elevado de nodos de los grafos propuestos hace computacionalmente inabordable en tiempos factibles el cálculo sobre los mismos de este estadístico.



(a) Subgrafo de *Spiders*. (b) Subgrafo de *Dendrite*.

Figura 5.6: representación de los subgrafos de *Spiders* (a) y de *Dendrite* (b) empleados en el estudio del nivel y la potencia para el estadístico de Cramer von Mises.

Con el objetivo de reducir este coste computacional, pero manteniendo la esencia de los modelos descritos anteriormente optamos por considerar "subgrafos" de *Spiders* y de *Dendrite* que tengan un menor número de nodos, pero sigan representando la estructura geométrica de los mismos. Los subgrafos que vamos a emplear pueden verse en la Figura 5.6, y se han obtenido limitándonos a un subconjunto (en el plano euclídeo) de los grafos originales. Notemos que el subgrafo de *Spiders* sigue teniendo ciclos, y que el de *Dendrite* sigue siendo un árbol. Esta acción no fue suficiente, por ello, para reducir todavía más el coste computacional reduciremos el número



(a) Subgrafo de *Spiders*.

(b) Subgrafo de *Dendrite*.

Figura 5.7: representación de la función de densidad asociada a la intensidad inhomogénea, λ_{Inh}, definida en la ecuación (5.2), en los subgrafos de *Spiders* (a) y *Dendrite* (b).

de permutaciones consideradas a $n_p = 5000$, elección que según [19] no pone en compromiso el proceso de calibrado.

Los modelos bajo la hipótesis nula que consideraremos para el estudio del nivel serán los que hemos detallado en la Sección 5.1: el modelo con intensidad homogénea definido en la ecuación (5.1), y en la ecuación (5.2) para el caso inhomogéneo. Las funciones de densidad relativas a estas funciones de intensidad pueden verse representadas en los subgrafos de *Spiders* y de *Dendrite* en la Figura 5.7, dónde vemos que el comportamiento espacial es el mismo que el observado en la Figura 5.2 para los grafos originales.

En la Tabla 5.5 y Tabla 5.6 mostramos los resultados del calibrado del nivel del estadístico Cramer von Mises para el subgrafo de *Spiders* y de *Dendrite* respectivamente. Podemos observar como a partir de tamaños muestrales esperados de orden $m = 100$ el contraste está bien calibrado, si bien es cierto que para el menor nivel de significación, $\alpha = 0.01$, presenta ciertas limitaciones, que se ven subsanadas al aumentar el tamaño muestral.

Para el estudio de la potencia del test de Cramer von Mises consideraremos los modelos bajo la hipótesis alternativa que hemos introducido en la Sección 5.2. De nuevo emplearemos los subgrafos de *Spiders* y *Dendrite* que hemos introducido en la Figura 5.6. En este caso los subconjuntos $\mathcal{L}_{\mathcal{G}}^{+}$ en los que se añaden en media a puntos homogéneamente vienen dados por:

$$\mathcal{L}^{+}_{\text{Subgrafo } \textit{Spiders}} = \{\boldsymbol{p} = (p_1, p_2) \in \mathcal{L}_{\text{Subgrafo } \textit{Spiders}} \colon \boldsymbol{p} \in [735.25,\ 984.375] \times [735.25,\ 956.25]\}\,,$$

y

$$\mathcal{L}^{+}_{\text{Subgrafo } \textit{Dendrite}} = \{\boldsymbol{p} = (p_1, p_2) \in \mathcal{L}_{\text{Subgrafo } \textit{Dendrite}} \colon p_2 > 295.9625\}\,.$$

	Modelo homogéneo			Modelo inhomogéneo		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$m = 20$	0.077	0.035	0.001	0.073	0.026	0.002
$m = 50$	0.076	0.037	0.005	0.087	0.038	0.004
$m = 100$	0.080	0.046	0.002	0.097	0.05	0.011
$m = 200$	0.107	0.044	0.003	0.098	0.051	0.006
$m = 500$	0.090	0.043	0.013	0.092	0.047	0.008
$m = 750$	0.101	0.044	0.009	0.081	0.038	0.008

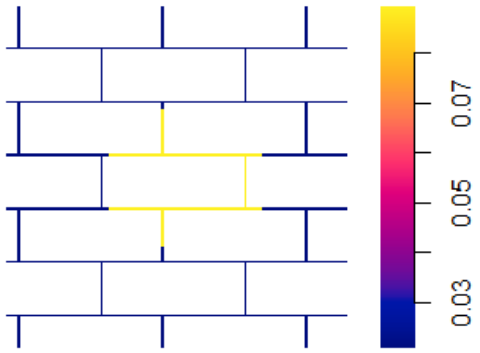
Tabla 5.5: proporción de rechazos bajo la hipótesis nula a distintos niveles de significación, $\alpha \in \{0.1,\ 0.05,\ 0.01\}$, para el test de Cramer von Mises en el subgrafo de *Spiders*, con modelos homogéneo e inhomogéneo y tamaños muestrales en media $m = 20,\ 50,\ 100,\ 200,\ 500$ y 750.

	Modelo homogéneo			Modelo inhomogéneo		
	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
$m = 20$	0.086	0.042	0.005	0.089	0.045	0.007
$m = 50$	0.091	0.040	0.006	0.088	0.048	0.003
$m = 100$	0.108	0.048	0.003	0.084	0.041	0.011
$m = 200$	0.099	0.039	0.013	0.100	0.048	0.010
$m = 500$	0.106	0.040	0.005	0.102	0.037	0.003
$m = 750$	0.089	0.048	0.008	0.098	0.048	0.008

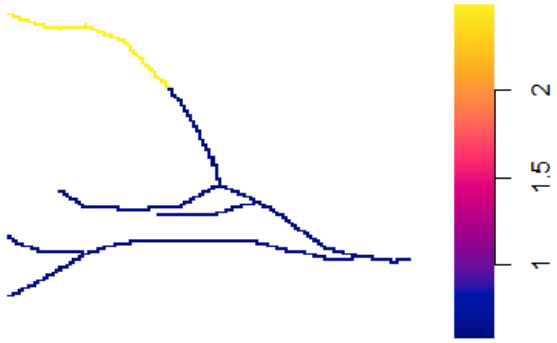
Tabla 5.6: proporción de rechazos bajo la hipótesis nula a distintos niveles de significación, $\alpha \in \{0.1,\ 0.05,\ 0.01\}$, para el test de Cramer von Mises en el subgrafo de *Dendrite*, con modelos homogéneo e inhomogéneo y tamaños muestrales en media $m = 20,\ 50,\ 100,\ 200,\ 500$ y 750.

Estas elecciones tienen como objetivo que $|\mathcal{L}_{\mathcal{G}}^{+}|/|\mathcal{L}_{\mathcal{G}}|$ sea aproximadamente el mismo en todos los casos. Atendiendo a estas definiciones, las funciones de intensidad bajo la hipótesis alternativa pueden verse representadas en la Figura 5.8 (caso homogéneo) y en la Figura 5.9 (escenario inhomogéneo).

En la Tabla 5.7 y en la Tabla 5.8 se encuentran los resultados del estudio de la potencia del contraste de Cramer von Mises, en los escenarios homogéneo e inhomogéneo respectivamente. Lo primero que debemos notar es que en este caso hemos tomado los mismos valores de a tanto para el subgrafo de *Spiders* como para el de *Dendrite*, ya que se alcanzan valores de potencia muy altos en ambos soportes para los mismos valores de a.

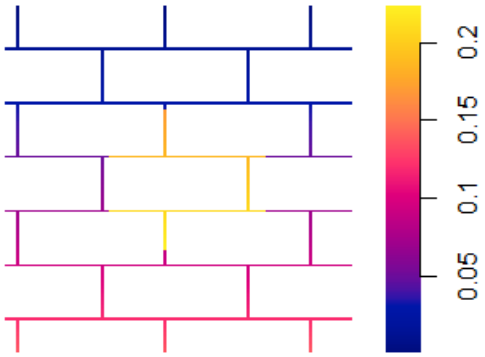


(a) Subgrafo de *Spiders*.

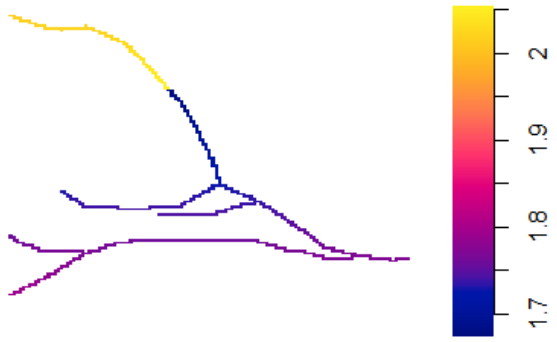


(b) Subgrafo de *Dendrite*.

Figura 5.8: representación de la función de intensidad homogénea perturbada de acuerdo a la ecuación (5.3) en los subgrafos de *Spiders* (a) y *Dendrite* (b) para el caso $m = 100$ y $a = 50$.



(a) Subgrafo de *Spiders*.



(b) Subgrafo de *Dendrite*.

Figura 5.9: representación de la función de intensidad inhomogénea perturbada de acuerdo a la ecuación (5.3) en los subgrafos de *Spiders* (a) y *Dendrite* (b) para el caso $m = 300$. En (a) hemos empleado $a = 100$ y en (b) $a = 10$.

	Subgrafo de *Spiders*			Subgrafo de *Dendrite*		
	$a = 3m/4$	$a = m/2$	$a = m/4$	$a = 3m/4$	$a = m/2$	$a = m/4$
$m = 50$	0.873	0.557	0.168	0.977	0.843	0.377
$m = 100$	0.999	0.927	0.378	1	0.986	0.669
$m = 200$	1	1	0.790	1	1	0.938
$m = 500$	1	1	0.994	1	1	1

Tabla 5.7: proporción de rechazos bajo la hipótesis alternativa al nivel de significación $\alpha = 0.05$ para el test de Cramer von Mises en los subgrafos *Spiders* y *Dendrite*, en el caso homogéneo ($\lambda_1 = \lambda_{Hom}$), para diferentes tamaños muestrales esperados del primer patrón de puntos ($m = 50,\ 100,\ 200$ y 500) y diferentes valores del número de puntos añadidos de forma esperada homogéneamente en $\mathcal{L}_{\mathcal{G}}^{+}$ en el segundo patrón de puntos a.

	Subgrafo de *Spiders*			Subgrafo de *Dendrite*		
	$a = 3m/4$	$a = m/2$	$a = m/4$	$a = 3m/4$	$a = m/2$	$a = m/4$
$m = 50$	0.905	0.612	0.184	0.985	0.848	0.394
$m = 100$	1	0.926	0.414	1	0.992	0.680
$m = 200$	1	0.999	0.738	1	1	0.994
$m = 500$	1	1	0.999	1	1	1

Tabla 5.8: proporción de rechazos bajo la hipótesis alternativa al nivel de significación $\alpha = 0.05$ para el test de Cramer von Mises en los subgrafos *Spiders* y *Dendrite*, en el caso inhomogéneo ($\lambda_1 = \lambda_{Inh}$), para diferentes tamaños muestrales esperados del primer patrón de puntos ($m =$ 50, 100, 200 y 500) y diferentes valores del número de puntos añadidos de forma esperada homogéneamente en $\mathcal{L}_{\mathcal{G}}^{+}$ en el segundo patrón de puntos a.

En ambas tablas vemos como para cada valor de m la potencia aumenta con a. Esto era de esperar, ya que nos alejamos de la hipótesis nula. Nótese además como, fijado a en función de m, según aumentamos también m aumenta la potencia, ya que tenemos más información para discernir si las funciones de intensidad de los procesos puntuales que generan los patrones observados son o no proporcionales. Estos comportamientos esperables se observan en los dos grafos que hemos estudiado. Por todo ello, podemos concluir que los resultados del estudio de la potencia para el test de Cramer von mises son satisfactorios.

Capítulo 6

Aplicación a datos reales

Los accidentes de carretera se han convertido, en el último medio siglo, en un serio problema de seguridad ciudadana a lo largo de todo el mundo: Brasil [1], India [18], Irán [15], Korea [29], etc. El estudio de la distribución de accidentes de tráfico a lo largo de una red de carreteras, analizando los puntos de mayor acumulación de accidentes (puntos negros), o cómo su distribución se ve afectada por factores externos, resulta una labor esencial de cara a mejorar la seguridad vial, y en esencia, salvar vidas. En concreto, el ayuntamiento de Río de Janeiro ha puesto en marcha un plan dedicado a mejorar la seguridad en la circulación por las carreteras de la ciudad. A raíz de esto, hemos conseguido un conjunto de datos reales sobre los que vamos a ilustrar las técnicas desarrolladas en el Capítulo 4[1].

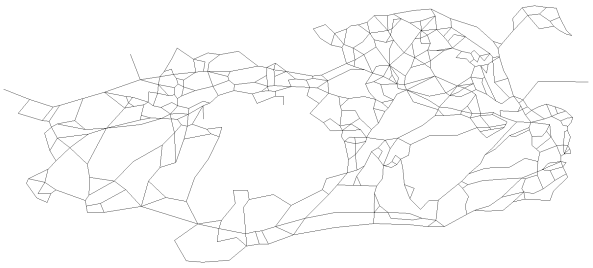
La base de datos cuenta con un total de 270908 entradas, cada una de las cuales corresponde con un accidente de tráfico en una carretera de Río de Janeiro entre el 21 de marzo de 2019 y el 4 de mayo de 2022. Estos eventos son reportados por conductores/transeúntes a través de la plataforma Waze[2]. Para cada evento se conocen: sus coordenadas geográficas, la fecha y hora en la que el accidente fue reportado, el tipo de vía en el que tuvo lugar, su dirección postal, e indicadores de la calidad de la medida asociados a posteriores notificaciones de que el accidente se ha reportado de forma correcta.

Para poder estudiar la distribución de los accidentes de tráfico a lo largo de la red de carreteras de Río de Janeiro, necesitamos un grafo lineal que la describa. Para que este grafo sea computacionalmente manejable, optamos por considerar únicamente las carreteras principales de la ciudad, obviando vías secundarias. Así, para la construcción del grafo emplearemos como referencia el mapa de carreteras de Río de Janeiro que puede verse en la Figura 6.1 (a). Toma-

[1]Agradecemos al profesor Rodrigo S. Targino de la Fundación Getulio Vargas por habernos proporcionado esta base de datos.

[2]La página principal de esta plataforma puede verse aquí.

(a) Mapa de las principales carreteras de Río, extraído de [17].



(b) Grafo lineal representando la red de carreteras principales de Río de Janeiro.

Figura 6.1: mapa de las principales carreteras de Río de Janeiro (a), y grafo lineal representativo, construido a partir del anterior (b).

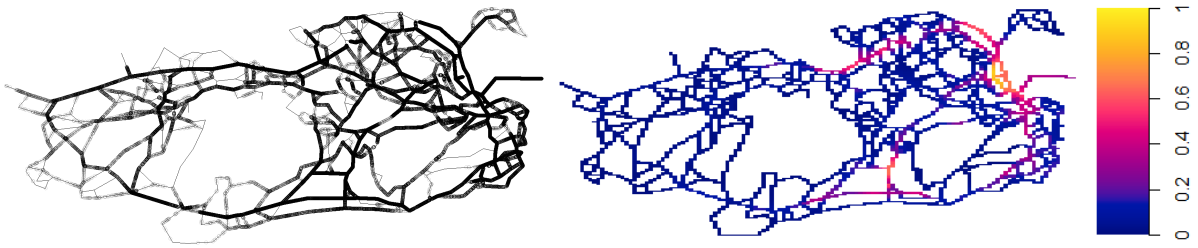
(a) Conjunto de datos original.

(b) Accidentes ocurridos en vías principales.

Figura 6.2: representación sobre el grafo lineal de las coordenadas geográficas de los accidentes de carretera recogidos en la base de datos (a), y de aquellos que tuvieron lugar en vías principales (b).

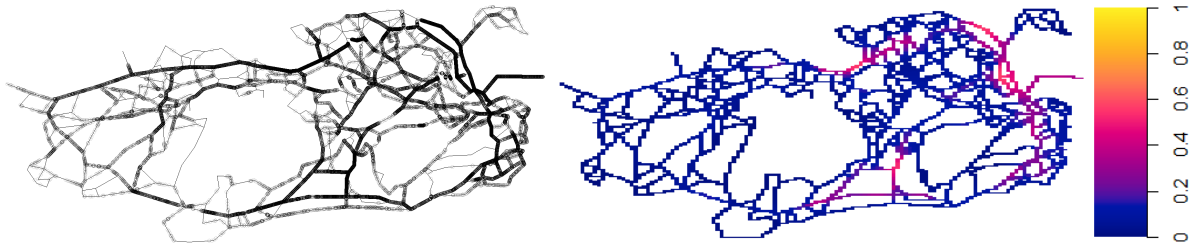
mos como conjunto de nodos los cruces de las carreteras de este mapa, y consideramos luego las aristas que se identifican con las carreteras que unen dichos nodos. Obtenemos así el grafo lineal que vemos en la Figura 6.1 (b), que posee 534 nodos y 806 aristas. Notemos que este no deja de ser una aproximación, ya que además de considerar solo las vías principales, aproxima tramos curvos por segmentos rectos.

Ahora bien, a pesar de que nosotros hemos considerado únicamente para la construcción de nuestro grafo lineal las carreteras principales de Río de Janeiro, la base de datos de la que disponemos contiene también accidentes en vías secundarias, como puede apreciarse en la Figura 6.2 (a). Como nuestra base de datos incluye una variable categórica que nos indica el tipo de vía en el que se reportó el accidente, de ahora en adelante nos restringiremos a aquellos accidentes que ocurran sobre las carreteras principales: 241804 que eventos pueden verse representados en la Figura 6.2 (b).



(a) Patrón de puntos. (b) Estimación de la densidad.

Figura 6.3: representación del patrón de puntos de los accidentes ocurridos en vías principales de lunes a viernes (a), junto con una estimación de su función de densidad asociada (b), obtenida empleando la ecuación del calor escogiendo el ancho de banda con la regla de Scott modificada.

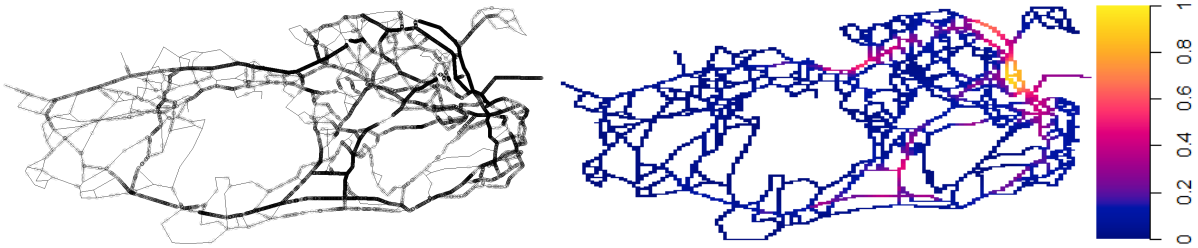


(a) Patrón de puntos. (b) Estimación de la densidad.

Figura 6.4: representación del patrón de puntos de los accidentes ocurridos en vías principales los sábados y domingos (a), junto con una estimación de su función de densidad (b), obtenida empleando la ecuación del calor escogiendo el ancho de banda con la regla de Scott modificada.

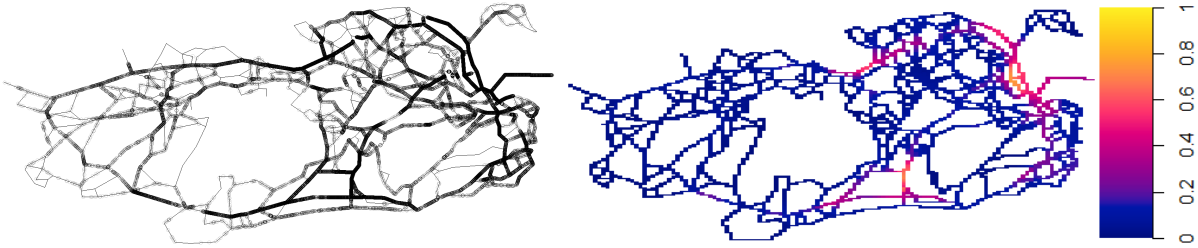
El primer problema de interés es comparar la distribución espacial de los accidentes que ocurren durante días laborables con los que ocurren en fin de semana. Para ello, extraemos los accidentes de lunes a viernes (189996 eventos) y los ocurridos a lo largo del fin de semana (51681 eventos). Estos dos patrones, junto con estimaciones no paramétricas de su función de densidad, pueden verse en la Figura 6.3 y en la Figura 6.4. Para este par de patrones de puntos calcularemos el nivel crítico asociado al contraste tanto de Kolmogorov-Smirnov como de Cramer von Mises, empleando $n_p = 10000$ permutaciones. Para el cómputo del estadístico de Kolmogorov-Smirnov se ha tomado como punto base del π-sistema en el grafo lineal de las carreteras principales de Río de Janeiro el punto que hemos representado en la Figura 6.7.

En ambos casos la estimación del nivel crítico ha sido de $5 \cdot 10^{-5}$; es decir, tenemos evidencias significativas en contra de la hipótesis nula. Podemos por ello concluir que existen evidencias significativas a favor de que la estructura espacial de los accidentes de tráfico en las carreteras principales de Río de Janeiro cambia de los días de semana a los fines de semana.



(a) Patrón de puntos. (b) Estimación de la densidad

Figura 6.5: representación del patrón de puntos de los accidentes ocurridos en vías principales de 10 a 13 horas (a), junto con una estimación de su función de densidad asociada (b), obtenida empleando la ecuación del calor escogiendo el ancho de banda con la regla de Scott modificada.



(a) Patrón de puntos. (b) Estimación de la densidad

Figura 6.6: representación del patrón de puntos de los accidentes ocurridos en vías principales de 20 a 23 horas (a), junto con una estimación de su función de densidad (b), obtenida empleando la ecuación del calor escogiendo el ancho de banda con la regla de Scott modificada.

El segundo problema que consideraremos será comparar la distribución espacial de los accidentes que ocurren en los dos tramos de hora punta. Extraemos para ello los accidentes que tienen lugar de 10 a 13 horas (45311 eventos) y de 20 a 23 horas (57159 eventos). Estos patrones de puntos se encuentran representados en la Figura 6.5 y en la Figura 6.6, respectivamente, junto con estimaciones no paramétricas de sus funciones de densidad asociadas. Nuevamente, para estos dos patrones de puntos estimamos el nivel crítico para los contrastes de Kolmogorov-Smirnov (empleando de nuevo como punto base el representado en la Figura 6.7) y de Cramer von Mises, empleando $n_p = 10000$ permutaciones.

De nuevo, ambas estimaciones arrojaron un p-valor de $5 \cdot 10^{-5}$. Teniendo evidencias en contra de la hipótesis nula, podemos concluir que existen evidencias significativas a favor de que la estructura espacial de los accidentes de tráfico en las carreteras principales de Río de Janeiro cambia de la franja horaria de 10 a 13 a la franja horaria de 20 a 23.

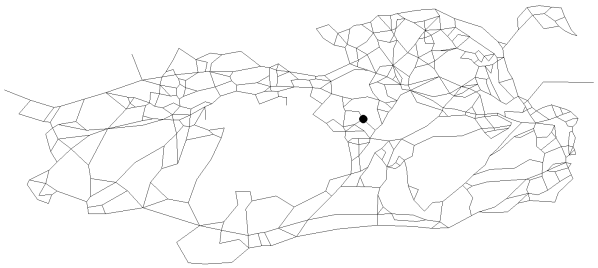


Figura 6.7: representación del punto base del π-sistema del grafo de las carreteras principales de Río de Janeiro empleado para el contraste de Kolmogorov-Smirnov. Este punto se ha calculado empleando el código descrito para tal efecto en la Sección A.1.1.

Los resultados obtenidos pueden resultar sorprendentes, ya que a primera vista no se observan grandes diferencias en las estimaciones de la densidad de las Figura 6.3 y la Figura 6.4, y tampoco entre las de la Figura 6.5 y la Figura 6.6. Ahora bien, las conclusiones obtenidas parecen tener sentido si pensamos, por ejemplo, que de lunes a viernes los accidentes de tráfico ocurrirán en mayor medida en las zonas de mayor actividad laboral, mientras que los fines de semana los eventos se concentrarán en mayor medida en las zonas de ocio, cuyo emplazamiento es diferente en la ciudad de Río de Janeiro. Un comportamiento similar es de esperar también de la estructura espacial de los accidentes en la red de carreteras cuando se comparan las dos horas punta.

Es importante notar que debido a los tamaños muestrales tan elevados que se manejan en este análisis, cualquier mínima variación entre los patrones puede ser detectada como significativa y derivar en un rechazo de la hipótesis nula. Sería interesante, y será objeto de trabajo futuro, resolver problemas en rangos de tiempo más específicos que den lugar a comparaciones más razonables.

Anexo A

Código

En este apéndice presentamos el código que hemos empleado en nuestro estudio de simulación.

A.1. Cálculo de los estadísticos

A.1.1. Test de Kolmogorov-Smirnov

Comenzamos incluyendo el código de la función que permite el cálculo del estadístico de Kolmogorov-Smirnov:

```
1  library(spatstat)
2  KStest=function(pp1,pp2,q){
3          #pp1 y pp2 son los patrones de puntos, y q es el punto base del
           pi-sistema considerado. Los tres han de pasarse como objetos lpp en
           el mismo grafo.
4          N1=npoints(pp1)
5          N2=npoints(pp2)
6          #estimamos omega globalmente
7          eta=N1/N2
8          #lo primero que vamos a hacer es calcular la estimación de xi.
9          #necesitamos saber cuantas aristas hay en el grafo
10         #y el número de puntos que hay en cada arista de cada patron de
           puntos
11         K=as.psp(domain(pp1))$n
```

```
12          N1L=rep(0,K)
13          X1=coords(pp1)[,3]
14          for(i in 1:length(X1)){N1L[X1[i]]=N1L[X1[i]]+1}
15          N2L=rep(0,K)
16          X2=coords(pp2)[,3]
17          for(i in 1:length(X2)){N2L[X2[i]]=N2L[X2[i]]+1}
18          #tenemos que determinar entonces en qué aristas se observa almenos
            un punto, y restringirnos a ellas
19          aristas=which(N1L+N2L>0)
20          #lo qque hacemos ahora entonces es restrinrgirnos a estas aristas
21          N1L=N1L[aristas]
22          N2L=N2L[aristas]
23          N1Lhat=(eta/(1+eta))*(N1L+N2L)
24          N2Lhat=N1Lhat/(eta)
25          #con esto ya podemos estimar xi:
26          xi=sum((N1L-N1Lhat)^2/N1Lhat+(N2L-N2Lhat)^2/N2Lhat)
27          xi=sqrt(xi/(length(N1L)-1))
28          #podemos empezar a construir nuestro estadístico
29          Dhat=sqrt(N1*N2/(N1+N2))/xi
30          #únicamente nos falta calcular el término que se expresa como un
            supremo, que en realidad es un máximo al ser finito el conjunto.
31          sup=0
32          #creamos el vector de distancias
33          dist=c(crossdist(q,pp1),crossdist(q,pp2))
34          #y creamos el vector indicador
35          ind=c(rep(1,N1),rep(2,N2))
36          #ahora lo que tenemos es que ordenar el vector de distancias de
            menor a mayor, y de la misma forma ordenar el vector de indicadores
37          y=cbind(dist,ind)
38          ind=y[order(y[,1]),][,2]
39          #con esto ya podemos calcular el supremo
40          N=c(0,0)
41          for (i in 1:(N1+N2)){
42                  N[ind[i]]=N[ind[i]]+1
43          sup=max(sup, abs( N[1]/N1-N[2]/N2))
44          }
45          Dhat=Dhat*sup
46          return(Dhat)
```

```
47 }
48 #EJEMPLO:
49 KStest(spiders[seq(1,48,2)], spiders[seq(2,48,2)],
   lpp(c(562.5,562.5),domain(spiders)))
```

Recordemos que la elección del punto base del π-sistema se hacía de forma distinta en función de si nuestro grafo es o no un árbol. Si es un árbol, la elección del punto base es clara como el nodo raíz. En el caso de que nuestro grafo tenga ciclos, una posible elección del punto base es el punto más cercano (en la distancia euclídea) al centroide de los nodos del grafo. En nuestro estudio de simulación, el punto base en el caso de que nuestro grafo tenga ciclos se ha calculado empleando la siguiente función:

```
1  library(rgeos)
2  library(sp)
3  library(spatstat)
4  CentroGrafo=function(graph){
5          #en el caso de pasar el patrón de puntos en vez del grafo permitimo
           que funcione también
6          if(sum(class(graph)=="lpp")==1){graph=domain(graph)}
7          #tomamos las coordenadas espaciales de los vértices del grafo en el
           formato apropiado para calcular el centroide
8          Nodos=SpatialPoints(coords(graph$vertices))
9          #y calculamos el centroide como un vector de coordenadas
10         cent=as.numeric(as.data.frame(gCentroid(Nodos)))
11         #una vez que tenemos esto, ya podemos hacer simplemente:
12         return(lpp(cent,graph))
13         #ya que al definir esto, la función lpp proyecta el punto cent
           ortogonalmente
14         a la arista más cercana de graph, usando la función
           project2segment. Esto permite que el punto q sea el punto del grafo
           (no necesariamente el vértice) más cercano al centroide de los
           nodos
15 }
16 #EJEMPLO
17 CentroGrafo(spiders)
18 plot(CentroGrafo(spiders))
```

A.1.2. Test de Cramer von Mises

Para calcular el estadístico de Cramer von Mises empleamos la siguiente función:

```
1  library(spatstat)
2  CvMtest=function(pp1,pp2,h,Heat=F,cont=F,ker="epa"){
3          #Recordemos que el test de Cramer von Mises calculará estimaciones
           de las funciones de densidad asociadas a los dos procesos puntuales
           a partir de los patrones de puntos observados, y calculará la
           distancia L2 al cuadrado entre ellas.
4          #pp1 y pp2 son los patrones de puntos en el mismo grafo.
5          #h es el ancho de banda para la estimación no paramétrica de la
           función de intensidad. Se admite que se pase un vector con dos
           entradas, con el ancho de banda para la estimación de la densidad
           de cada proceso.
6          #Heat es una variable binaria que nos indica si la estimación de la
           función de intensidad se hace (Heat=T) o no (Heat=F) empleando la
           estimación basada en la ecuación del calor. En caso de que Heat=F,
           la estimación se hace empleando núcleos equitativos.
7          #cont es una variable binaria. En el caso de que Heat=F, la
           estimación de la función de intensidad se hace empelando núcleos
           equitativos discontinuos si cont=F y continuos si cont=T
8          #ker indica la función núcleo básica con la que se construyen los
           núcleos equitativos en el caso de que Heat=F. Por defecto se toma
           el núcleo de Epanechnikov.
9          #notemos como por defecto la estimación de las funciones de
           intensidad se hace empleando núcleos equitativos discontinuos
           tomando como función núcleo básica el núcleo de Epanechnikov.
10         #si no se especifica ancho de banda, este se toma para cada proceso
           puntual empleando la regla de Scott modificada
11         if (missing(h)){
12                 h=c(0,0)
13                 h[1]=sqrt(sum(diag(cov(coords(as.ppp(pp1))))))
14                 h[1]=h[1]*(3*npoints(pp1))^(-1/5)
15                 h[2]=sqrt(sum(diag(cov(coords(as.ppp(pp2))))))
16                 h[2]=h[2]*(3*npoints(pp2))^(-1/5)
17         }
```

```
18          #si se pasa un único ancho de banda, se entiende que es el mismo
            para las dos estimaciones
19          if (length(h)==1){h=c(h,h)}
20          #calculamos el número de puntos de caba patrón observado
21          N1=npoints(pp1)
22          N2=npoints(pp2)
23          #y estimamos entonces cada una de las funciones de intensidad,
            distinguiendo el tipo de estimación
24          if (Heat==F){
25                  lambda1=densityEqualSplit(pp1, sigma=h[1], continuous =
                    cont, kernel=ker, verbose=F, savehistory = F)
26                  lambda2=densityEqualSplit(pp2, sigma=h[2], continuous =
                    cont, kernel=ker, verbose=F, savehistory = F)
27          }else{
28                  lambda1=densityHeat.lpp(pp1, sigma=h[1], verbose=F, iterMax
                    = 1e+12)
29                  lambda2=densityHeat.lpp(pp2, sigma=h[2], verbose=F, iterMax
                    = 1e+12)
30          }
31          #hecho esto, podemos calcular el estadístico de contraste
            expandiendo el cuadrado de la diferencia
32          That=integral.linim(lambda1^2)/N1^2
33          That=That+integral.linim(lambda2^2)/N2^2
34          That=That-2*integral.linim(lambda1*lambda2)/(N1*N2)
35          return(That)
36  }
37  #EJEMPLO
38  #empleando núcleos equitativos discontinuos tomando como función núcleo
    básica el núcleo de Epanechnikov
39  CvMtest(spiders[seq(1,48,2)], spiders[seq(2,48,2)])
40  #o empleando la ecuación del calor
41  CvMtest(spiders[seq(1,48,2)], spiders[seq(2,48,2)], Heat = T)
```

A.1.3. Test no paramétrico basado en la función de riesgo relativo

Finalmente, para el cálculo del estadístico del test no paramétrico basado en la función de riesgo relativo emplearemos la siguiente función:

```
1  library(spatstat)
2  NPtest=function(pp1,pp2,h,Heat=F,cont=F,ker="epa",CV=F){
3          #Recordemos que el test no paramétrico calculará estimaciones de la
           función de riesgo relativo en los puntos de ambos patrones
           observados, y luego planteará un test de efecto en la regresión de
           estos valores estimados frente a las posiciones de los puntos en el
           grafo, para lo que se realizará una regresión no paramétyrica
           empleando una generalización a grafos lineales del estimador de
           Nadaraya-Watson.
4          #pp1 y pp2 son los patrones de puntos en el mismo grafo.
5          #h es el ancho de banda para la estimación no paramétrica de la
           función de intensidad y para la construcción de la matriz de
           suavizado. Se admite que se pase un vector con tres entradas, las
           dos primeras con el ancho de banda para la estimación de la
           densidad de cada proceso, y la tercera para el cómputo de la matriz
           de suavizado
6          #Heat es una variable binaria que nos indica si la estimación de la
           función de intensidad y los núcleos empleando en la regresión no
           paramétrica se hace (Heat=T) o no (Heat=F) empleando la estimación
           basada en la ecuación del calor. En caso de que Heat=F, se
           consideran núcleos equitativos.
7          #cont es una variable binaria. En el caso de que Heat=F, se
           emplearán núcleos equitativos discontinuos si cont=F y continuos si
           cont=T
8          #ker indica la función núcleo básica con la que se construyen los
           núcleos equitativos en el caso de que Heat=F. Por defecto se toma
           el núcleo de Epanechnikov.
9          #CV es una variable binaria que nos indica si la estimación de la
           función de riesgo relativo en los puntos observados se hace
           empleando validación cruzada (CV=T) o no (CV=F)
10         #notemos como por defecto la estimación de las funciones de
           intensidad y el cálculo de la matriz de suavizado se hace empleando
           núcleos equitativos discontinuos tomando como función núcleo básica
           el núcleo de Epanechnikov, y la estimación de la función de riesgo
           relativo se hace sin emplear validación cruzada.
11         #si no se especifica ancho de banda, este se toma para cada proceso
           puntual empleando la regla de Scott modificada, y para el caso de
           la matriz de suavizado se consideran los dos patrones superpuestos
```

```
12          if (missing(h)){
13                  h=numeric(3)
14                  h[1]=sqrt(sum(diag(cov(coords(as.ppp(pp1))))))
15                  h[1]=h[1]*(3*npoints(pp1))^(-1/5)
16                  h[2]=sqrt(sum(diag(cov(coords(as.ppp(pp2))))))
17                  h[2]=h[2]*(3*npoints(pp2))^(-1/5)
18                  #teniendo en cuenta que h[3] lo vamos a usar para estimar
                    la intensidad empleando tanto puntos de
19                  #ambos patrones, parece una elección natural usar el ancho
                    de banda dado por la regla se scott superponiendo ambos
                    patrones de puntos:
20                  pp3=superimpose(pp1,pp2)
21                  h[3]=sqrt(sum(diag(cov(coords(as.ppp(pp3))))))
22                  h[3]=h[3]*(3*npoints(pp3))^(-1/5)
23          }
24          # si se especifica un único ancho de banda, se entiende el mismo
            para todos los casos
25          if (length(h)==1){h=c(1,1,1)*h}
26          N1=npoints(pp1)
27          N2=npoints(pp2)
28          #recordemos que las estimaciones de las funciones de intensidad de
            los procesos puntuales no pueden anularse en ninguno de los puntos
            observados. Como no hemos impuesto ninguna condición sobre el ancho
            de banda para garantizar esto, lo que haremos será imponer un
            treshold a las estimaciones de la intensidad, tal que estas nunca
            sean menores que el valor de la estimación homogénea de la
            intensidad del proceso en cuestión. Estimamos la intensidad de
            forma homogénea como:
29          I1=N1/volume(domain(pp1))
30          I2=N2/volume(domain(pp2))
31          #calculamos entonces las estimaciones de la intensidad de cada
            proceso puntual en los puntos. Tenemos que distinguir en función de
            los valores de Heat y de CV
32          if (Heat==F){
33                  #estimamos las funciones de intensidad empleando la
                    ecuación del calor. Para simplificar su evaluación, las
                    tomamos como objetos linfun
```

```
34              lambda1=as.linfun(densityEqualSplit(pp1, sigma=h[1],
                continuous = cont, kernel=ker, verbose=F, savehistory = F))
35              lambda2=as.linfun(densityEqualSplit(pp2, sigma=h[2],
                continuous = cont, kernel=ker, verbose=F, savehistory = F))
36              if (CV==F){
37                      #de no requerir validación cruzada evaluamos
                        directamente
38                      l1=c(lambda1(coords(pp1)),lambda1(coords(pp2)))
39                      l2=c(lambda2(coords(pp1)),lambda2(coords(pp2)))
40                      #en aquellos puntos en los que se estime una
                        intensidad menor que el estimador homogéneo,
                        sustituimos la estimacuón por la homogénea
41                      l1[l1<I1]=I1
42                      l2[l2<I2]=I2
43                      #y estimamos entonces el logaritmo de la función de
                        riesgo relativo
44                      rho=log((N2/N1)*(l1/l2))
45              }else{
46                      #de querer emplear validación cruzada, las
                        estimaciones que tenemos nos valen para el otro
                        patrón de puntos, pero para el respectivo tenemos
                        que calcularlas empleando validación cruzada:
47                      l11=densityEqualSplit(pp1, sigma=h[1], continuous =
                        cont, kernel=ker, verbose=F, savehistory = F,
                        at="points", leaveoneout = T)
48                      attr(l11,"sigma")=NULL
49                      l22=densityEqualSplit(pp2, sigma=h[2], continuous =
                        cont, kernel=ker, verbose=F, savehistory = F,
                        at="points", leaveoneout = T)
50                      attr(l22,"sigma")=NULL
51                      l12=lambda1(coords(pp2))
52                      l21=lambda2(coords(pp1))
53                      #nuevamente imponemos el treshold de la intensidad
                        homogénea
54                      l11[l11<I1]=I1
55                      l12[l12<I1]=I1
56                      l21[l21<I2]=I2
57                      l22[l22<I2]=I2
```

```
58                          #y estimamos
59                          rho=c(log((N2/(N1-1))*(l11/l21)),\n
                            log(((N2-1)/N1)*(l12/l22)))
60                  }
61          } else{
62                  #si Heat=F, todo es análogo cambiando la forma en la que se
                    estima la intensidad:
63                  lambda1=as.linfun(densityHeat.lpp(pp1,sigma=h[1],verbose=F))
64                  lambda2=as.linfun(densityHeat.lpp(pp2,sigma=h[2],verbose=F))
65                  if (CV==F){
66                          l1=c(lambda1(coords(pp1)),lambda1(coords(pp2)))
67                          l2=c(lambda2(coords(pp1)),lambda2(coords(pp2)))
68                          l1[l1<I1]=I1
69                          l2[l2<I2]=I2
70                          rho=log((N2/N1)*(l1/l2))
71                  }else{
72                          l11=densityHeat.lpp(pp1, sigma=h[1], verbose=F,
                            at="points", leaveoneout = T)
73                          attr(l11,"sigma")=NULL
74                          l12=lambda1(coords(pp2))
75                          l21=lambda2(coords(pp1))
76                          l22=densityHeat.lpp(pp2, sigma=h[2], verbose=F,
                            at="points", leaveoneout = T)
77                          attr(l22,"sigma")=NULL
78                          l11[l11<I1]=I1
79                          l12[l12<I1]=I1
80                          l21[l21<I2]=I2
81                          l22[l22<I2]=I2
82                          rho=c(log((N2/(N1-1))*(l11/l21)),
83                                  log(((N2-1)/N1)*(l12/l22)))
84                  }
85          }
86          #una vez estimada la función de riegso relativo, llega el momento
            de efectuar la regresión. Bajo la hipótesis nula del test de efecto
87          mu=mean(rho)
88          #bajo la hipótesis alternativa del test de efecto, estimamos la
            función de regresión a través de la matriz de suavizado S
89          S=matrix(0,nrow=N1+N2,ncol=N1+N2)
```

```
90          #ahora calculamos cada elemento, que es S_ij=G_{p_j}(p_i). Tenemos
            que distinguir como queremos calcular estos núcleos.
91          if (Heat==F){
92                  #empezamos con los puntos de pp1
93                  for (j in 1:N1){
94                          #calculamos G_{p_j} como una linfun
95                          Gj=as.linfun(densityEqualSplit(pp1[j],sigma=h[3],
96                          continuous = cont,kernel = ker,verbose =
                            F,savehistory = F))
97                          #y la evaluamos en los puntos de los patrones
                            observados
98                          S[,j]=c(Gj(coords(pp1)),Gj(coords(pp2)))
99                  }
100                 #y ahora hacemos lo mismo basando en los puntos de pp2
101                 for (j in 1:N2){
102                         #calculamos G_{p_j} como una linfun
103                         Gj=as.linfun(densityEqualSplit(pp2[j],sigma=h[3],
104                         continuous = cont,kernel = ker,verbose =
                            F,savehistory = F))
105                         S[,j+N1]=c(Gj(coords(pp1)),Gj(coords(pp2)))
106                 }
107         }else{
108                 #prodecemos de fora análoga pero estimando mediante la
                    ecuación del calor
109                 for (j in 1:N1){
110                         Gj=as.linfun(densityHeat.lpp(pp1[j], sigma=h[3],
                            verbose = F))
111                         S[,j]=c(Gj(coords(pp1)),Gj(coords(pp2)))
112                 }
113                 for (j in 1:N2){
114                         Gj=as.linfun(densityHeat.lpp(pp2[j], sigma=h[3],
                            verbose = F))
115                         S[,j+N1]=c(Gj(coords(pp1)),Gj(coords(pp2)))
116                 }
117         }
118         #ahora que ya tenemos la matriz S, debemos dividir cada elemento de
            S entre la suma de los elementos de su fila
119         S=sweep(S,1,apply(S,1,sum),"/")
```

```
120         #una vez ya tenemos S podemos calcular los residuos cuadráticos de
            cada modelo de regresión bajo la hipótesis nula y alternativa.
121         #bajo la hipótesis nula
122         RSS0=sum((rho-mu)^2)
123         #bajo la hipótesis alternativa
124         RSSa=sum((rho-S%*%rho)^2)
125         #los grados de libertad bajo la hipótesis nula son
126         df0=N1+N2-1
127         #ya que se estima un único parámetro. Bajo la hipótesis
            alternativa, en analogía con el modelo lineal general
128         dfa=N1+N2-sum(diag(S))
129         #Con todo esto ya podemos calcular el estadístico
130         Fhat=((RSS0-RSSa)/(df0-dfa))/(RSSa/dfa)
131         return(Fhat)
132 }
133 #EJEMPLOS
134 #empleando la ecuación del calor
135 NPtest(spiders[seq(1,48,2)], spiders[seq(2,48,2)], Heat = T)
136 #empleando la ecuación del calor, y estimando el log-riesgo relativo
    mediante validación cruzada
137 NPtest(spiders[seq(1,48,2)], spiders[seq(2,48,2)], Heat = T, CV = T)
```

A.2. Estimación de niveles críticos empleando el test de permutaciones

Habiendo especificado ya el código que permite calcular los tres estadísticos de contraste que hemos propuesto en el Capítulo 4, debemos especificar el código empleado para nuestro estudio de simulación. Primeramente debemos introducir las funciones que ejecutan el test de permutaciones para estimar el nivel crítico asociado al valor de un estadístico de contraste.

A.2.1. Test de Kolmogorov-Smirnov

Para el estadístico del test de Kolmogorov-Smirnov hemos empleado la siguiente función:

```
1 library(spatstat)
2 PermutationsKS=function(pp1,pp2,q,nboots){
3         #pp1 y pp2 son los patrones de puntos observados
```

```
4       #q es el punto base del pi-sistema que se desea emplear en el
        cálculo del estadístico de Kolmogorov-Smirnov
5       #nboots es el número de permutaciones que se van a considerar para
        estimar el estadístico.
6       t =KStest(pp1,pp2,q)
7       na = npoints(pp1)
8       nb = npoints(pp2)
9       n = nb + na
10      #queremos ahora combinar los dos patrones de puntos en uno solo
11      comb = superimpose(pp1,pp2)
12      #en el caso de que nboots no se haya pasado como entero
13      nboots = as.integer(nboots)
14      reps = bigger = 0L
15      boot_t=numeric(nboots)
16      for(idx in 1:nboots){
17              #tomamos los subconjutnos de índices de forma aleatoria
18              e = sample.int(n, na, T)
19              f = sample.int(n, nb, T)
20              #calculamos el estadístico para las muestras permutadas
21              boot_t[idx] = KStest(comb[e], comb[f], q)
22              #si el valor del estadístico es mayor que el obtenido en
                las muestras originales, acumulamos
23              if (boot_t[idx] >= t){bigger = 1L + bigger }
24      }
25      #lo que vamos a devolver es el estadístico y  el p-valor, que
        calculamos como la fracción de estadísticos calculados a partir de
        muestras permutadas mayores que el de partida
26      out = c(t, bigger/nboots)
27      #en el caso de que nuestro p-valor estimado sea cero, admitimos que
        este condicide con la mitad de la resolución
28      if (out[2] == 0) {out[2] = 1/(2 * nboots)}
29      #damos alrgo de formato
30      details = c(na, n - na, nboots)
31      names(details) = c("n1", "n2", "n.boots")
32      attributes(out) = list(details = details)
33      names(out) = c("Test Stat", "P-Value")
34      #finalmente le damos formato a la salida
35      out2=list("Test Stat"=out[1],"Pvalue"=out[2])
```

```
36          return(out2)
37  }
38  #EJEMPLO
39  #EJEMPLO:
40  PermutationsKS(spiders[seq(1,48,2)], spiders[seq(2,48,2)],
    lpp(c(562.5,562.5), domain(spiders)), nboots = 5000L)
```

A.2.2. Test de Cramer von Mises

Para el test de Cramer von Mises y el test no paramétrico basado en la función de riesgo relativo las funciones que efectúan el test de permutaciones son análogas a la anterior, ya que únicamente hay que cambiar los parámetros de entrada y la llamada a las funciones que efectúan el cálculo del estadístico. Para el test de Cramer von Mises tenemos que:

```
1   library(spatstat)
2   PermutationsCvM=function(pp1,pp2,h,contin=F,kernel="epa",Heateq=F,nboots){
3           #pp1 y pp2 son los patrones de puntos observados
4           #q es el punto base del pi-sistema que se desea emplear en el
            cálculo del estadístico de Kolmogorov-Smirnov
5           #nboots es el número de permutaciones que se van a considerar para
            estimar el estadístico.
6           t =CvMtest(pp1,pp2,h,cont=contin,ker=kernel,Heat=Heateq)
7           na = npoints(pp1)
8           nb = npoints(pp2)
9           n = nb + na
10          #queremos ahora combinar los dos patrones de puntos en uno solo
11          comb = superimpose(pp1,pp2)
12          #en el caso de que nboots no se haya pasado como entero
13          nboots = as.integer(nboots)
14          reps = bigger = 0L
15          boot_t=numeric(nboots)
16          for(idx in 1:nboots){
17                  #tomamos los subconjutnos de índices de forma aleatoria
18                  e = sample.int(n, na, T)
19                  f = sample.int(n, nb, T)
20                  #calculamos el estadístico para las muestras permutadas
21                  boot_t[idx] = CvMtest(comb[e], comb[f], h, cont=contin,
                    ker=kernel,Heat=Heateq)
```

```
22                  #si el valor del estadístico es mayor que el obtenido en
                    las muestras originales, acumulamos
23                  if (boot_t[idx] >= t){bigger = 1L + bigger }
24          }
25          #lo que vamos a devolver es el estadístico y  el p-valor, que
            calculamos como la fracción de estadísticos calculados a partir de
            muestras permutadas mayores que el de partida
26          out = c(t, bigger/nboots)
27          #en el caso de que nuestro p-valor estimado sea cero, admitimos que
            este condicide con la mitad de la resolución
28          if (out[2] == 0) {out[2] = 1/(2 * nboots)}
29          #damos alrgo de formato
30          details = c(na, n - na, nboots)
31          names(details) = c("n1", "n2", "n.boots")
32          attributes(out) = list(details = details)
33          names(out) = c("Test Stat", "P-Value")
34          #finalmente le damos formato a la salida
35          out2=list("Test Stat"=out[1],"Pvalue"=out[2])
36          return(out2)
37  }
38  #EJEMPLO
39  PermutationsCvM(spiders[seq(1,48,2)], spiders[seq(2,48,2)], Heateq = T ,
    nboots = 5000L)
```

A.2.3. Test no paramétrico basado en la función de riesgo relativo

Finalmente para el test no paramétrico basado en la función de riesgo relativo tenemos:

```
1   library(spatstat)
2   PermutationsNP=function(pp1,pp2,h,con=F,k="epa",Heq=F,cross=F,nboots){
3           #pp1 y pp2 son los patrones de puntos observados
4           #q es el punto base del pi-sistema que se desea emplear en el
            cálculo del estadístico de Kolmogorov-Smirnov
5           #nboots es el número de permutaciones que se van a considerar para
            estimar el estadístico.
6           t =NPtest(pp1,pp2,h,cont=con,ker=k,Heat=Heq,CV=cross)
7           na = npoints(pp1)
8           nb = npoints(pp2)
```

```
9           n = nb + na
10          #queremos ahora combinar los dos patrones de puntos en uno solo
11          comb = superimpose(pp1,pp2)
12          #en el caso de que nboots no se haya pasado como entero
13          nboots = as.integer(nboots)
14          reps = bigger = 0L
15          boot_t=numeric(nboots)
16          for(idx in 1:nboots){
17                  #tomamos los subconjutnos de índices de forma aleatoria
18                  e = sample.int(n, na, T)
19                  f = sample.int(n, nb, T)
20                  #calculamos el estadístico para las muestras permutadas
21                  boot_t[idx] = NPtest(comb[e], comb[f], h, cont=con,
                    ker=k,Heat=Heq,CV=cross)
22                  #si el valor del estadístico es mayor que el obtenido en
                    las muestras originales, acumulamos
23                  if (boot_t[idx] >= t){bigger = 1L + bigger }
24          }
25          #lo que vamos a devolver es el estadístico y  el p-valor, que
            calculamos como la fracción de estadísticos calculados a partir de
            muestras permutadas mayores que el de partida
26          out = c(t, bigger/nboots)
27          #en el caso de que nuestro p-valor estimado sea cero, admitimos que
            este condicide con la mitad de la resolución
28          if (out[2] == 0) {out[2] = 1/(2 * nboots)}
29          #damos alrgo de formato
30          details = c(na, n - na, nboots)
31          names(details) = c("n1", "n2", "n.boots")
32          attributes(out) = list(details = details)
33          names(out) = c("Test Stat", "P-Value")
34          #finalmente le damos formato a la salida
35          out2=list("Test Stat"=out[1],"Pvalue"=out[2])
36          return(out2)
37  }
38  #EJEMPLO
39  PermutationsNP(spiders[seq(1,48,2)], spiders[seq(2,48,2)], Heq = T , nboots
    = 5000L)
```

A.3. Estudio de simulación

Una vez que hemos construido las funciones que nos permiten llevar a cabo el test de permutaciones y estimar niveles críticos asociados a nuestros estadísticos de contraste, podemos especificar el código empleado para llevar a cabo nuestro estudio de simulación.

A.3.1. Test de Kolmogorov-Smirnov

Comenzamos con el estudio del nivel para el test de Kolmogorov-Smirnov. Detallamos primeramente el caso homogéneo en *Spiders* con $m = 100$.

```
library(spatstat)
#comenzamos cargando el grafo. Para el estudio de simulación en Spiders tomamos
        L=domain(spiders)
#calculamos el centro del grafo
        q=CentroGrafo(L)
#comenzamos definiendo la intensidad.
#Tomamos
        lambda1=function(x,y){1+0*x}
#hecho esto, divimos la función anterior entre su integral para obtener la función de densidad asociada
lambda1.im=as.linim(lambda1,L)
m1=integral.linim(lambda1.im)
lambda10.im=lambda1.im/m1
#consideramos el tamaño muestral esperado
        m=100
#y definimos el vector de P-valores
Pvalues=numeric(1000)
#consideramos entonces las 1000 réplicas Monte Carlo
for (i in 1:1000){
        set.seed(2*i)
        pp1=rpoislpp(m*lambda10.im,L)
        pp2=rpoislpp(m*lambda10.im,L)
        #y para cada una de ellas calculamos el p-valor asociado al contraste de Kolmogorov-Smirnov.
        Pvalues[i]=PermutationsKS(pp1,pp2,q,nboots = 7500L)$Pvalue
}
```

Para obtener los niveles críticos para un distinto tamaño muestral, basta cambiar el valor de m en la línea 14. Para realizar el estudio de simulación tomando como soporte el grafo *Dendrite* hay que sustituir la línea 3 por `L=domain(dendrite)`. En este caso debemos cambiar también el punto base del π-sistema considerado. Para ello, en lugar de la línea 5 definimos el punto base como `q=lpp(c(243.4,300),domain(dendrite))`. En el caso de querer considerar el modelo con intensidad inhomogénea, se ha de sustituir la línea 8 del código anterior por `lambda1=function(x,y){1-(y/1125)^0.1}`.

Podemos proceder entonces con el estudio de la potencia del test de Kolmogorov-Smirnov. Presentamos a continuación el código para el estudio de la potencia en *Spiders*, en el caso homogéneo, con $m = 100$ y $a = 75$.

```
1  library(spatstat)
2  #comenzamos definiendo el grafo que vamos a emplear
3          L=domain(spiders)
4  #calculamos el centro del grafo
5          q=CentroGrafo(L)
6  #comenzamos definiendo la intensidad.
7  #Tomamos
8  lambda1=function(x,y){1+0*x}
9  #hecho esto, divimos la función anterior entre su integral para obtener la
   función de densidad asociada
10 lambda1.im=as.linim(lambda1,L)
11 m1=integral.linim(lambda1.im)
12 lambda10.im=lambda1.im/m1
13 #el siguiente paso es determinar la región en la que queremos añadir en
   media a puntos homogéneamente.
14 xwin=as.owin(L)$xrange
15 ywin=as.owin(L)$yrange
16         limits=c(0.3, 0.7, 0.3, 0.7)
17 xmin=xwin[1]+limits[1]*(xwin[2]-xwin[1])
18 xmax=xwin[1]+limits[2]*(xwin[2]-xwin[1])
19 ymin=ywin[1]+limits[3]*(ywin[2]-ywin[1])
20 ymax=ywin[1]+limits[4]*(ywin[2]-ywin[1])
21 #calculamos la longitud de la región del grafo en la que queremos añadir
   los a puntos en media
22 area=volume(L[owin(xrange=c(xmin,xmax),yrange=c(ymin,ymax))])
23 #definimos entonces m y a
```

```
24          m=100
25          a=3*m/4
26  #y definimos la función que, cuando multipliquemos por m, será la función
    de intensidad del segundo proceso puntual
27  lambda2=function(x,y){lambda1(x,y)/m1 +
    (1/(m*area))*a*(x>xmin)*(x<xmax)*(y>ymin)*(y<ymax)}
28  lambda2.im=as.linim(lambda2,L)
29  #definimos el vector de p-valores
30  Pvalues=numeric(1000)
31  #y tomamos 1000 réplicas Monte Carlo
32  for (i in 1:1000){
33          set.seed(2*i)
34          pp1=rpoislpp(m*lambda10.im,L)
35          pp2=rpoislpp(m*lambda2.im,L)
36          #y para cada una de ellas calculamos el p-valor del estadístico de
            Kolmogorov-Smirnov
37          Pvalues[i]=PermutationsKS(pp1,pp2,q,nboots = 7500L)$Pvalue
38  }
```

Al igual que para el estudio del nivel, si queremos estudiar la potencia para distintos valores de m y a, basta cambiar estos valores en las líneas 24 y 25. Para considerar el modelo inhomogéneo, únicamente hay que sustituir la línea 8 por `lambda1=function(x,y){1-(y/1125)^0.1}`. Finalmente, para considerar como soporte el grafo *Dendrite* hay que cambiar tanto el dominio en la línea 3: `L=domain(dendrite)`, como la región en la que añadimos en media a puntos, para lo que hay que sustituir la línea 16 por `limits=c(0.3, 0.575, 0.6, 0.9)`. Además, al igual que para el estudio del nivel, hay que especificar el punto base del π-sistema apropiado en *Dendrite*, lo que se consigue sustituyendo la línea 5 por `q=lpp(c(243.4,300),domain(dendrite))`.

A.3.2. Test de Cramer von Mises

Detallamos entonces el código empleado para el estudio del nivel y de la potencia del test de Cramer von Mises. Lo primero que debemos recordar es que en este caso no hemos empleado como soporte los grafos de *Spiders* y *Dendrite*, sino unos determinados subgrafos de ellos, cuya construcción especificaremos en el código. Comenzamos entonces con el código para el estudio del nivel del test de Cramer con Mises. Presentamos el código para el estudio del nivel en el subgrafo de *Spiders*, en el caso homogéneo con $m = 100$ puntos en media:

```
1  library(spatstat)
2  #Comenzamos definiendo el subgrafo con el que vamos a trabajar
3          L=domain(spiders)
4          win=as.owin(L)
5          L=L[owin(xrange=c(0.5,1)*win$xrange[2],
           yrange=c(0.5,1)*win$yrange[2])]
6  #el siguiente paso es construir la función de densidad asociada al modelo
   que hayamos elegido
7          lambda=function(x,y){1+0*x}
8  lambda.im=as.linim(lambda,L=L)
9  m0=integral.linim(lambda.im)
10 lambda0.im=lambda.im/m0
11 #definimos el número esperado de puntos de cada patrón
12         m=100
13 #tomamos el vector de p-valores
14 Pvalues=numeric(1000)
15 #consideramos entonces 1000 réplicas Monte Carlo
16 for (i in 1:1000){
17         set.seed(2*i)
18         pp1=rpoislpp(m*lambda0.im,L)
19         pp2=rpoislpp(m*lambda0.im,L)
20         #y para cada una calculamos el p-valor asociado
21         Pvalues[i]=PermutationsCvM(pp1,pp2,Heateq = T,nboots =
           5000L)$Pvalue
22 }
```

Para realizar el estudio del nivel con distintos valores de m, basta cambiar el valor de este parámetro en la línea 12. Si se desea emplear el modelo inhomogéneo, se ha de sustituir la función definida en la línea 7 por `lambda=function(x,y){1-(y/1125)^0.1}`. Finalmente, para realizar el estudio de simulación del test de Cramer von Mises en el subgrafo de *Dendrite*, en el código anterior hay que sustituir las líneas 3-5 por el código siguiente:

```
1  L=domain(dendrite)
2  win=as.owin(L)
3  subdom=L[owin(xrange=win$xrange[1] +
   c(0.125,0.36)*(win$xrange[2]-win$xrange[1]) , yrange=win$yrange[1] +
   c(0.7,0.95)*(win$yrange[2]-win$yrange[1]))]
4  vertex=coords(vertices(subdom))
```

```
eliminar=-which(0.676232*vertex[,1]+vertex[,2]>334.258 |
-0.468425*vertex[,1]+vertex[,2]<230.452)
L=thinNetwork(subdom , retainvertices = eliminar)
```

Finalmente, presentamos el código para el estudio de la potencia del contraste de Cramer von Mises. Ejemplificamos el modelo homogéneno en el subgrafo de *Spiders* con $m = 100$ y $a = 50$:

```
library(spatstat)
#construimos el soporte
        L=domain(spiders)
        win=as.owin(L)
        L=L[owin(xrange=c(0.5,1)*win$xrange[2],yrange=c(0.5,1)*win$yrange[2])]
#y definimos la función de densidad del primer proceso puntual asociada al
modelo que hayamos considerado
        lambda1=function(x,y){1+0*x}
lambda1.im=as.linim(lambda1,L)
m1=integral.linim(lambda1.im)
lambda10.im=lambda1.im/m1
#el siguiente paso es especificar la región en la que se añaden a puntos en
media de forma homogénea
xwin=as.owin(L)$xrange
ywin=as.owin(L)$yrange
        limits=c(0.3, 0.75, 0.3, 0.7)
xmin=xwin[1]+limits[1]*(xwin[2]-xwin[1])
xmax=xwin[1]+limits[2]*(xwin[2]-xwin[1])
ymin=ywin[1]+limits[3]*(ywin[2]-ywin[1])
ymax=ywin[1]+limits[4]*(ywin[2]-ywin[1])
#y calculamos la longitud de esta región
area=volume(L[owin(xrange=c(xmin,xmax),yrange=c(ymin,ymax))])
#definimos los parámetros asociados a los números de puntos en media
        m=100
        a=50
#y construimos la función necesaria para obtener la función de intensidad
asociada al segundo proceso puntual
lambda2=function(x,y){lambda1(x,y)/m1+(1/(m*area))*a*(x>xmin)*(x<xmax)*(y>ymin)*(y<ymax)]
lambda2.im=as.linim(lambda2,L)
#Definimos el vector de p-valores
Pvalues=numeric(1000)
```

```
29 #consideramos entonces 1000 réplicas Monte Carlo
30 for (i in 1:1000){
31         set.seed(2*i)
32         pp1=rpoislpp(m*lambda10.im,L)
33         pp2=rpoislpp(m*lambda2.im,L)
34         #y para cada una de ellas calculamos el nivel crítico
35         Pvalues[i]=PermutationsCvM(pp1,pp2,Heateq=T,nboots = 5000L)$Pvalue
36 }
```

Nuevamente, para realizar el estudio completo de la potencia del contraste de Cramer von Mises debemos variar los valores de m y a en las líneas 22 y 23 respectivamente. Para considerar el modelo inhomogéneo hemos de sustituir la línea 7 por `lambda=function(x,y){1-(y/1125)^0.1}`, y para realizar el estudio de la potencia en el subgrafo de *Dendrite* se han se intercambiar las líneas 3-5 por las apropiadas para la construcción del subgrafo de *Dendrite*, tal como hicimos para el estudio del nivel, junto con definir `limits=c(0.3, 0.575, 0.75, 1)` en la línea 14.

Índice de notación

$\mathbf{0}$	vector cero
$\mathbf{1}_{\{\cdot\}}$	función indicadora
$\lvert A\rvert$	área o longitud del conjunto A
$\mathcal{A}$	conjunto de aristas de un grafo lineal
AMISE	error cuadrático medio integrado asintótico
β_m	σ-álgebra de Borel en $\mathbb{R}^m$
$\#A$	cardinal del conjunto A
CSR	aleatoriedad espacial completa
$d_{\mathcal{G}}$	distancia del camino más corto en el grafo lineal $\mathcal{G}$
$\mathcal{G}$	grafo lineal
$\mathbb{E}$	esperanza
F	función de distribución
f	función de densidad
$\widehat{f}$	estimación de la función de densidad
$G_{\boldsymbol{p},h}$	función núcleo equitativo centrada en $\boldsymbol{p}$ con ancho de banda h
∇g	vector gradiente de una función g que tome valores escalares
$deg(\cdot)$	grado de un nodo de un grafo lineal
h	ancho de banda
h_{AMISE}	ancho de banda óptimo respecto al AMISE
h_{Norm}	ancho de banda óptimo respecto al AMISE para poblaciones normales
$\boldsymbol{H}$	matriz de ancho de banda
Hg	matriz Hessiana de una función g que tome valores escalares
$\mathcal{H}_a$	hipótesis alternativa

$\mathcal{H}_0$	hipótesis nula
$\boldsymbol{I}_d$	matriz identidad d-dimensional
K	función núcleo multidimensional
k	función núcleo unidimensional
L	densidad de probabilidad bivariante, isótropa, unimodal y centrada en $\mathbf{0}$
$\mathcal{L}_{\mathcal{G}}$	conjunto de puntos del plano del grafo lineal $\mathcal{G}$
l	arista de un grafo lineal
λ	función de intensidad
λ_c	función de intensidad condicionada
$\widehat{\lambda}$	estimación de la función de intensidad
ln	logaritmo en base e
M	número de réplicas Monte Carlo
$\mathfrak{m}$	marca de un proceso puntual
MISE	error cuadrático medio integrado
MSE	error cuadrático medio
$N(\cdot)$	medida de contar
$N_i(\cdot)$	medida de contar restringida al i-ésimo patrón de puntos
n	tamaño muestral determinista
n_p	número de permutaciones empleado en la estimación del nivel crítico
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	distribución normal de vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$
$\|\| \cdot \|\|$	norma euclídea
Ω	espacio muestral
ω	función núcleo básica
$\mathbf{P}$	proceso puntual en un grafo lineal
$\mathbb{P}$	probabilidad
$\mathcal{P}$	π-sistema
$\mathbf{p}$	patrón de puntos en un grafo lineal
$\boldsymbol{p}$	punto de un grafo lineal
$\mathfrak{p}$	nivel crítico o p-valor
$\mathrm{P}(A)$	conjunto de partes del conjunto A
$\mathbb{R}^m$	espacio euclídeo m-dimensional
$\mathbb{R}^+$	conjunto de los números reales positivos

RSS	suma de residuos cuadráticos
$\boldsymbol{S}$	matriz de suavizado
S	función núcleo apropiada para la regresión no paramétrica
T	estadístico
T_{KS}	estadístico de Kolmogorov-Smirnov
T_{CvM}	estadístico de Cramer von Mises
T_{NP}	estadístico no paramétrico basado en la función de riesgo relativo
$\mathcal{V}$	conjunto de nodos de un grafo lineal
$\mathbb{V}$ar	varianza
W	región de observación de un proceso puntual en el plano
$\mathbf{X}$	proceso puntual en el plano euclídeo
$\boldsymbol{X}$	vector aleatorio
X	variable aleatoria
$\mathbf{x}$	patrón de puntos en el plano euclídeo
$\boldsymbol{x}$	punto del espacio euclídeo multidimensional
x	número real
$\mathbf{y}$	patrón de puntos marcado en el plano
$Z(\cdot)$	covariable a un proceso puntual

Bibliografía

[1] Bacchieri, G., y Barros, A. J. (2011). Traffic accidents in Brazil from 1998 to 2010: many changes and few effects. *Revista de Saúde Pública*, 45, 949-963.

[2] Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., y Davies, T. M. (2021). Analysing point patterns on networks — A review. *Spatial Statistics*, 42.

[3] Baddeley, A., Rubak, E. y Turner, R. (2015). *Spatial Point Patterns: Methodology and Applications with R.* Chapman & Hall.

[4] Baddeley, A. y Turner, R. (2005). Spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software*, 12(6), 1-42.

[5] Borrajo, M. I., González-Manteiga, W. y Martínez-Miranda, M. D. (2020). Testing for significant differences between two spatial patterns using covariates. *Spatial Statistics*, 40.

[6] Bowman, A. W. y Azzalini, A. (1997). *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations.* Oxford Science Publications.

[7] Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2), 353-360.

[8] Dantzig, G. B. y Thapa, M. N. (1997). *Linear programming 1: introduction.* Springer.

[9] Dantzig, G. B. y Thapa, M. N. (2003). *Linear programming 2: Theory and Extensions.* Springer.

[10] de Burgos, J. (2007). *Cálculo Infinitesimal de una variable (Segunda edición).* McGraw-Hill.

[11] DeGroot, M.H. (1988). *Probabilidad y estadística (Segunda edición).* Addison-Wesley Iberoamericana

[12] Diggle, P.J. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society (Series C)*, 34, 138-147.

[13] Diggle, P.J. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* Chapman & Hall.

[14] Fuentes-Santos, I., González-Manteiga, W. y Mateu, J. (2021). Testing similarity between first-order intensities of spatial point processes. A comparative study. *Communications in Statistics-Simulation and Computation*, 1-21.

[15] Ghadirzadeh, M. R., Shojaei, A., Khademi, A., Khodadoost, M., Kandi, M., Alaeddini, F., y Moradi, S. (2015). Status and trend of deaths due to traffic accidents from 2001 to 2010 in Iran. *Iranian Journal of Epidemiology,* 11(2), 13-22.

[16] Gómez-Villegas, M. A. (2005). *Inferencia estadística.* Ediciones Díaz de Santos.

[17] Google Maps (*https://www.google.es/maps/@-22.9140693,-43.5860658,11z?hl=es*). Consultado el 05/06/2022

[18] Gopalakrishnan, S. (2012). A public health perspective of road traffic accidents. *Journal of family medicine and primary care,* 1(2), 144.

[19] Marozzi, M. (2004). Some remarks about the number of permutations one should consider to perform a permutation test. *Statistica*, 64(1), 193-201.

[20] McSwiggan, G., Baddeley, A. y Nair, G. (2017). Kernel density estimation on a linear network. *Scandinavian Journal of Statistics*, 44(2), 324-345.

[21] Okabe, A., Satoh, T. y Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, 23(1), 7-32.

[22] Okabe, A. y Sugihara, K. (2012). *Spatial analysis along networks: statistical and computational methods.* John Wiley & Sons.

[23] Rakshit, S., Davies, T., Moradi, M. M., McSwiggan, G., Nair, G., Mateu, J. y Baddeley, A. (2019). Fast kernel smoothing of point patterns on a large network using two-dimensional convolution. *International Statistical Review*, 87(3), 531-556.

[24] Rodríguez, G. (2003). *Diferenciación de Funciones de Varias Variables Reales.* Servizo de Publicacións e Intercambio Científico da USC.

[25] Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons.

[26] Venables, W. N. y Ripley, B. D. (2002) *Modern Applied Statistics with S. (Cuarta edición).* Springer.

[27] Wand, M.P. y Jones, M.C. (1995). *Kernel Smoothing.* Chapman & Hall.

[28] Yamada, I. y Thill, J. C. (2004). Comparison of planar and network K-functions in traffic accident analysis. *Journal of Transport Geography*, 12(2), 149-158.

[29] Yang, B. M., y Kim, J. (2003). Road traffic accidents and policy interventions in Korea. *Injury control and safety promotion,* 10(1-2), 89-94.

[30] Zhang, T. y Zhuang, R. (2017). Testing proportionality between the first-order intensity functions of spatial point processes. *Journal of Multivariate Analysis*, 155, 78-82.