

**Análisis Estadístico de Datos
Direccionales.**
Aplicaciones Medioambientales
Proyecto Fin de Máster
Máster en Técnicas Estadísticas

María Leyenda Rodríguez

Resumen

En diversos campos surgen problemas estadísticos donde los datos son dados mediante medidas angulares dando la orientación o ángulos en el plano (datos circulares) o en el espacio (datos esféricos). Datos circulares es el caso más simple de esta categoría de datos llamada datos direccionales, donde la única respuesta no es escalar, pero es angular o direccional. La suposición estadística básica es que los datos son una muestra aleatoria de una población de direcciones. Para trabajar con datos de esta naturaleza es necesario construir nuevos estadísticos pues los estadísticos usuales empleados para datos lineales son inapropiados en los datos direccionales, pues no tienen en cuenta la naturaleza circular de esta clase de datos.

En este trabajo vamos a analizar la dirección de viento recogida por la estación meteorológica A Mourela que pertenece al Sistema de Control Suplementario de la Contaminación Atmosférica de la U.P.T. de As Pontes propiedad de Endesa Generación S.A. Debido a la naturaleza circular de esta variable introduciremos los datos circulares según [9]. Comenzaremos con la estimación no paramétrica tipo núcleo de la densidad circular haciendo un especial hincapié en la selección del parámetro de suavizado. Notemos que estos datos de dirección de viento son minutales y recogidos a lo largo del año 2010 lo que nos llevó a estudiar el modelo de Möebius de series de tiempo.

Debido al interés por estudiar la relación entre las concentraciones de SO_2 y la dirección de viento realizaremos la estimación no paramétrica tipo núcleo de la función de regresión donde la selección del parámetro de suavizado también es de vital importancia. Notemos que para este apartado se va a trabajar con la dirección de viento recogida en las estaciones de medida automáticas B1, B2, C9, F2 y G2 que también forman parte del Sistema de Control Suplementario de la Contaminación Atmosférica de la U.P.T de As Pontes; pues para realizar la estimación de la función de regresión necesitamos que tanto los datos de SO_2 y dirección de viento estén recogidos por la misma estación.

Contenidos

Resumen	1
1 Introducción	1
1.1 Medidas de localización	1
1.2 Medidas de concentración y dispersión	3
1.3 Distribuciones notables	4
1.3.1 Distribución lattice	5
1.3.2 Distribución uniforme	5
1.3.3 Distribución von Mises	5
1.3.4 Distribución Cardioide	9
1.3.5 Distribución Normal proyectada	10
1.3.6 Distribución Wrapped	10
2 Función de densidad circular	13
2.1 Estimación no paramétrica tipo núcleo de la densidad circular . . .	13
2.2 Selección del parámetro de suavizado.	14
2.2.1 Regla plug-in escala von Mises	15
2.2.2 Validación cruzada	15
2.2.3 Ilustración	16
3 Función de regresión circular-lineal	17
3.1 Estimación no paramétrica tipo núcleo de la regresión circular-lineal	17
3.2 Selección del parámetro de suavizado	18
4 El modelo de Möbius de series de tiempo	19
4.1 El modelo Möbius de series de tiempo	19
4.2 Función de máxima verosimilitud	21
5 Aplicación. Datos Medioambientales	23
5.1 Datos perturbados	24
5.2 Análisis de los datos de dirección de viento en A Mourela	25

5.3	Estimación tipo núcleo de la función de densidad circular	30
5.4	Estimación del modelo de Möbius de series de tiempo	42
5.5	Estimación tipo núcleo de la función de regresión circular-lineal . .	47
5.5.1	Estación B1	54
5.5.2	Estación B2	55
5.5.3	Estación C9	57
5.5.4	Estación F2	58
5.5.5	Estación G2	60
Bibliografía		63

Capítulo 1

Introducción

Los datos circulares surgen de varias formas. Las dos principales corresponden a los instrumentos de medición circular clásicos: la brújula y el reloj. Entre las observaciones típicas medidas por la brújula están las direcciones de viento y las direcciones migratorias de los pájaros. Entre las típicas medidas por el reloj están los tiempos de llegada (en un reloj de 24 horas) de los pacientes a una unidad de urgencias de un hospital. Conjuntos de datos similares surgen al considerar veces en un año (o veces en un mes) de los eventos correspondientes.

Las direcciones en el plano se pueden observar como vectores unitarios en el plano como puntos en el círculo unidad. Aunque hay otras dos formas muy útiles de observar las direcciones- como ángulos y como números complejos; escogiendo una dirección y orientación inicial (Esto es equivalente a escoger un sistema de coordenadas ortogonal en el plano). Luego cada punto $\mathbf{x}$ en el círculo unidad puede ser representado por un ángulo θ o por un número complejo unitario z .

$$\begin{aligned}x &= (\cos \theta, \sin \theta) \\z &= e^{i\theta} = \cos \theta + i \sin \theta\end{aligned}$$

En este capítulo, presentaremos la descripción de los datos circulares siguiendo [9]. Comenzaremos por estudiar las medidas de localización, concentración y dispersión, secciones 1.1, 1.2., y finalmente nos centraremos en las principales distribuciones circulares, sección 1.3.

1.1 Medidas de localización

Sean $x_1, \dots, x_n$ cuyos ángulos correspondientes son θ_i , $i = 1, \dots, n$. La dirección media $\bar{\theta}$ de $\theta_1, \dots, \theta_n$ es la dirección de $x_1 + \dots + x_n$ que es el centro de masa $\bar{x}$ de $x_1, \dots, x_n$.

Por tanto si las coordenadas cartesianas de x_j son $(\cos \theta_j, \text{sen} \theta_j)$, entonces $(\bar{C}, \bar{S})$ son las coordenadas cartesianas del centro de masas.

$$\bar{C} = \frac{1}{n} \sum_{j=1}^n \cos \theta_j; \quad \bar{S} = \frac{1}{n} \sum_{j=1}^n \text{sen} \theta_j$$

Además $\bar{\theta}$ es la solución de las ecuaciones (1.1,1.2) y es la dirección de $x_1 + \dots + x_n$, denominada, dirección media

$$\bar{C} = \bar{R} \cos \bar{\theta} \tag{1.1}$$

$$\bar{S} = \bar{R} \text{sen} \bar{\theta} \tag{1.2}$$

(supuesto $\bar{R} > 0$), dónde la longitud media resultante $\bar{R}$ es dada por

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2} \tag{1.3}$$

Notemos que $\bar{\theta}$ no está definida cuando $\bar{R} = 0$ y cuando $\bar{R} > 0$, $\bar{\theta}$ es dada por

$$\bar{\theta} = \begin{cases} \tan^{-1}(\bar{S}/\bar{C}) & \text{si } \bar{C} \geq 0 \\ \tan^{-1}(\bar{S}/\bar{C}) + \pi & \text{si } \bar{C} < 0 \end{cases}$$

En el contexto de estadística circular $\bar{\theta}$ no es la media $(\theta_1 + \dots + \theta_n)/n$. Sin embargo, la denominamos media muestral de la dirección y cumple las siguientes propiedades

- es equivariante bajo rotación.
- Diferentes estadísticos utilizando diferentes sistemas de coordenadas estarán de acuerdo en dónde está la media la muestral, a pesar de que puede usar números diferentes para describir su posición

Por otra parte, la mediana muestral de la dirección $\bar{\theta}$ de los ángulos $(\theta_1, \dots, \theta_n)$ es algún ángulo ϕ tal que la mitad de los puntos se encuentran en el arco $[\phi, \phi + \pi)$ y la mayoría de los puntos están más cerca de ϕ que de $\phi + \pi$.

- Cuando n es par, la mediana coincide con uno de ellos.
- Cuando n es impar, es conveniente tomar la mediana como el punto medio de dos puntos adyacentes adecuados.

1.2 Medidas de concentración y dispersión

La media de longitud resultante $\bar{R}$ introducida en (1.3) es la medida de dispersión más importante en datos direccionales.

Luego, si $x_1, \dots, x_n$ son vectores unitarios, entonces $0 \leq \bar{R} \leq 1$. Por lo que, podemos concluir que

- Si las direcciones $\theta_1, \dots, \theta_n$ están estrechamente agrupadas luego $\bar{R} = 1$.
- Si $\theta_1, \dots, \theta_n$ están muy dispersas luego $\bar{R}$ será prácticamente 0.

Además esta medida de concentración tiene las siguientes propiedades

- $\bar{R}$ es invariante bajo rotaciones.
- La longitud resultante $\bar{R}$ es la longitud del vector resultante $x_1 + \dots + x_n$.
- $R = n\bar{R}$

A veces emplearemos otras medidas de dispersión de datos circulares, para comparar con datos en línea: la medida más simple es la varianza circular muestral,

$$V = 1 - \bar{R}$$

o la desviación circular estándar,

$$v = \sqrt{-2\log \bar{R}} \quad (1.4)$$

También podemos considerar como medida de dispersión la distancia entre dos ángulos θ y ξ

$$\min(\theta - \xi, 2\pi - (\theta - \xi)) = \pi - |\pi - |\theta - \xi||$$

o definir la medida de dispersión de los ángulos $\theta_1, \dots, \theta_n$ sobre un ángulo dado α es

$$d_0(\alpha) = \frac{1}{n} \sum_{i=1}^n (\pi - |\pi - |\theta_i - \alpha||)$$

la función d_0 toma el mínimo en la mediana muestral $\tilde{\theta}$. La desviación circular media es $d_0(\tilde{\theta})$

1.3 Distribuciones notables

Una forma de especificar una distribución en el círculo unidad es por medio de su función de distribución. Suponemos que ha sido escogida una dirección y orientación inicial en el círculo unidad. Luego la distribución puede ser considerada como que los ángulos aleatorios θ , y su función de distribución F es definida como la función en la recta real dada por

$$F(x) = Pr(0 < \theta \leq x), \quad 0 \leq x \leq 2 * \pi$$

y

$$F(x + 2\pi) - F(x) = 1, \quad -\infty < x < \infty \quad (1.5)$$

La ecuación (1.5) solo afirma que todo arco en el círculo unidad de longitud 2π tiene probabilidad uno (este arco es la totalidad de la circunferencia en el círculo). Para $\alpha \leq \beta \leq \alpha + 2\pi$,

$$Pr(\alpha < \theta \leq \beta) = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} dF(x) \quad (1.6)$$

dónde la integral es una integral de Lebesgue-Stieltjes. La función de distribución F es continua por la derecha. En contraste con las distribuciones en la recta real,

$$\lim_{x \rightarrow \infty} F(x) = \infty, \quad \lim_{x \rightarrow -\infty} F(x) = -\infty$$

Por definición

$$F(0) = 0, \quad F(2\pi) = 1$$

Notemos que aunque la función F dependa de la elección de la elección inicial, (1.6) muestra que $F(\beta) - F(\alpha)$ es independiente de esta elección. por tanto cambiar la dirección inicial es simplemente añadir una constante a F . Pues, si la función de distribución F es absolutamente continua tiene como función de densidad a f tal que

$$\int_{\alpha}^{\beta} f(x)dx = F(\beta) - F(\alpha), \quad -\infty < \alpha \leq \beta < \infty$$

Una distribución circular es una distribución de probabilidad la cual está concentrada en la circunferencia de círculo unidad. Las distribuciones circulares son de dos tipos:

1. Discretas- asignan masas de probabilidad solo a un número de direcciones
2. Absolutamente continuas

Una función f es la función de densidad de una distribución absolutamente continua si y solo si

1. $f(\theta) \geq 0$ en casi todo $(-\infty, \infty)$
2. $\int_0^{2\pi} f(\theta) d\theta = 1$
3. $f(\theta) = f(\theta + 2\pi)$ en casi todo $(-\infty, \infty)$

1.3.1 Distribución lattice

La distribución lattice (1.7) es una distribución discreta que toma valores en los vértices de un polígono de m lados inscrito en el círculo unidad, $\frac{2\pi r}{m}$. Si los pesos (1.8) son $p_r = \frac{1}{m}$, entonces se le denomina distribución uniforme discretizada en m puntos.

$$Pr(\theta = v + \frac{2\pi r}{m}) = p_r \quad r = 0, 1, \dots, m-1 \quad (1.7)$$

$$p_r \geq 0, \quad \sum_{r=0}^{m-1} p_r = 1 \quad (1.8)$$

1.3.2 Distribución uniforme

Es la distribución más básica en el círculo y a menudo es usada como modelo nulo. Esta es la única distribución en el círculo la cual es invariante bajo rotación y reflexión. Su función de densidad es

$$f(\theta) = \frac{1}{2\pi} \quad (1.9)$$

Por tanto, para $\alpha \leq \beta \leq \alpha + 2\pi$,

$$Pr(\alpha < \theta \leq \beta) = \frac{\beta - \alpha}{2\pi}$$

es decir, es proporcional a la longitud del arco.

1.3.3 Distribución von Mises

Desde el punto de la inferencia estadística, la distribución von Mises (1.10) es la mejor distribución en el círculo. Pues esta es unimodal y es simétrica sobre $\theta = \mu$ (Figura 1.1). Además la moda se encuentra en $\theta = \mu$ y la antimoda en $\theta = \mu + \pi$.

La relación de la moda de la densidad y la antimoda viene dada por $e^{2\kappa}$, así que cuanto mayor sea el valor de κ , mayor es el agrupamiento acerca de la moda.

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)} \quad (1.10)$$

$$F(\theta; 0, \kappa) = \frac{1}{I_0(\kappa)} \int_0^\theta e^{\kappa \cos u} du$$

- I_0 denota la función de Bessel modificada de primer tipo y orden 0,

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos \theta} d\theta, \quad I_0(\kappa) = \sum_{r=0}^{\infty} \frac{1}{(r!)^2} \left(\frac{\kappa}{2}\right)^{2r}$$

- El parámetro μ es la media de las direcciones
- κ es el parámetro de concentración. En (Figura 1.1) se observa que cuanto más grande sea el parámetro de concentración su función de densidad von Mises estará más concentrada en su media circular.

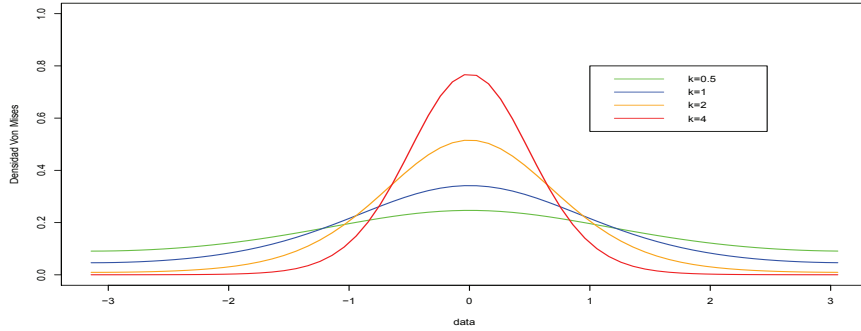


Figura 1.1: Representación de la función de densidad de Distribución von Mises $\text{vM}(0, \kappa)$, $\kappa=0.5, 1, 2, 3, 4$.

Como el parámetro κ es una medida de concentración en el caso de que nuestra muestra siga una von Mises; nos hemos planteado explicar explicar la relación entre diferentes medidas de dispersión como son la media de longitud resultante (1.3), la desviación circular estándar (1.4), la varianza definida de forma análoga que en el caso lineal (1.11) y el parámetro de concentración de la distribución von Mises κ .

Para ello,

1. Generamos una muestra $vM(\pi, \kappa)$ de tamaño 3000 para distintos valores del parámetro κ (75 valores).
2. Para cada una de las muestras calculamos
 - La varianza (1.11) con respecto a la distancia euclídea (1.12) y a la distancia definida por la longitud de arco (1.13),

$$S_n^2(\theta) = \frac{1}{n} \sum_{i=1}^n d^2(\theta_i, \theta) \quad (1.11)$$

$$\theta = \min_y S_n^2(y)$$

$$d^2(\theta_i, \theta_j) = 2(1 - (\cos(\theta_i - \theta_j))) \quad (1.12)$$

$$d^2(\theta_i, \theta_j) = \pi - |\pi - |\theta_i - \theta_j|| \quad (1.13)$$

- el parámetro de concentración $\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}$ y
 - la desviación circular estándar $\sqrt{-2\log(\bar{R})}$.
3. Representamos el estimador local-lineal para estudiar las relaciones de estas medidas de concentración y dispersión, frente a diferentes valores de κ .

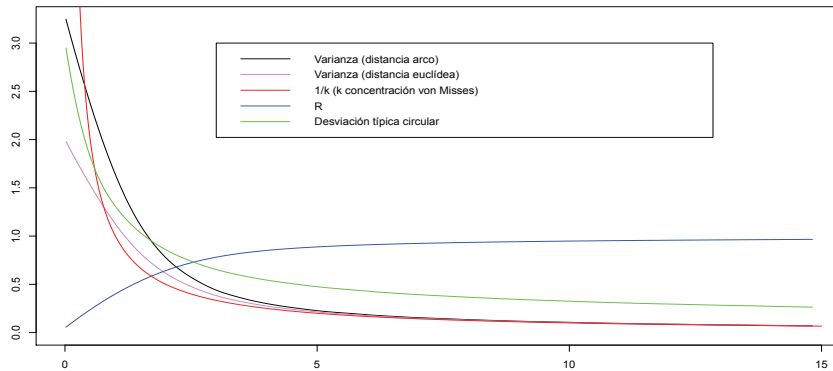


Figura 1.2: Relación entre $\bar{R}, v, S_n^2$ y κ

Mediante este estudio (Figura 1.2) se puede observar que v , S_n^2 (obtenida mediante cualquiera de las distancias) se comportan de forma similar. De modo que disminuye casi de modo exponencial a medida que κ aumenta. Mientras que $\bar{R}$ se comporta de modo diferente; aumenta a medida que κ aumenta aunque presenta una asíntota en 1.

La distribución von Mises juega un papel crucial en el campo de los datos direccionales. De hecho, juega el mismo papel que la distribución gaussiana en los datos lineales. Debido a esto nos va a interesar tanto aproximar como relacionar la distribución von Mises con otras distribuciones. Así pues,

- Si $\kappa = 0$, entonces $vM(\mu, \kappa)$ es la distribución uniforme (1.9). La aproximación $\exp(x) \cong 1 + x$, muestra que para κ pequeños $vM(\mu, \kappa) \cong C(\mu, \kappa/2)$, donde $C(\mu, \kappa/2)$ denota una distribución Cardioide (1.16). Por tanto, una distribución von Mises con parámetro de concentración pequeño puede ser aproximado por una distribución Cardioide con la misma dirección media y mismo parámetro de concentración dividido por 2.
- Cuando $\kappa \rightarrow \infty$ la distribución $vM(\mu, \kappa)$ está concentrada en el punto $\theta = \mu$. Si κ es grande, $\xi = \kappa^{1/2}(\theta - \mu)$. Luego, de (1.10) la función de densidad de ξ es proporcional a

$$\exp\{-\kappa[1 - \cos(\kappa^{-1/2}\xi)]\} \quad (1.14)$$

Para κ grande,

$$1 - \cos(\kappa^{-1/2}\xi) = \frac{1}{2}\kappa^{-1}\xi^2 + O(\kappa^{-2})$$

así pues, de (1.14), $\xi \sim N(0, 1)$. Luego para grandes valores de κ ,

$$\theta \sim vM(\mu, \kappa) \Rightarrow \kappa^{-1/2}(\theta - \mu) \sim N(0, 1), \quad \kappa \rightarrow \infty \quad (1.15)$$

- De foma más general, cualquier von Mises puede ser aproximada por una distribución wrapped normal (1.17).

$$vM(\mu, \kappa) \cong WN(\mu, A(\kappa)), \quad \kappa \rightarrow \infty$$

$$A_\kappa = I_1(\kappa)/I_0(\kappa)$$

Respecto a la convolución, notemos que la convolución de dos von Mises no es una distribución von Mises. En cambio, la convolución de dos distribuciones wrapped normal, $WN(\mu_1, A(\kappa_1))$ y $WN(\mu_2, A(\kappa_2))$ es la distribución wrapped normal $WN(\mu_1 + \mu_2, A(\kappa_1)A(\kappa_2))$. La cual puede ser aproximada por $vM(\mu_1 + \mu_2, A^{-1}(A(\kappa_1)A(\kappa_2)))$

$$\theta_1 + \theta_2 \sim vM(\mu_1 + \mu_2, A^{-1}(A(\kappa_1)A(\kappa_2)))$$

1.3.4 Distribución Cardioide

La perturbación de la densidad uniforme por la función coseno da lugar a una distribución Cardioide $C(\mu, \rho)$, cuya función de densidad es,

$$f(\theta) = \frac{1}{2\pi}(1 + 2\rho \cos(\theta - \mu)) \quad |\rho| < \frac{1}{2} \quad (1.16)$$

- La media de longitud resultante es ρ
- La media de la dirección es μ
- La distribución es simétrica y unimodal en μ (si $\rho > 0$) (Figura 1.3).
- Si $\rho = 0$ la distribución se reduce a la distribución uniforme (Figura 1.3).
- El principal uso de estas distribuciones es como aproximaciones de poca concentración a las distribuciones von Mises

$$\theta_i \sim C(\mu_i, \rho_i) (i = 1, 2) \Rightarrow \theta_1 + \theta_2 \sim C(\mu_1 + \mu_2, \rho_2 \rho_1)$$

Notemos que el conjunto de distribuciones Cardioides es cerrado bajo convolución.

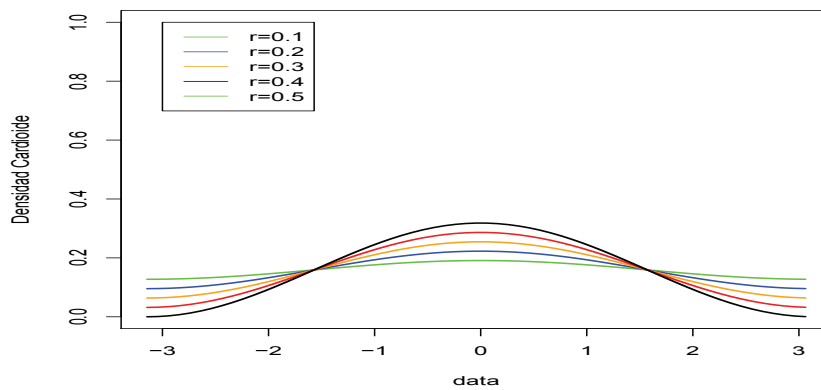


Figura 1.3: Representación de distribuciones cardiodes $C(0, \rho)$, $\rho=(0.1, 0.2, 0.3, 0.4, 0.5)$

1.3.5 Distribución Normal proyectada

Las distribuciones en el círculo pueden ser obtenidas mediante proyección radial de la distribución en el plano. Sea x un vector aleatorio bidimensional, $\Pr(x=0)=0$. Luego $\|x\|^{-1}x$ es un punto aleatorio sobre el círculo unidad.

$$p(\theta; \mu, \Sigma) = \frac{\phi(\theta; 0, \Sigma) + |\Sigma|^{-1/2} D(\theta) \phi(D(\theta)) \phi(|\Sigma|^{-1/2} (x^T \Sigma^{-1} x)^{-1/2} \mu \wedge x)}{x^T \Sigma^{-1} x}$$

$\phi(\cdot; 0, \Sigma)$ denota la función de densidad de $\mathbf{N}_2(0, \Sigma)$, ϕ y Φ denotan la función densidad y la función de densidad acumulada de $\mathbf{N}(0,1)$, $x = (\cos\theta, \sin\theta)^T$

$$D(\theta) = \frac{\mu^T \Sigma^{-1} x}{(x^T \Sigma^{-1} x)^{1/2}}$$

$$\mu \wedge x = \mu_1 \sin\theta - \mu_2 \cos\theta \quad \mu = (\mu_1, \mu_2)^T$$

1.3.6 Distribución Wrapped

Dada una distribución en la línea, se puede envolver alrededor de la circunferencia del círculo de radio uno. Si x es una variable aleatoria en la línea, la variable aleatoria correspondiente x_w de la distribución wrapped viene dada por

$$x_w = x(\text{mod } 2\pi)$$

Si el círculo es identificado con el conjunto de números complejos de módulo la unidad luego el mapa wrapping $x \rightarrow x_w$ pueden ser escrito como

$$x \rightarrow e^{2\pi i x}$$

Si x tiene como función de distribución $F \Rightarrow$ la función de distribución F_w de x_w viene dada por

$$F_w(\theta) = \sum_{\kappa=-\infty}^{\infty} \{F(\theta + 2\pi\kappa) - F(2\pi\kappa)\}$$

Si x tiene como función de masa de probabilidad $f \Rightarrow$ la función de densidad f_w de x_w es

$$f_w(\theta) = \sum_{\kappa=-\infty}^{\infty} f(\theta + 2\pi\kappa)$$

Las distribuciones Wrapped tienen las siguientes propiedades:

1. $(x + y)_w = x_w + y_w$
2. Si la función característica de x es ϕ entonces la función característica $\{\phi_p : p = 0, \pm 1, \dots\}$ de x_w es dada por $\phi_p = \phi(p)$
3. Si ϕ es integrable entonces x tiene función una densidad y

$$f_w(\theta) = \sum_{\kappa=-\infty}^{\infty} f(\theta + 2\pi\kappa) = \frac{1}{2\pi} \left[1 + 2 \sum_{p=1}^{\infty} (\alpha_p \cos p\theta + \beta_p \sin p\theta) \right]$$

$$\Phi(p) = \alpha_p + i\beta_p$$

4. Si x es infinitamente divisible luego x_w es infinitamente divisible.

Hay muchas distribuciones lineales las cuales pueden ser envueltas en cualquier distribución dada en el círculo. Sea g la función de densidad de una distribución en el círculo y define una función de densidad en la línea por

$$f(x) = p_r g(x)$$

$$2\pi r < x \leq 2\pi(r + 1) \quad r = 0, \pm 1, \pm 2, \dots,$$

p_r son números no negativos tales que, $\sum_{r=-\infty}^{\infty} p_r = 1$. Luego, $f_w = g$.

Hay diversos tipos de distribuciones wrapped. Uno de ellos es la distribución wrapped normal $WN(\mu, \rho)$ la cual se obtiene al "envolver" la distribución $N(\mu, \sigma^2)$ en el círculo, donde

$$\sigma^2 = -2 \log \rho, \quad \rho = e^{-\sigma^2/2}$$

Su función de densidad viene dada por

$$\phi_w(\theta; \mu, \rho) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{p=1}^{\infty} \rho^{p^2} \cos p(\theta - \mu) \right\} \quad 0 \leq \rho \leq 1 \quad (1.17)$$

$$(1.18)$$

Capítulo 2

Función de densidad circular

En este capítulo, describiremos una estimación no paramétrica tipo núcleo de la densidad circular, sección 2.1. Al igual que en el caso lineal, la selección del parámetro de suavizado será crucial. Se estudiarán tres métodos de selección; un método usando la técnica plug-in y dos mediante validación cruzada que surgen de minimizar la función de pérdida dada por minimizar el error cuadrático medio (L2) o minimizar la función de pérdida dada por Kullback-Leibler.

2.1 Estimación no paramétrica tipo núcleo de la densidad circular

En el caso elemental de la estimación no paramétrica tipo núcleo de la densidad univariante $f(x)$ usando observaciones con valores reales $x_1, \dots, x_n$, con núcleo K viene dada por el estimador de Parzen-Rosemblat

$$\hat{f}(x) = \frac{1}{nh} \sum K\left(\frac{x - x_i}{h}\right)$$

dónde

- h es el parámetro de suavizado o ventana.
- Usualmente se supone que la función núcleo es una función de masa de probabilidad simétrica, por ejemplo, la densidad Gausiana.
- La estimación de la densidad tipo núcleo se extiende fácilmente a datos circulares, aunque se debe tener cuidado en la selección de la función núcleo.

Sin embargo, cuando usamos datos en el círculo, no podemos emplear la distancia en el espacio Euclídeo, así que las diferencias $\theta - \theta_i$ serán remplazadas por el ángulo dado por la diferencia entre dos vectores,

$$d_i = ||\theta - \theta_i|| = \min(|\theta - \theta_i|, 2\pi - |\theta - \theta_i|)$$

Esto también puede escribirse como $d_i = \cos^{-1}(x^T x_i)$, donde $x^T = (\cos\theta, \sin\theta)$ es un vector unitario. Una elección natural de la función núcleo es usar una función de densidad circular como la dada por la distribución de von Mises. Esto permite una representación alternativa de la estimación no paramétrica tipo núcleo de la densidad,

$$\hat{f}(x) = \frac{1}{nh} \sum K\left(\frac{1 - x^T x_i}{h}\right)$$

Dada una muestra aleatoria de ángulos $\theta_1, \dots, \theta_n \in [0, 2\pi]$. Si consideraremos que el estimador no paramétrico tipo núcleo de la densidad con función núcleo la distribución von Mises (2.1), viene dado por

$$\hat{f}(\theta, \nu) = \frac{c_0(\nu)}{n} \sum_{i=1}^n L(\nu \cos(\theta - \theta_i)) \quad (2.1)$$

$$c_0(\nu) = \frac{1}{2\pi I_0(\nu)} \quad L(x) = e^x \quad (2.2)$$

- Dónde $I_r(\nu)$ es la función de Bessel modificada de orden r .
- El parámetro de concentración ν ahora ha asumido el papel de la inversa del parámetro de suavizado.

Para cualquier θ y entero j , los ángulos θ y $\theta + 2\pi j$ son idénticos una función de densidad puede ser necesaria para el núcleo K . Una posible opción planteada por [9] es la distribución von Mises (1.10). El parámetro de concentración controla el grado de suavidad en estimación no paramétrica tipo núcleo de la densidad circular y es análogo al parámetro de suavizado excepto que valores altos de ν proporcionan menos suavidad y pequeños valores de ν conducen a mayor suavidad.

2.2 Selección del parámetro de suavizado.

A continuación, nos centramos en la elección del parámetro de suavizado. Existen varios métodos para seleccionar el parámetro de suavizado subjetivamente, regla plug-in escala von Mises, o automáticamente, validación cruzada.

2.2.1 Regla plug-in escala von Mises

Si suponemos que $f(\cdot)$ es von Mises con concentración κ y (sin pérdida de generalidad) media circular $\mu = 0$. El error asintótico cuadrático medio integrado es

$$AMISE(\nu) = 3\kappa^2 I_2(2\kappa) / \{32\pi\nu^2 I_0(\kappa)^2\} + \nu^{1/2} / (2n\pi) \quad (2.3)$$

dónde el primer sumando es la expresión asintótica del sesgo cuadrático integrado y el segundo la expresión asintótica de la varianza cuadrática integrada.

Por tanto, el error asintótico cuadrático medio integrado (2.3) es de la forma $a\nu^{-2} + b\nu^{1/2}$ la cual puede ser minimizada diferenciando respecto de ν e igualando a cero. Esto permite una "Regla plug-in escala von Mises" para el parámetro de suavizado ν basado en la estimación de κ [10].

$$\nu = [3n\hat{\kappa}^2 I_2(2\hat{\kappa}) \{4\pi^{1/2} I_0(\hat{\kappa})^2\}^{-1}]^{2/5} \quad (2.4)$$

Por tanto el parámetro de suavizado, así estimado, tiene una expresión de la forma $\nu = Cn^{2/5}$ dónde

$$C = [3\hat{\kappa}^2 I_2(2\hat{\kappa}) \{4\pi^{1/2} I_0(\hat{\kappa})^2\}^{-1}]^{2/5}$$

2.2.2 Validación cruzada

Introducimos la validación cruzada para minimizar la función pérdida dada por el error cuadrático medio y la función de pérdida dada por Kullback-Leibler. Consideremos la estimación no paramétrica tipo núcleo de la densidad construída dejando fuera el valor θ_j de la muestra.

$$\begin{aligned} \hat{f}_j(\theta, \nu) &= \frac{c_0(\nu)}{n} \sum_{i \neq j} L(\nu \cos(\theta - \theta_i)) \\ c_0(\nu) &= \frac{1}{2\pi I_0(\nu)} \quad L(x) = e^x \end{aligned}$$

Sea

$$\begin{aligned} cv_2(\nu) &= 2n^{-1} \sum_{i=1}^n \hat{f}_i(\theta_i, \nu) - \int \hat{f}^2(\theta, \nu) d\theta \\ cv_{KL} &= n^{-1} \sum_{i=1}^n \hat{f}_i(\theta_i, \nu) \end{aligned}$$

Luego $-cv_2(\nu) + \int f^2$ y $-cv_{KL}(\nu) + \int f \log(f)$ son estimaciones insesgadas de la pérdida del error cuadrático $L_2(\nu)$ y la pérdida dada por Kullback-Leibler [6], respectivamente.

$$v_2 = \operatorname{argmin}_{\nu \geq 0} -cv_2(\nu) \quad (2.5)$$

$$v_{KL} = \operatorname{argmin}_{\nu \geq 0} -cv_{KL}(\nu) \quad (2.6)$$

De estas dos expresiones obtenemos dos posibles valores para el parámetro de suavizado.

2.2.3 Ilustración

Con el fin de estudiar como se comportan los tres parámetros de suavizado, obtenidos en este capítulo, para la estimación no paramétrica tipo núcleo de la densidad (2.1). Hemos simulado una muestra de tamaño 100 que sigue una mixtura de von Mises y se ha representado su distribución teórica junto a las tres estimaciones de la densidad obtenidas tras seleccionar diferentes parámetros de suavizado (2.4, 2.5, 2.6).

En (Figura 2.1) se observa como las estimaciones no paramétricas tipo núcleo de la densidad (2.1) obtenidas tras estimar el parámetro de validación cruzada (2.5, 2.6) son mejores que la dada por el parámetro de suavizado obtenido mediante plug-in (2.4). Pues ambas estimaciones se parecen notablemente a la teórica.

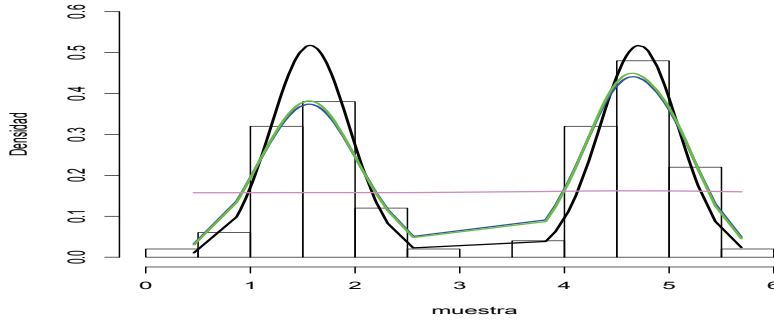


Figura 2.1: Representación de la densidad teórica de una mixtura de von Mises $vM(\pi/2,15), vM(3\pi/2,15)$ (negra) y la estimación no paramétrica de la densidad seleccionando el parámetro de suavizado mediante las técnicas estudiadas. L2 (azul), Kullback-Leibler (verde), plug-in (violeta)

Capítulo 3

Función de regresión circular-lineal

En este capítulo describiremos una estimación no paramétrica tipo núcleo de la regresión circular-lineal, sección 3.1, que se puede utilizar para investigar la relación entre una variable explicativa direccional y una variable respuesta lineal. Mediante validación cruzada obtendremos razonables parámetros de suavizado, sección 3.2.

La estimación no paramétrica tipo núcleo de la función de regresión es comunmente usada como una forma de resumir la relación entre dos variables sin requerir el supuesto de una forma paramétrica para describir esta relación.

3.1 Estimación no paramétrica tipo núcleo de la regresión circular-lineal

Dadas n observaciones de una variable direccional explicativa $\theta_1, \dots, \theta_n$ y una variable respuesta lineal $y_1, \dots, y_n$, suponemos que $y_i = m(\theta_i) + \epsilon_i$, dónde ϵ_i son variables con media cero, independientes e idénticamente distribuidos. Luego, la estimación no paramétrica tipo núcleo de la regresión circular-lineal, $m(\theta)$, viene dada por [5]

$$\hat{m}(\theta; \kappa) = \frac{\sum_{i=1}^n y_i g(\theta - \theta_i, 0, \kappa)}{\sum_{i=1}^n g(\theta - \theta_i, 0, \kappa)} \quad (3.1)$$

La cual es una estimación (3.1) análoga al estimador de Nadaraya-Watson para caso lineal.

3.2 Selección del parámetro de suavizado

La elección del grado de suavizado es crucial para la estimación no paramétrica tipo núcleo de la regresión, así como para la estimación no paramétrica tipo núcleo de la densidad. Existen varios métodos para seleccionar el parámetro de suavizado subjetivamente o automáticamente y pueden ser extendidos al caso de la estimación no paramétrica tipo núcleo de la regresión.

Por ejemplo, cuando el núcleo de suavizado evaluado sobre datos lineales es un núcleo Gausiano, Silverman recomienda un parámetro de suavizado de $0.9\hat{\sigma}n^{-1/5}$ y el valor $h_s = 1.06\hat{\sigma}n^{-1/5}$ es sugerido por Scott. (En ambos casos $\hat{\sigma}$ es típicamente dado por el mínimo de la desviación típica muestral y el rango intercuartílico dividido por 1.34). Ambas pueden estar conectadas con el parámetro κ en la distribución de von Mises usando el siguiente enlace conocido entre la distribución de von Mises y la distribución Gausiana propiedad de la distribución von Mises [9].

Si θ es distribuída de acuerdo con la distribución von Mises centrada en μ y con parámetro κ , luego

$$\kappa^{-1/2}(\theta - \mu) \rightarrow_D N(0, 1) \quad \kappa \rightarrow \infty$$

Esto sugiere la elección del parámetro de concentración κ mediante

$$\kappa = \frac{1}{h_s^2} \quad (3.2)$$

Otra opción es seleccionar κ por validación cruzada de mínimos cuadrados (LCV). Como en el caso lineal, escogemos κ de modo que minimize a la función.

$$CV(\kappa) = n^{-1} \sum_{j=1}^n [y_j - \hat{m}^{-j}(\theta_j; \kappa)]^2 \quad (3.3)$$

dónde

$$\hat{m}^{-j}(\theta_j; \kappa) = \frac{\sum_{i \neq j}^n y_i g(\theta_j - \theta_i, 0, \kappa)}{\sum_{i \neq j}^n g(\theta_j - \theta_i, 0, \kappa)}$$

son las estimaciones sin el dato j -ésimo.

La estimación no paramétrica tipo núcleo seleccionando el parámetro de suavizado mediante LCV puede ser inconsistente bajo una variedad de circunstancias. En particular, para datos discretos con múltiples valores repetidos, validación cruzada tiende a sugerir parámetros de suavizado que suavizan muy poco, es decir, κ tiende a ser muy grande. Por tanto valores de κ escogidos de acuerdo (3.2) serán preferidos en este caso.

Capítulo 4

El modelo de Möbius de series de tiempo

En este capítulo un modelo de regresión circular propuesto en [3] es adaptado al contexto de series de tiempo siguiendo la metodología propuesta en [7], sección 4.1. La distribución de θ_t dado θ_{t-1} es modelada usando una distribución von Mises particular. En la sección 4.1 comentaremos el modelo de regresión estudiado por [3] y lo adaptaremos a una serie de tiempo. La función de máxima verosimilitud (condicionada a la primera observación) es obtenida en la sección 4.2, así como la estimación de los parámetros.

4.1 El modelo Möbius de series de tiempo

La componente determinística del modelo de regresión estudiado por [3] une a la variable angular dependiente v con la variable angular independiente u mediante

$$\tan \frac{1}{2}(v - \beta) = \omega \tan \frac{1}{2}(u - \alpha) \quad (4.1)$$

dónde $\omega \in [-1, 1]$ es el parámetro que determina la pendiente y $-\pi \leq \alpha, \beta < \pi$ son parámetros que determinan la localización angular. La ecuación (4.1) nos conduce a

$$v = \beta + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(u - \alpha) \right\} \quad (4.2)$$

Luego, aplicaremos la ecuación (4.2) a las series de datos temporales, reemplazando el ángulo v por θ_t y el ángulo u por θ_{t-1} , $t = 2, \dots, n$. Esta sustitución nos sugiere la existencia de un único parámetro de localización, $\alpha = \beta$, obteniendo así

$$\tan \frac{1}{2}(\theta_t - \alpha) = \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \quad (4.3)$$

y

$$\theta_t = \alpha + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\} \quad (4.4)$$

ecuaciones análogas a (4.1) y (4.2).

Para el modelo de serie de tiempo dado en (4.4) se asume que $\theta_t|\theta_{t-1}$ sigue una distribución von Mises

$$\theta_t|\theta_{t-1} \sim vM \left(\alpha + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\}, k \right) \quad (4.5)$$

Por tanto el modelo de serie de tiempo se convierte en

$$\theta_t = \alpha + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\} + \epsilon_t \quad (4.6)$$

dónde $\epsilon_t \sim vM(0, k)$. Nos referiremos a la distribución condicionada de la media direccional de θ_t dado θ_{t-1} como μ_t . Esto es

$$\mu_t = \alpha + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\} \quad (4.7)$$

El valor de ω se restringe al intervalo $[-1, 1]$ así que α es único de identificación. Esto es, restamos π al valor de α en la ecuación (4.3). Luego

$$\tan \frac{1}{2}(\theta_t - \alpha + \pi) = \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha + \pi) \quad (4.8)$$

Ahora tengamos en cuenta que $\frac{1}{2}(\phi + \pi) = -\cot \frac{1}{2}\phi$, (4.8) es equivalente a

$$-\cot \frac{1}{2}(\theta_t - \alpha) = -\omega \cot \frac{1}{2}(\theta_{t-1} - \alpha)$$

de lo que se obtiene

$$\tan \frac{1}{2}(\theta_t - \alpha) = \frac{1}{\omega} \tan \frac{1}{2}(\theta_{t-1} - \alpha) \quad (4.9)$$

La equivalencia de (4.8) y (4.9) muestra que, si ω no estuviese restringido al intervalo $[-1, 1]$, y si $(\hat{\alpha}, \hat{\omega})$ es una solución de (4.3), luego $(\hat{\alpha} - \pi, \frac{1}{\hat{\omega}})$ sería una solución equivalente.

4.2 Función de máxima verosimilitud

Como

$$\theta_t|\theta_{t-1} \sim vM(\alpha + 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\}, k)$$

se obtiene que

$$f(\theta_t|\theta_{t-1}) = 2\pi I_0(\kappa)^{-1} \exp \kappa \cos(\theta_t - \mu_t)$$

dónde μ_t es dada por (4.7)

Luego la función de masa de probabilidad de $\theta_2, \dots, \theta_n$ dado θ_1 es

$$\begin{aligned} f(\theta_2, \dots, \theta_n|\theta_1) &= f(\theta_2|\theta_1)f(\theta_3, \dots, \theta_n|\theta_1, \theta_2) \\ &= f(\theta_2|\theta_1)f(\theta_3|\theta_1, \theta_2)f(\theta_4, \dots, \theta_n|\theta_1, \theta_2, \theta_3) \\ &= \dots = f(\theta_2|\theta_1)f(\theta_3|\theta_1, \theta_2)f(\theta_4|\theta_1, \theta_2, \theta_3) \cdots f(\theta_n|\theta_1, \dots, \theta_{n-1}) \end{aligned}$$

Pero en ecuación (4.5) el valor de θ_t depende solo del valor θ_{t-1} . Además,

$$f(\theta_t|\theta_1, \dots, \theta_{t-1}) = f(\theta_t|\theta_{t-1}),$$

$\forall t = 2, \dots, n$. La función de máxima verosimilitud condicionada es

$$L_C(\alpha, \omega, \kappa) = \{2\pi I_0(\kappa)\}^{-(n-1)} \exp \left\{ \kappa \sum_{t=2}^n \cos \left[\theta_t - \alpha - 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\} \right] \right\}$$

dando la verosimilitud

$$l_C(\alpha, \omega, \kappa) = \text{const.} - (n-1) \log I_0 \kappa + \kappa \sum_{t=2}^n \cos \left[\theta_t - \alpha - 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\} \right].$$

La cual es maximizada respecto los parámetros α y ω desconocidos

$$l_C(\alpha, \omega) = \sum_{t=2}^n \cos \left[\theta_t - \alpha - 2 \tan^{-1} \left\{ \omega \tan \frac{1}{2}(\theta_{t-1} - \alpha) \right\} \right]. \quad (4.10)$$

La maximización de (4.10) la haremos usando la función *nlm* de R pero minimizando la función $-l_C(\alpha, \omega)$ usando un algoritmo tipo Newton. Una vez que $l(\alpha, \omega)$ ha sido maximizada con respecto a α y ω , una aproximación mediante verosimilitud puede ser usada para obtener una estimación de máxima verosimilitud de κ , maximizando

$$l_C(\hat{\alpha}, \hat{\omega}, \kappa) = \text{const.} - (n-1) \log I_0 \kappa + \kappa l_{\hat{\alpha}, \hat{\omega}} \quad (4.11)$$

respecto de κ . Diferenciando (4.11) respecto de κ y teniendo en cuenta que $d(I_0(\kappa))/d\kappa = I_1(\kappa)$, la función de Bessel de primer tipo y de orden uno, obtenemos

$$\frac{\partial}{\partial \kappa}[l_C(\hat{\alpha}, \hat{\omega}, \kappa)] = -(n-1) \frac{I_1(\kappa)}{I_0(\kappa)} + l_{\hat{\alpha}, \hat{\omega}},$$

así que κ es la solución de

$$\frac{I_1(\hat{\kappa})}{I_0(\kappa)} = \frac{l_{\hat{\alpha}, \hat{\omega}}}{n-1}.$$

Capítulo 5

Aplicación. Datos Medioambientales

En este capítulo, realizaremos un análisis exploratorio de la dirección de viento tomada en la estación meteorológica, A Mourela, durante el año 2010. Para ello realizaremos un análisis descriptivo de esta variable (sección 5.2.) así como realizaremos una estimación no paramétrica tipo núcleo de la densidad (sección 5.3.) y estimaremos el modelo de Möbius de series de tiempo (sección 5.4.).

Además, estudiaremos que tipo de relación hay entre el SO_2 y la dirección de viento en las estaciones de medida automáticas B1, B2, C9, F2 Y G2, construyendo un modelo de regresión circular-lineal (sección 5.5).

Cabe mencionar que tanto la estación meteorológica A Mourela como las estaciones de medida automáticas pertenecen al Sistema de Control Suplementario de la Contaminación Atmosférica de la U.P.T. de As Pontes propiedad de Endesa Generación S.A.

Por otra parte, notemos que para llevar a cabo este capítulo fue necesaria realizar una perturbación en los datos, la cual está descrita en la sección 5.1. Por tanto, en este capítulo no vamos a trabajar con los datos minutales recogidos en la estación meteorológica y en las estaciones de medida automáticas B1, B2, C9, F2, G2; sino que vamos a trabajar con los datos horarios perturbados.

5.1 Datos perturbados

Al analizar los datos obtenidos por los dispositivos de dirección se observó que había valores repetidos tanto en la dirección de viento como en las concentraciones de SO_2 . Esto pueden ser debido a que estos dispositivos no presentan una precisión tan alta como para evitarlo. La aparición de medidas repetidas va a ser un problema a la hora de implementar ciertas técnicas, como la validación-cruzada para escoger el parámetro de suavizado.

Con el fin de solucionar dicho problema, haremos una perturbación en ambas variables. La perturbación en el caso lineal, viene dada por [1], una muestra artificial de la concentración de SO_2 viene dada por:

$$\tilde{X}_i = X_i + b\epsilon_i$$

, dónde X_i denota los valores observados de SO_2 , $b = 1.03\hat{\sigma}n^{-1/3}$ y $\epsilon_i, i = 1, \dots, n$ son variables aleatorias independientes e idénticamente distribuidas siguiendo el núcleo de Epanechnikov en $(-\sqrt{5}, \sqrt{5})$. $\hat{\sigma}$ es un estimador robusto de la varianza, el cual puede ser obtenido usando el rango intercuartílico estandarizado. [1] muestra que esta elección de b para los datos perturbados permite una estimación consistente de la función de distribución, obteniendo un error cuadrático medio con la misma magnitud que el de la función empírica de distribución acumulada.

En el caso circular, la perturbación se realiza de forma similar al caso lineal. La muestra artificial de la dirección del viento es

$$\tilde{\theta}_i = \theta_i + d\epsilon_i$$

, dónde θ_i denota los valores observados de la dirección de viento y $\epsilon_i, i = 1, \dots, n$ son variables aleatorias independientes generadas de la distribución von Mises con $\mu = 0$ y $\kappa = 1$. En [4] se toma $d = n^{-1/5}$ basándose en los resultados de [8] para la estimación tipo núcleo de la distribución multivariante. Esta perturbación resuelve el problema de los valores repetidos y no afecta a la función de distribución subyacente.

5.2 Análisis de los datos de dirección de viento en A Mourela

En esta sección, vamos a analizar los datos de viento recogidos durante el año 2010 por la estación meteorológica A Mourela situada en As Pontes (Figura 5.1), que van a ser usados en todo el capítulo excepto en la sección 4.4.

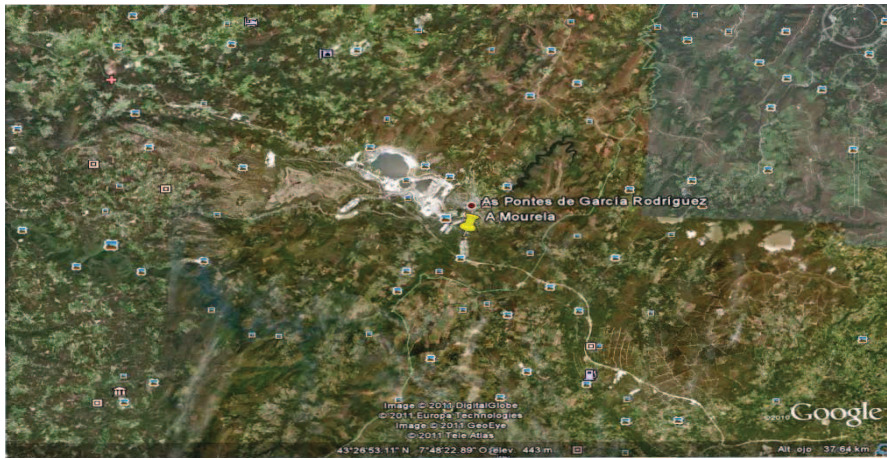


Figura 5.1: Estación Meteorológica A Mourela

Esta estación recoge datos minutales pero ,en este caso, trabajaremos con las medias horarias con el objetivo de reducir la cantidad de datos. Aunque a la hora de presentar los resultados de los datos horarios de viento nos seguimos encontrando con una gran cantidad de datos, 8360. Con el objetivo de que los gráficos sean más fáciles de visualizar hemos dividido el año 2010 en cuatro periodos de 2090 datos (Tabla 5.1).

En lo que se refiere a la notación, cabe mencionar que la dirección de viento va a tomar valores entre $[0, 2\pi)$ y cada valor representa el ángulo que forma respecto al Este. Por tanto, vamos a usar la siguiente codificación $0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$ representan las direcciones Este, Norte, Oeste y Sur, respectivamente.

En este análisis exploratorio de los datos , observamos que en el primer y el cuarto periodo del año 2010 la dirección media es la de Sureste y que la dirección media del segundo y tercer periodo del año 2010 es la de Noreste (Tabla 5.1).

	n	Media direccional	Medida de dispersión
Primer periodo	2090	5.3634	0.2177
Segundo periodo	2090	0.6446	0.2850
Tercer periodo	2090	0.2134	0.1750
Cuarto periodo	2090	5.7374	0.2952

Tabla 5.1: Estudio de los datos de dirección de viento 2010 mediante una medida de localización (media direccional) y una medida de dispersión ($\bar{R}$)

Se observa que los datos están muy dispersos (Figuras 5.2, 5.4, 5.6, 5.8) lo que ya se podía deducir de (Tabla 5.1) dónde se muestra que $\bar{R}$ es próxima a 0 en los cuatro periodos en los que hemos dividido el año.

Además, a la vista de los diagramas de rosa (Figuras 5.3, 5.5, 5.7, 5.9) podemos decir que la distribución de la dirección de viento en los 4 periodos en los que hemos dividido el año 2010 es prácticamente uniforme. Aunque cabe destacar que el cuarto periodo del año 2010 presenta una ligera predominancia de la dirección sureste.

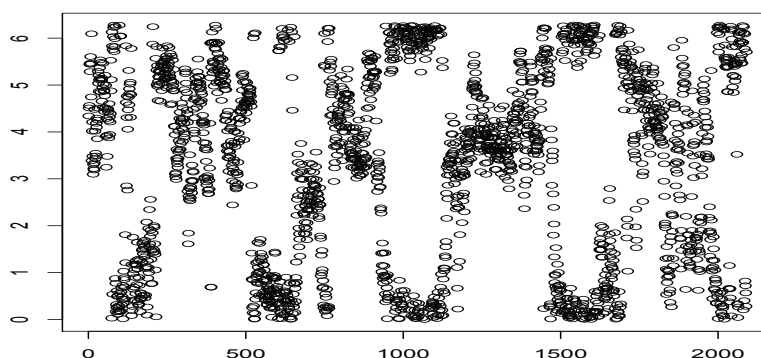


Figura 5.2: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010.

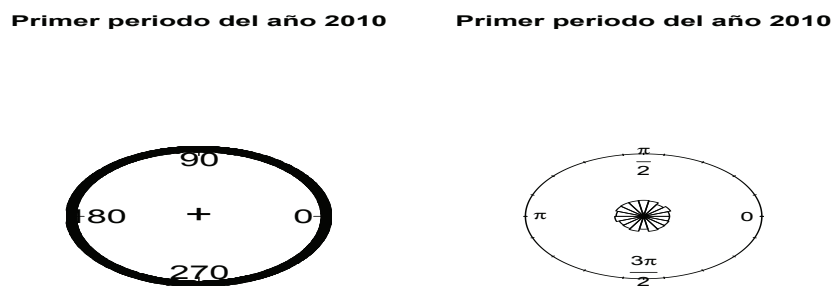


Figura 5.3: Gráfico de dispersión circular y diagrama de rosa de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010.

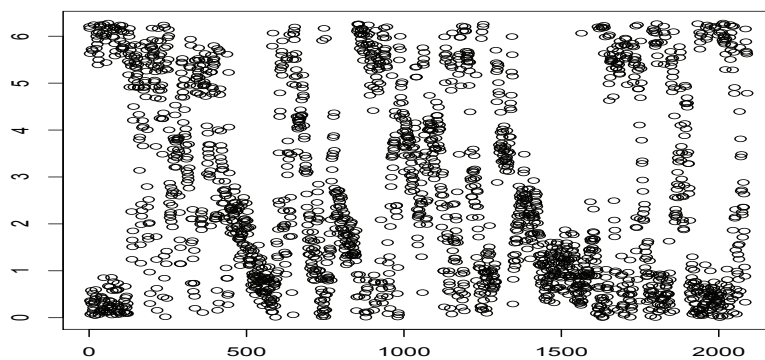


Figura 5.4: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010.

Segundo periodo del año 2010 Segundo periodo del año 2010

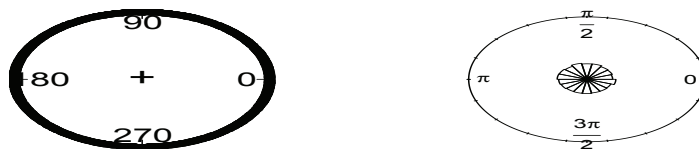


Figura 5.5: Gráfico de dispersión circular y diagrama de rosa de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010.

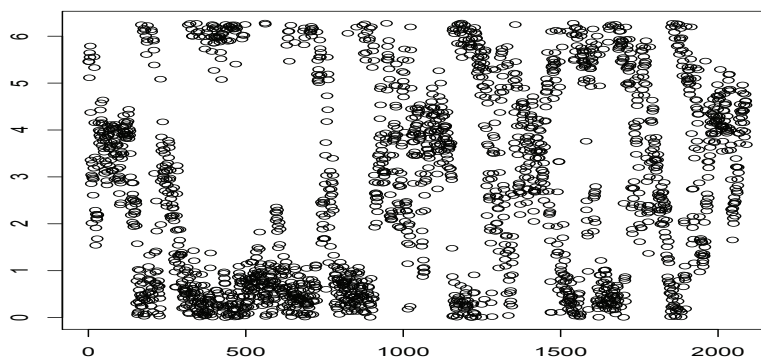


Figura 5.6: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010.

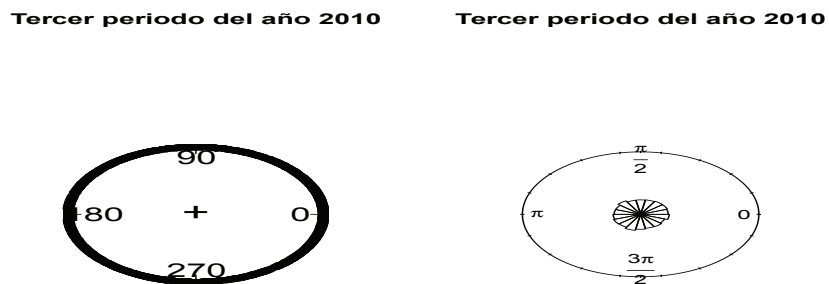


Figura 5.7: Gráfico de dispersión circular y diagrama de rosa de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010.

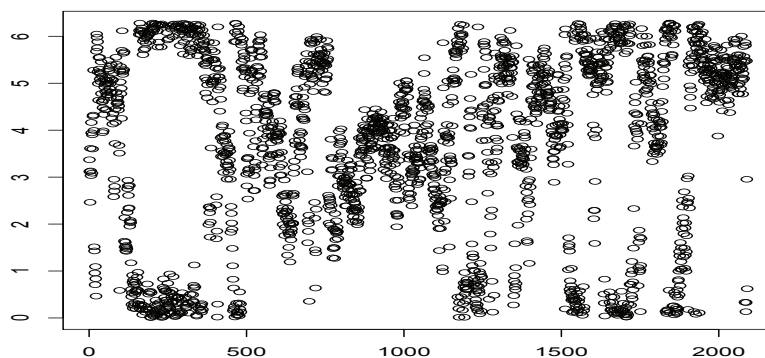


Figura 5.8: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010.

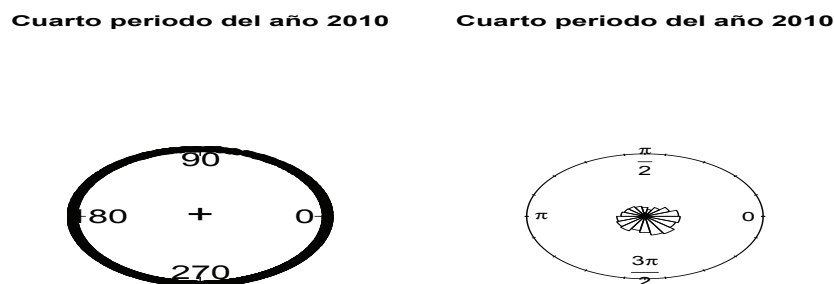


Figura 5.9: Gráfico de dispersión circular y diagrama de rosa de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010.

5.3 Estimación tipo núcleo de la función de densidad circular

En esta sección obtendremos la estimación no paramétrica tipo núcleo de la densidad circular (2.1) para cada uno de los cuatro periodos en los que hemos dividido el año 2010. Para ello hemos implementado tres diferentes elecciones del parámetro de suavizado; mediante la técnica plug-in (2.4), mediante la técnica de validación cruzada dada por la función de pérdida L2 (2.5) y mediante la técnica de validación cruzada dada por la función de pérdida Kullback-Leibler (2.6) para la correspondiente estimación no paramétrica tipo núcleo de la densidad circular en cada uno de los periodos en los que hemos dividido el año 2010.

Los parámetros de suavizado obtenidos (Tabla 5.2) mediante validación cruzada (2.5,2.6) toman valores del mismo orden pero más altos que los obtenidos mediante la técnica plug-in (2.4) en todos los periodos. Por lo que las estimaciones no paramétricas tipo núcleo de la densidad obtenidas seleccionando el parámetro de suavizado mediante la técnica plug-in van a ser más suaves que las dadas al seleccionar el parámetro de suavizado mediante validación cruzada.

	plug-in	L2	Kullback-Leibler
Primer periodo	3.10	28.56	26.20
Segundo periodo	4.87	63.57	77.17
Tercer periodo	2.17	41.44	36.17
Cuarto periodo	5.16	35.56	34.31

Tabla 5.2: Estimaciones del parámetro de suavizado para la estimación no paramétrica tipo núcleo de la función de densidad circular.

En primer lugar, al representar las estimaciones no paramétricas tipo núcleo de la densidad circular seleccionando el parámetro de suavizado mediante plug-in en los cuatro periodos; se observa que estamos ante una situación de sobresuavizado (Figuras 5.10, 5.11, 5.12, 5.13) como ya apuntaban los valores obtenidos en (5.2).

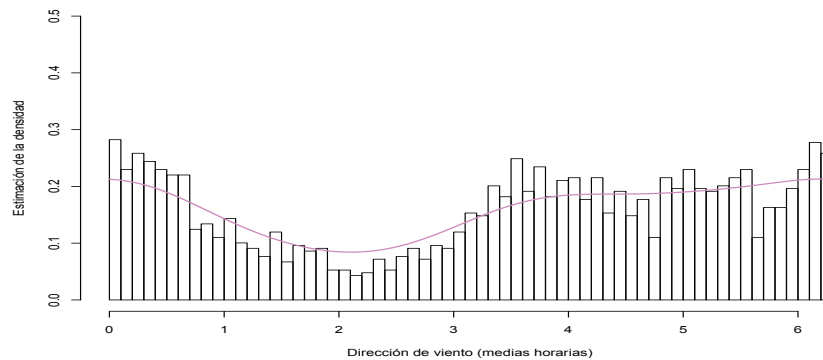


Figura 5.10: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010. Parámetro de suavizado obtenido mediante plug-in.

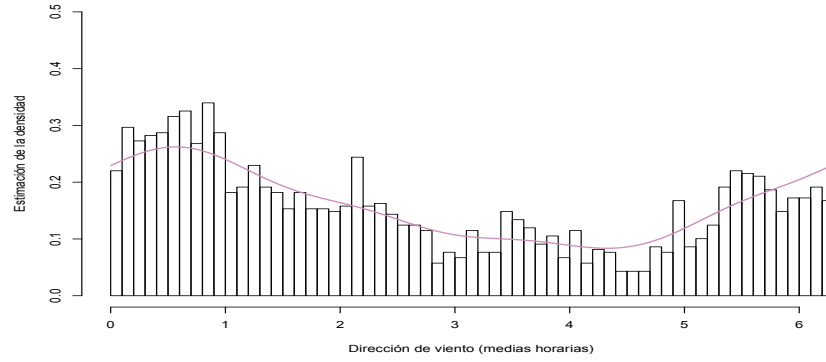


Figura 5.11: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010. Parámetro de suavizado obtenido mediante plug-in.

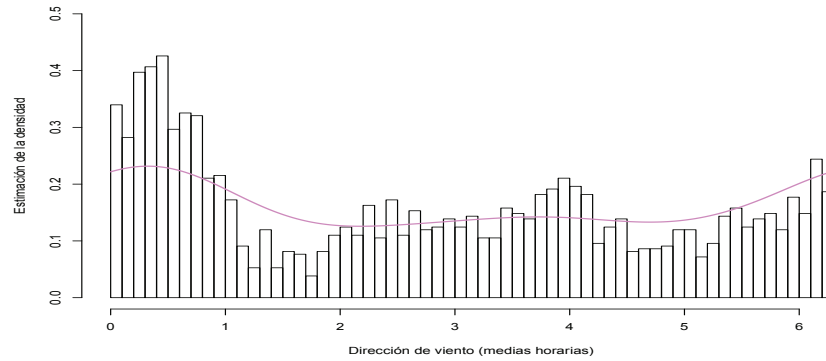


Figura 5.12: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010. Parámetro de suavizado obtenido mediante plug-in.

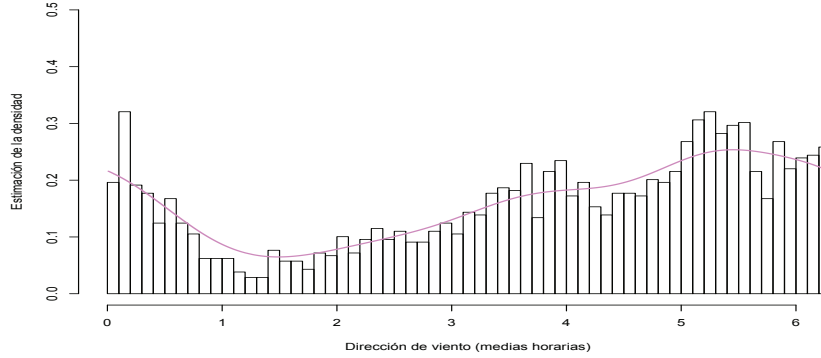


Figura 5.13: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010 . Parámetro de suavizado obtenido mediante plug-in.

A continuación, estudiamos la estimación de la densidad circular seleccionando el parámetro de suavizado mediante validación cruzada. A la vista de los gráficos de validación cruzada correspondiente a cada uno de los periodos (Figuras 5.14, 5.3, 5.3, 5.3) se observa que tanto la validación cruzada para minimizar la función pérdida dada por error cuadrático medio (L2) como la dada por la función de pérdida de Kullback-Leibler (KL) convergen, es decir, alcanzan un mínimo (2.5,2.6). Nótese que hemos implementado la validación cruzada que maximiza la función de pérdida cambiada de signo en ambos casos.

Observamos que estimando la función de densidad seleccionando el parámetro de suavizado mediante validación cruzada con cualquiera de las dos funciones de pérdida, anteriormente definidas, obtenemos la prácticamente la misma estimación a lo largo de los cuatro periodos en los que hemos dividido el año 2010. En cambio, la estimación de la densidad seleccionando el parámetro de suavizado mediante plug-in, se obtiene una estimación sobresuavizada de la densidad circular (Figuras 5.17, 5.21, 5.25, 5.29).

En definitiva, para cada uno de los periodos nos quedamos con cualquiera de las dos estimaciones de la densidad circular obteniendo el parámetro de suavizado mediante la función de validación cruzada para minimizar la función pérdida dada por error cuadrático medio (Figuras 5.15, 5.19, 5.23, 5.27) o la dada por la función de pérdida dada por Kullback-Leibler (Figuras 5.16, 5.20, 5.24, 5.28).

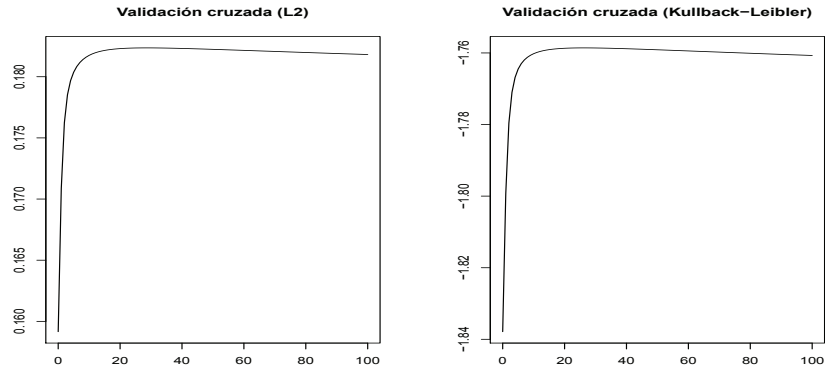


Figura 5.14: Validación cruzada del primer periodo de 2010

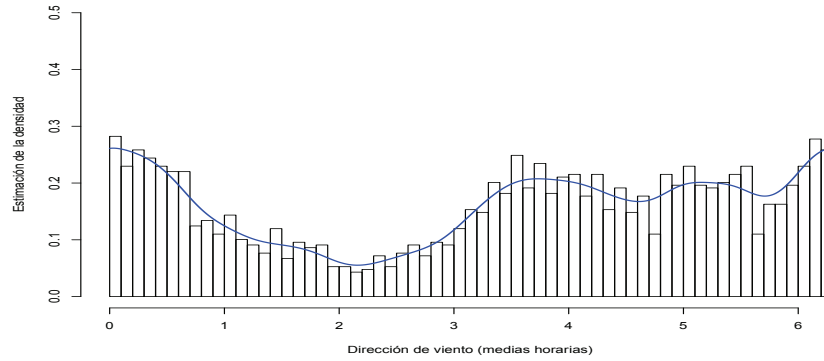


Figura 5.15: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida L2.

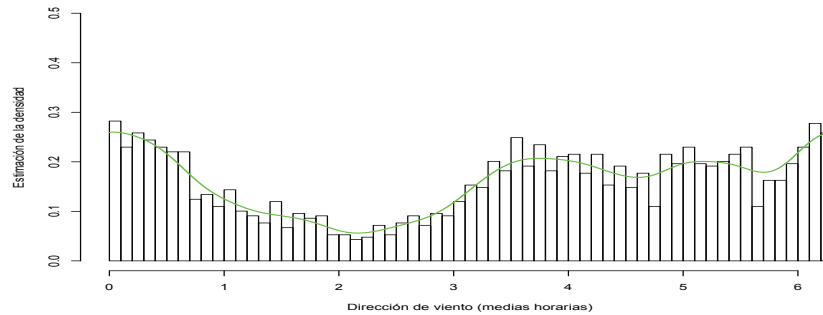


Figura 5.16: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida Kullback-Leibler.

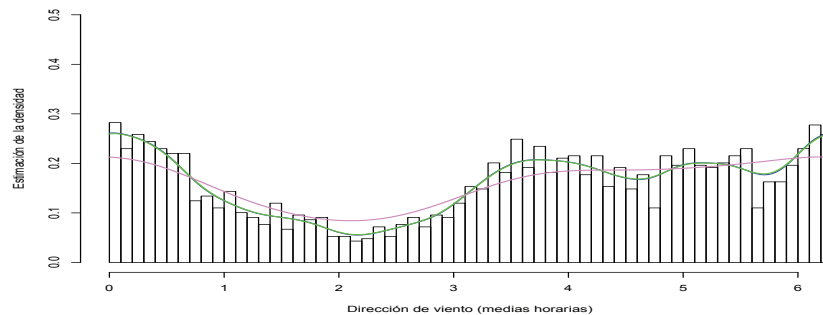


Figura 5.17: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L2 (azul) y por la función de pérdida Kullback-Leibler (verde).

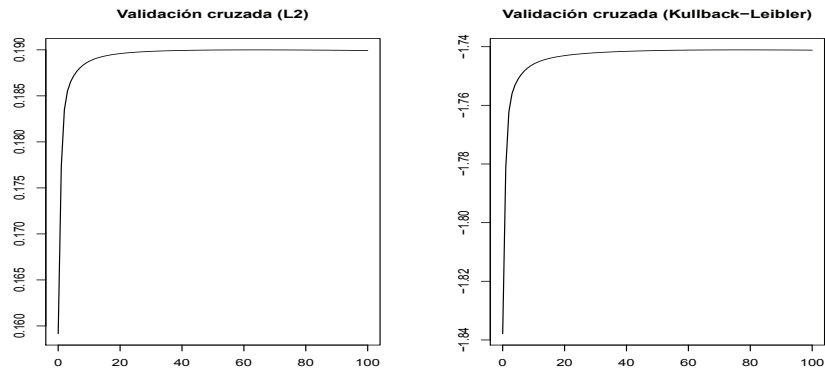


Figura 5.18: Validación cruzada del segundo periodo de 2010

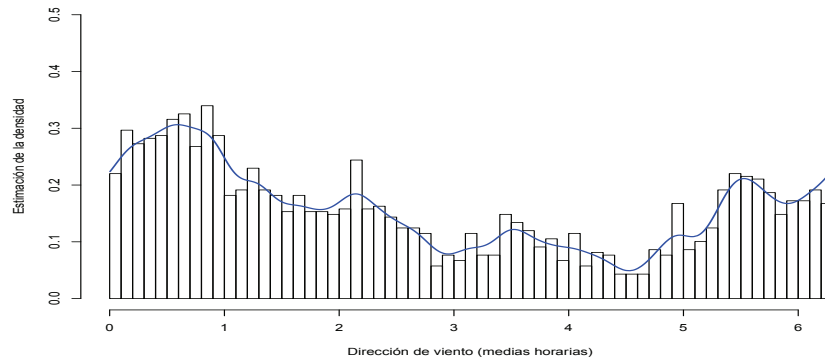


Figura 5.19: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida L2.

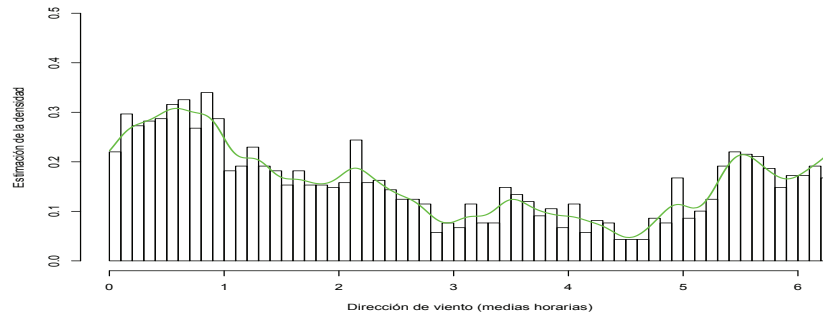


Figura 5.20: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida Kullback-Leibler.

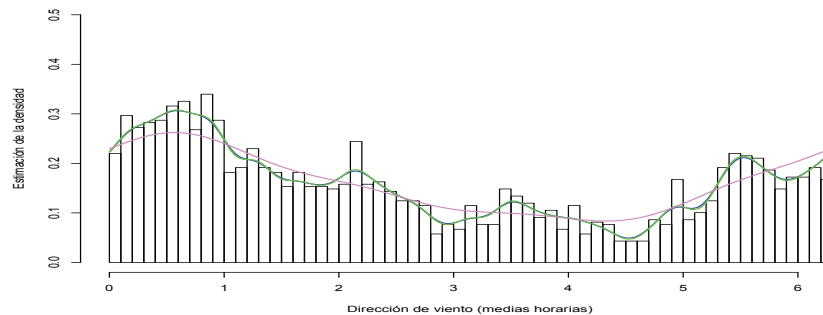


Figura 5.21: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L2 (azul) y por la función de pérdida Kullback-Leibler (verde).

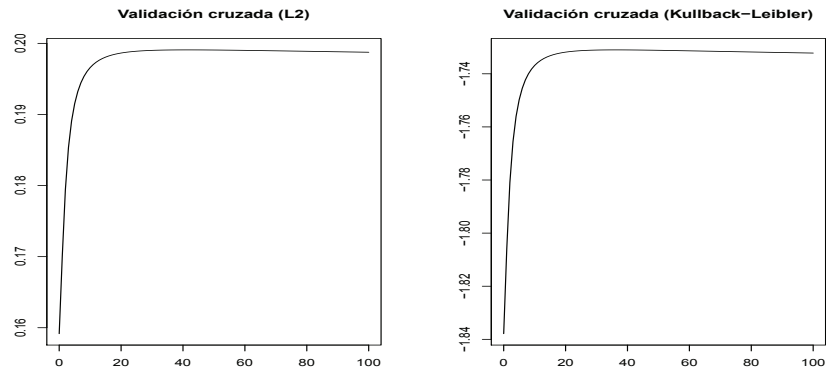


Figura 5.22: Validación cruzada del tercer periodo de 2010

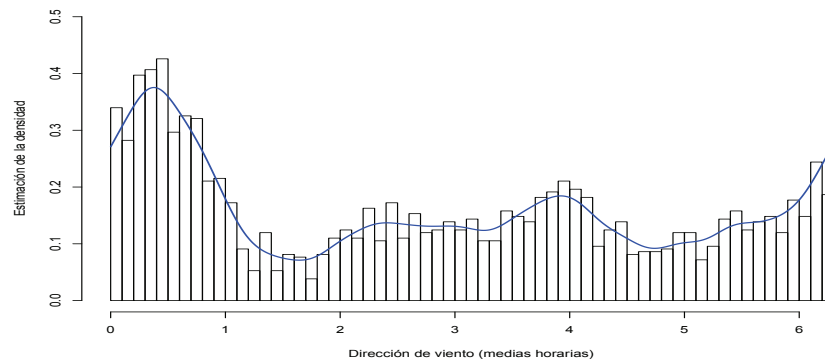


Figura 5.23: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida L2.

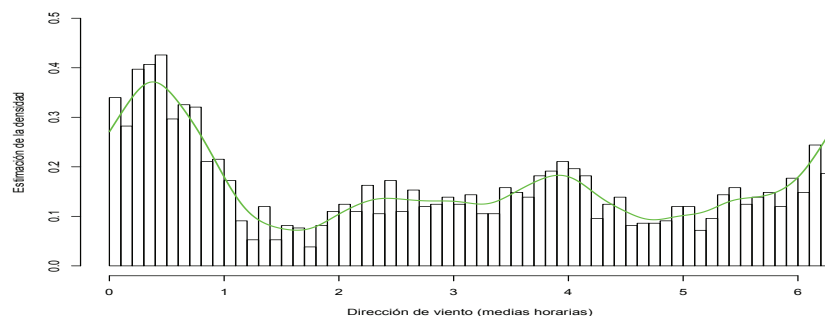


Figura 5.24: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida Kullback-Leibler.

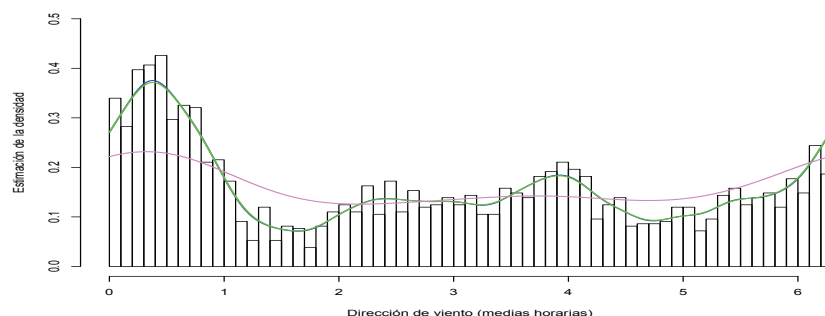


Figura 5.25: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L2 (azul) y por la función de pérdida Kullback-Leibler (verde).

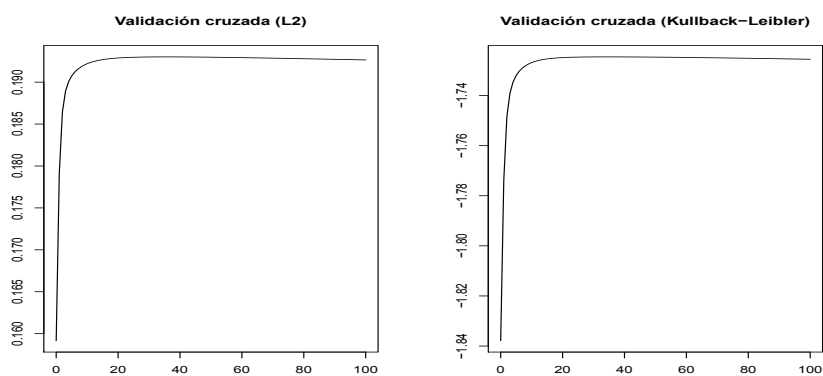


Figura 5.26: Validación cruzada del cuarto periodo de 2010

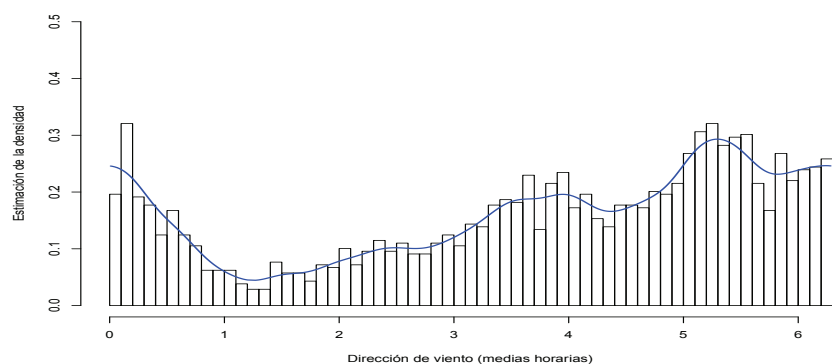


Figura 5.27: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida L2.

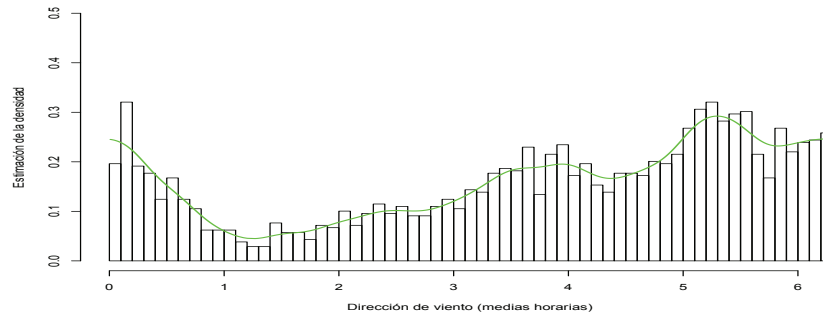


Figura 5.28: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010. Parámetro de suavizado obtenido mediante validación cruzada dada por la función de pérdida Kullback-Leibler.

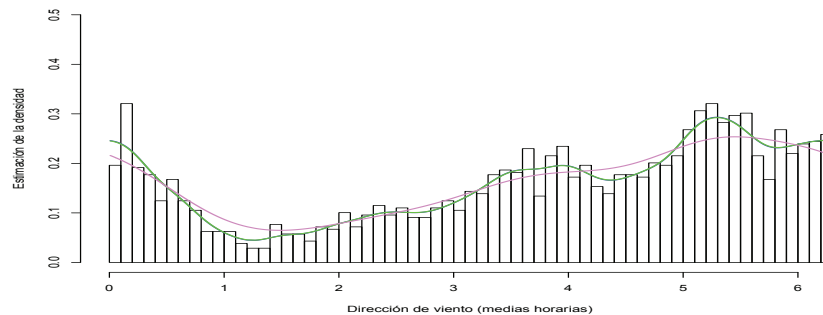


Figura 5.29: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L2 (azul) y por la función de pérdida Kullback-Leibler (verde).

5.4 Estimación del modelo de Möbius de series de tiempo

En esta sección realizaremos la estimación del modelo de Möbius de series de tiempo (4.4) para los datos horarios perturbados de la dirección de viento recogida en la estación meteorológica A Mourela. Comenzamos esta sección realizando un estudio de la dependencia en los datos de dirección de viento. Para ello estudiamos el coeficiente de autocorrelación basado en la estimación de la correlación circular entre dos variables aleatorias Θ, Φ introducido por Fisher y Lee (1983).

$$\rho_T = \frac{\mathbf{E} [\text{sen}(\Theta_1 - \Theta_2) \text{sen}(\Phi_1 - \Phi_2)]}{\{\mathbf{E} [\text{sen}^2(\Theta_1 - \Theta_2)] \mathbf{E} [\text{sen}^2(\Phi_1 - \Phi_2)]\}^{1/2}}$$

Si tomamos (Θ_1, Φ_1) y (Θ_2, Φ_2) independientes e uniformemente distribuídas como (Θ, Φ) obtenemos el estimador de ρ_T

$$\hat{\rho}_T^k = \frac{\sum_{1 \leq i < j \leq n} \text{sen}(\theta_i - \theta_j) \text{sen}(\phi_i - \phi_j)}{\left[\sum_{1 \leq i < j \leq n} \text{sen}^2(\theta_i - \theta_j) \sum_{1 \leq i < j \leq n} \text{sen}^2(\phi_i - \phi_j) \right]^{1/2}}$$

Por conveniencia tomamos $\phi_i = \theta_{i+k}$, $i = 1, \dots, n-k$ luego tenemos $(n-k)$ pares

$$(\phi_1, \theta_1), (\phi_2, \theta_2), \dots, (\phi_{n-k}, \theta_{n-k})$$

Por tanto, podemos calcular la autocorrelación circular k-lag $\hat{\rho}_T^k$ como

$$\hat{\rho}_T^k = \frac{\sum_{1 \leq i < j \leq n-k} \text{sen}(\theta_i - \theta_j) \text{sen}(\phi_i - \phi_j)}{\left[\sum_{1 \leq i < j \leq n-k} \text{sen}^2(\theta_i - \theta_j) \sum_{1 \leq i < j \leq n-k} \text{sen}^2(\phi_i - \phi_j) \right]^{1/2}}$$

Luego si representamos $\hat{\rho}_T^k$ frente a k obtendremos el correlograma circular (Figuras 5.30, 5.32, 5.34, 5.36). En estos gráficos se observa que las medias horarias de las direcciones de viento correspondientes primer, segundo y cuarto periodo son muy dependientes, respectivamente. En cambio las medias horarias de las direcciones de viento en el tercer periodo presentan una dependencia no tan destacada.

Al construir el modelo de Möbius de series de tiempo (4.4) hemos tenido que obtener los parámetros α , ω , κ (Tabla 5.3) dónde $\omega \in [-1, 1]$ y $-\pi < \alpha < \pi$. El modelo de series de tiempo obtenido en cada uno de los periodos modeliza bien nuestros datos de dirección de viento (Figuras 5.31, 5.33, 5.35, 5.4).

Se debe tener en cuenta que las fuertes subidas y bajadas que presenta el modelo de Möbius de series de tiempo, en los cuatro periodos, son debidas a cuando pasamos de valores entorno al cero a valores entorno a 2π . Obviamente, en el entorno de datos circulares, ambos valores representan a la misma observación.

	α	ω	κ
Primer periodo	-0.7492	0.9634	4.12
Segundo periodo	0.5133	0.9296	4.09
Tercer periodo	-0.1431	0.9644	4.11
Cuarto periodo	-1.3190	0.9382	4.11

Tabla 5.3: Estimaciones de los parámetros que están involucrados en la obtención del modelo de Möbius de series de tiempo

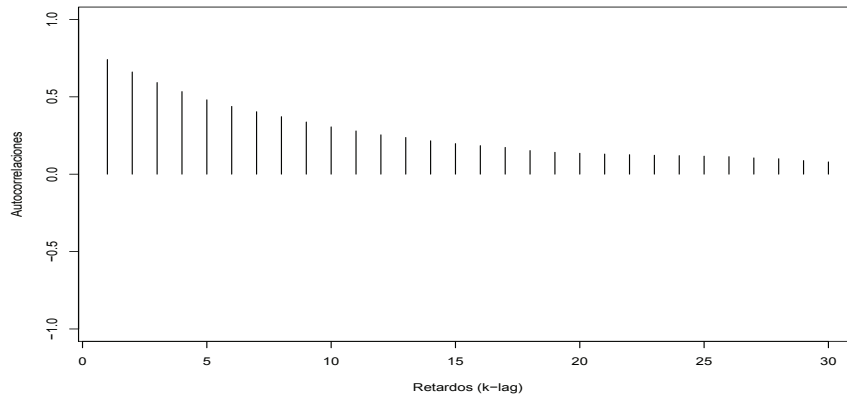


Figura 5.30: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010.

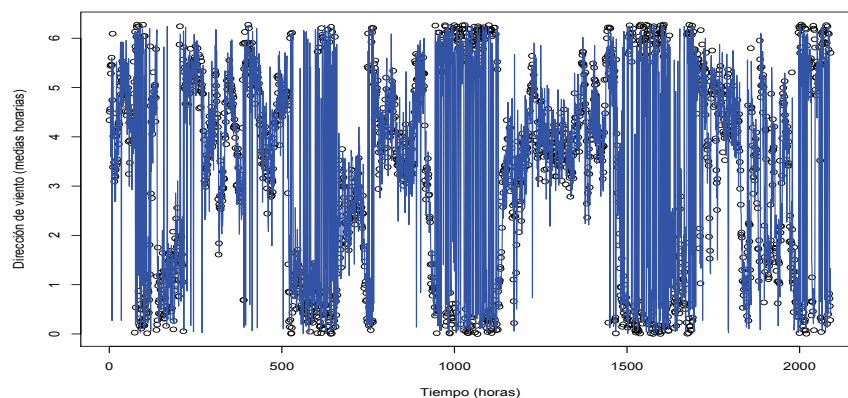


Figura 5.31: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010

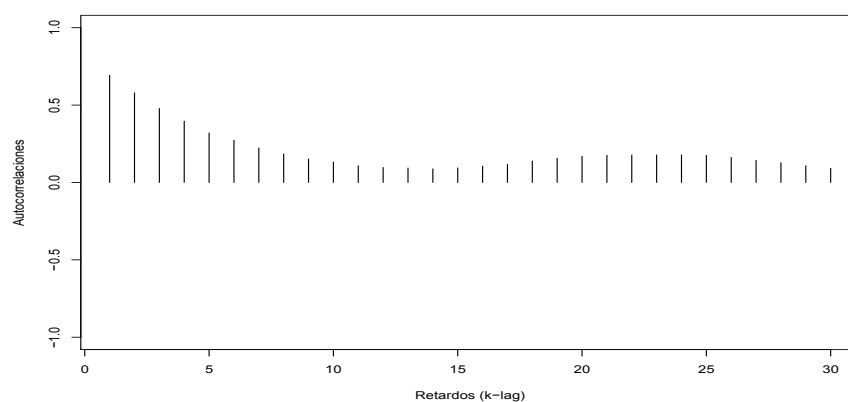


Figura 5.32: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010.

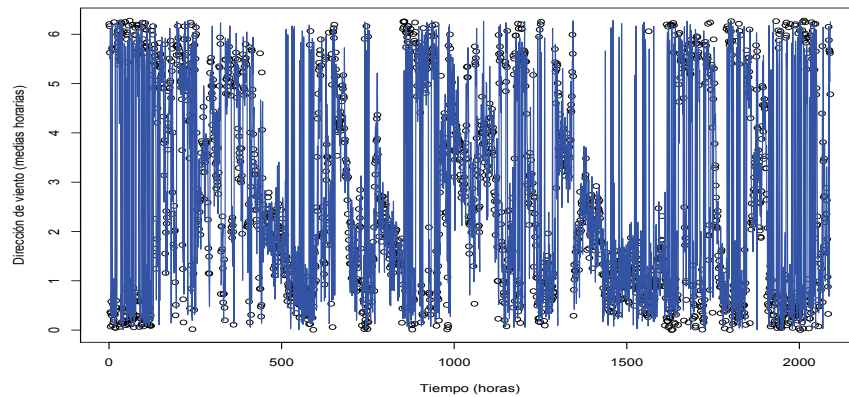


Figura 5.33: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010

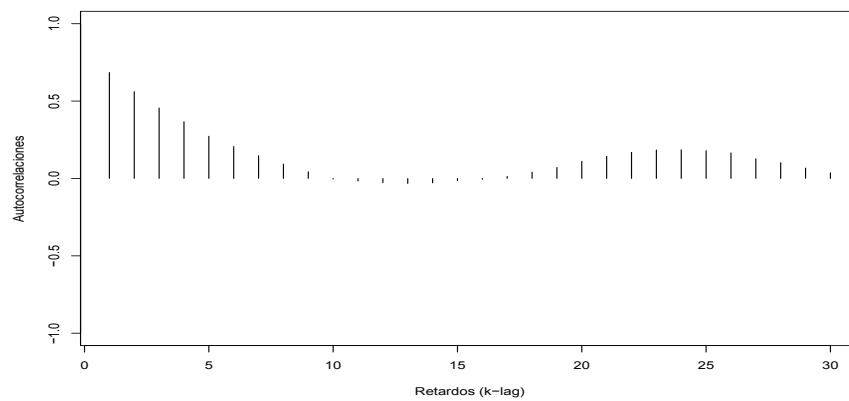


Figura 5.34: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010.

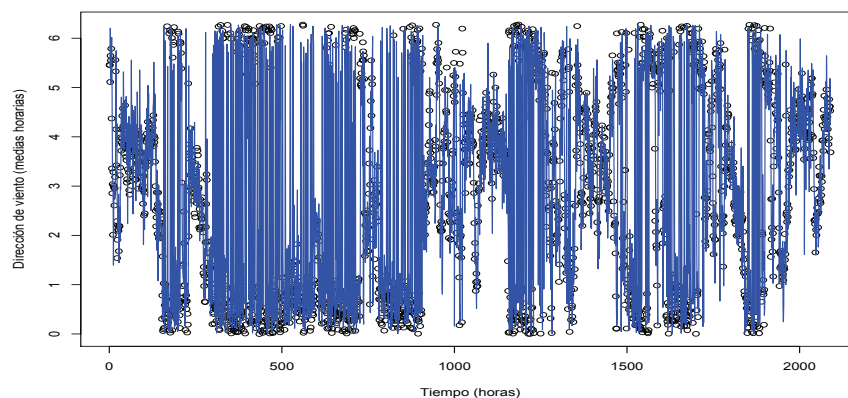


Figura 5.35: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010

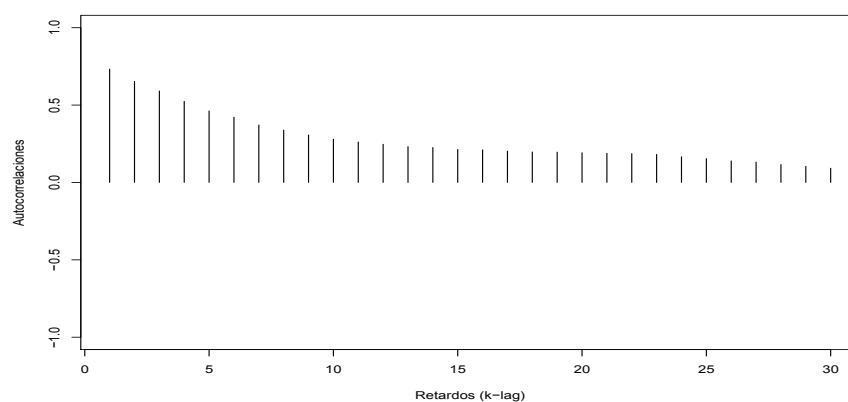


Figura 5.36: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010.

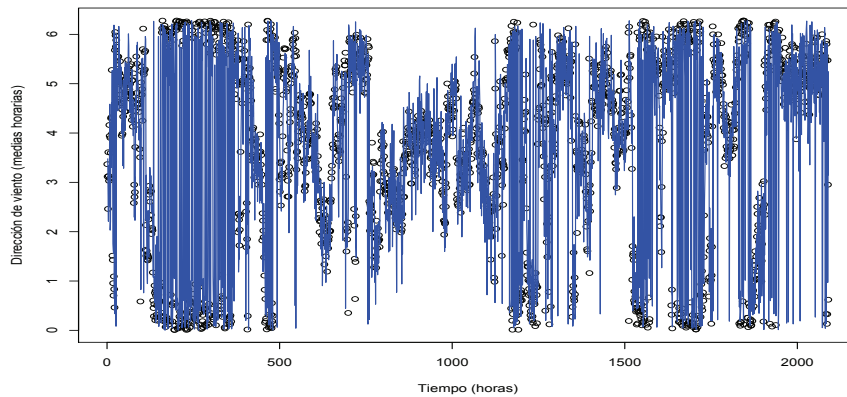


Figura 5.37: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010

5.5 Estimación tipo núcleo de la función de regresión circular-lineal

En esta sección vamos a realizar la estimación de la regresión de las media horarias transformadas mediante la función logaritmo de las concentraciones de SO_2 frente a las direcciones horarias de viento en las 5 estaciones de medición automática, B1, B2, C9, F2 y G2. Notemos que estas cinco estaciones que pertenecen a la Red de Vigilancia y Control de la Calidad Atmosférica la cual forma parte del Sistema de Control Suplementario de la Contaminación Atmosférica propiedad de la U.P.T. de As Pontes.

Como en estas estaciones se mide tanto la dirección de viento como las concentraciones de SO_2 , las utilizaremos para estudiar la relación entre las concentraciones de SO_2 y la dirección de viento mediante la estimación no paramétrica tipo núcleo de la función de regresión. Pues para realizar esta estimación necesitamos que tanto los datos de SO_2 y dirección de viento estén recogidos por la misma estación.

Vamos a utilizar en esta sección los datos dados por estas cinco estaciones que pertenecen a la Red de Vigilancia y Control de la Calidad Atmosférica la cual forma parte del Sistema de Control Suplementario de la Contaminación Atmosférica propiedad de la U.P.T. de As Pontes pues para realizar la estimación no paramétrica tipo núcleo de la función de regresión necesitamos que tanto los datos de SO_2 y dirección de viento estén recogidos por la misma estación.

En (Figura 5.38) observamos que la estación de medición B1 y F2 denominadas Magdalena y Fraga Redonda, respectivamente, se encuentran en el municipio de As Pontes, B2 denominada Louseiras se encuentra en el municipio de Muras, C9 denominada Mouranza se encuentra en el municipio de Vilalba y G2 denominada Vilanova se encuentra en el municipio de San Sadurniño.

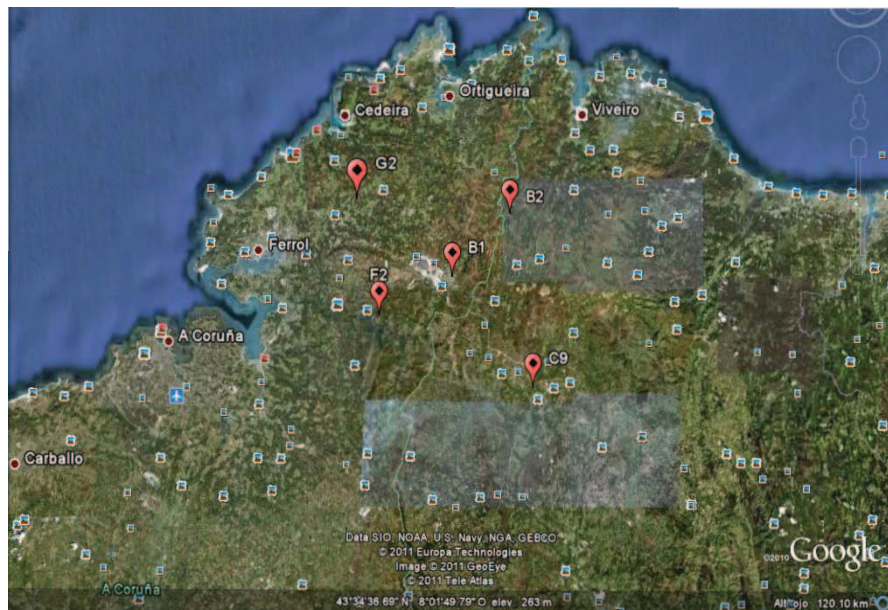


Figura 5.38: Estaciones de medición automática B1, B2, C9, F2, G2

Comenzamos analizando los valores que hemos obtenido para el parámetro de suavizado en la estimación de la función de regresión circular. En primer lugar, empleando validación cruzada de mínimos cuadrados (3.3) se han obtenido, en general, unos parámetros de suavizado razonables (Tabla 5.4).

	B1	B2	C9	F2	G2
Primer periodo	3.33	2.87	30.28	37.69	26.42
Segundo periodo	13.07	8.83	7.07	10.73	20.13
Tercer periodo	6.61	25.56	2.33	11.97	18.43
Cuarto periodo	0.00	22.63	4.87	11.24	17.74

Tabla 5.4: Estimaciones del parámetro de suavizado obtenidas mediante validación cruzada

Por otra parte, hemos obtenido el parámetro de suavizado mediante la expresión dada por (3.2). En este caso los parámetros de suavizado obtenidos son pequeños (Tabla 5.5), por lo que, se concluye que estamos ante una situación de sobre-suavizado (Figuras 5.46, 5.49, 5.52, 5.55, 5.58).

	B1	B2	C9	F2	G2
Primer periodo	4.81	6.28	5.95	4.73	5.12
Segundo periodo	5.30	6.74	6.41	4.14	5.21
Tercer periodo	5.08	6.81	6.13	3.85	4.75
Cuarto periodo	4.51	6.77	5.60	4.51	4.91

Tabla 5.5: Estimaciones del parámetro de suavizado obtenidas mediante la expresión $\kappa = 1/h_s^2$

Observamos que las medias horarias de SO_2 transformadas mediante la función logaritmo están muy concentrados entorno al uno, es decir, que las medias horarias de SO_2 originales están concentradas entorno al 3 lo que se observa en (Figuras 5.39, 5.40, 5.41, 5.42, 5.43).

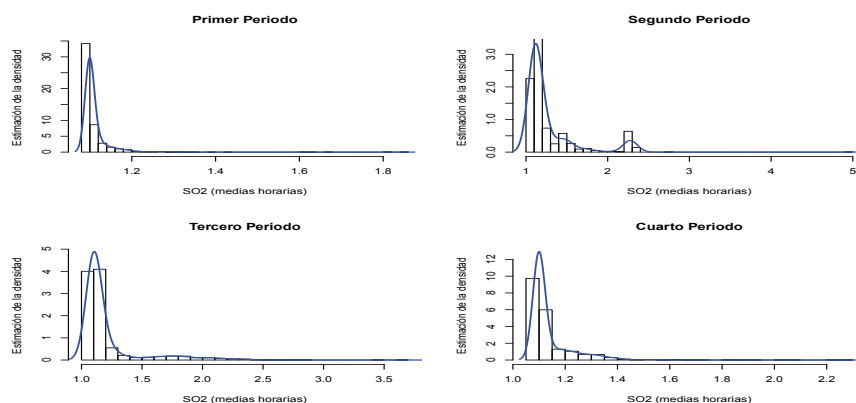


Figura 5.39: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación B1.

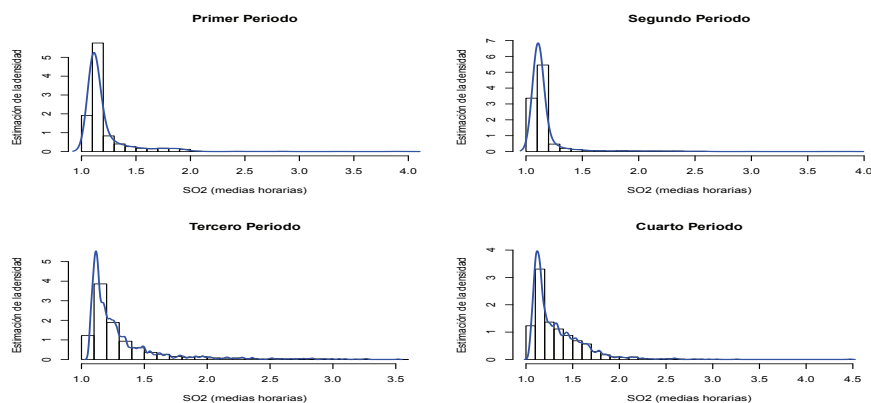


Figura 5.40: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación B2.

5.5 ESTIMACIÓN TIPO NÚCLEO DE LA FUNCIÓN DE REGRESIÓN CIRCULAR-LINEAL51

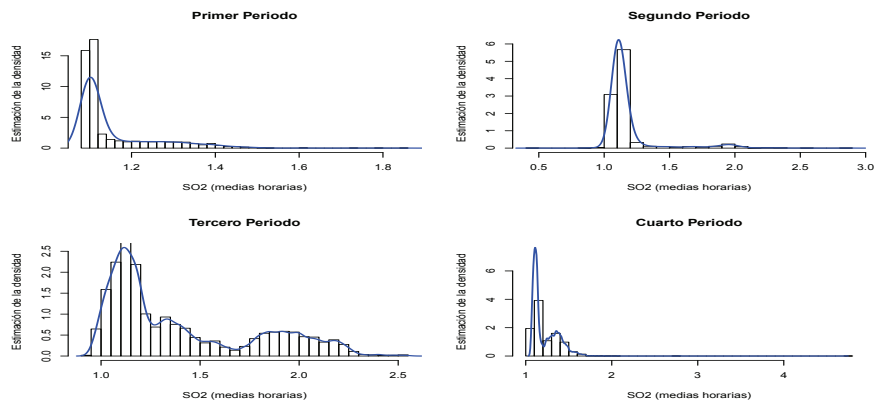


Figura 5.41: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación C9.

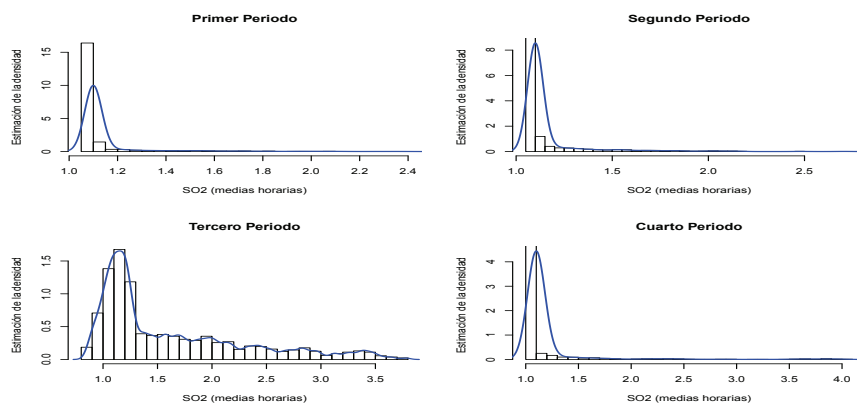


Figura 5.42: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación F2.

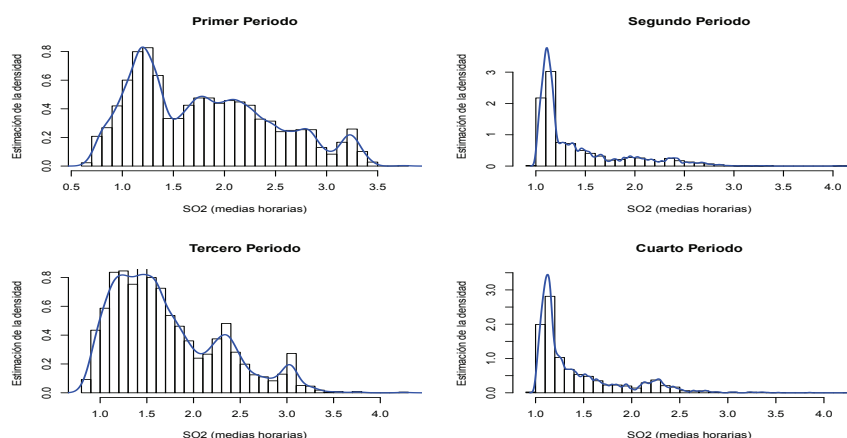


Figura 5.43: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la G2.

A continuación, presentamos la estimación no paramétrica tipo núcleo de la función regresión en cada una de las estaciones de medición automáticas.

En todos los casos, se observa que en todas las estaciones las estimaciones no dependen de la elección del parámetro de suavizado, es decir, las estimaciones dadas con el parámetro de suavizado obtenido mediante validación cruzada (Figuras 5.45, 5.48, 5.51, 5.54, 5.57) son muy similares a las obtenidas seleccionando el parámetro de suavizado mediante la técnica plugin (Figuras 5.46, 5.49, 5.52, 5.55, 5.58). Aunque notemos que escogiendo este último las estimaciones son un poco más suaves.

De todos modos, apuntemos que las estimaciones se ven muy afectadas debido a la gran cantidad de datos de SO_2 concentrados entorno al 3. Al ser las medias horarias transformadas de SO_2 tan bajas y tan concentradas la estimación no paramétrica tipo núcleo de la función de regresión no es capaz de seguir a los datos ya que estas observaciones tienen más peso que las demás. Por ello, no se puede apreciar si hay realmente relación entre la dirección de viento y las concentraciones de SO_2 .

5.5.1 Estación B1

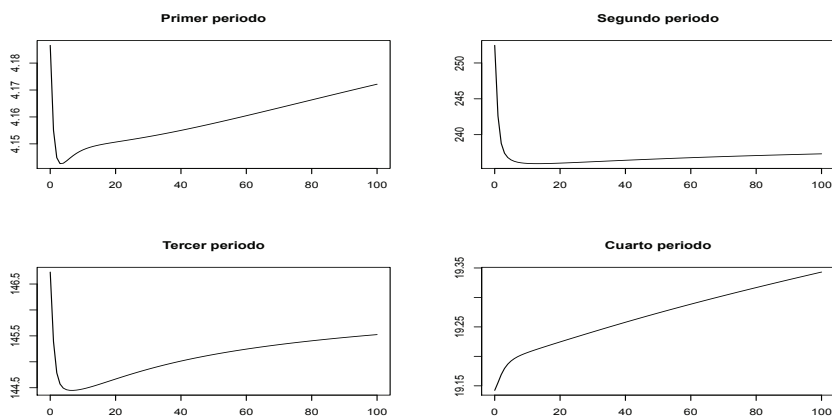
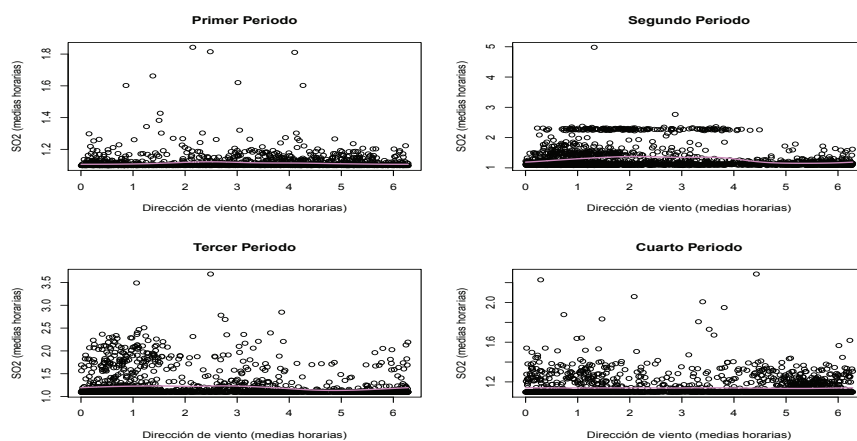


Figura 5.44: Validación cruzada en la estación B1.

Figura 5.45: Estimación de la función de regresión circular-lineal en B1. Elección del parámetro de suavizado mediante $1/h_s^2$.

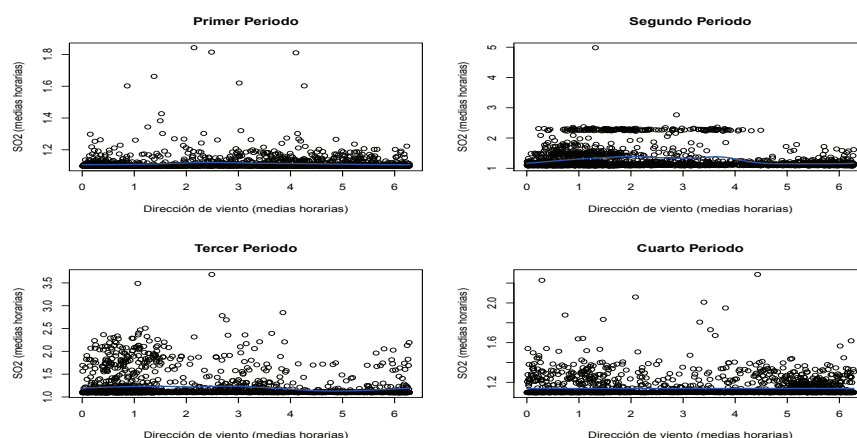


Figura 5.46: Estimación de la función de regresión circular-lineal en B1. Elección del parámetro de suavizado mediante validación cruzada.

5.5.2 Estación B2

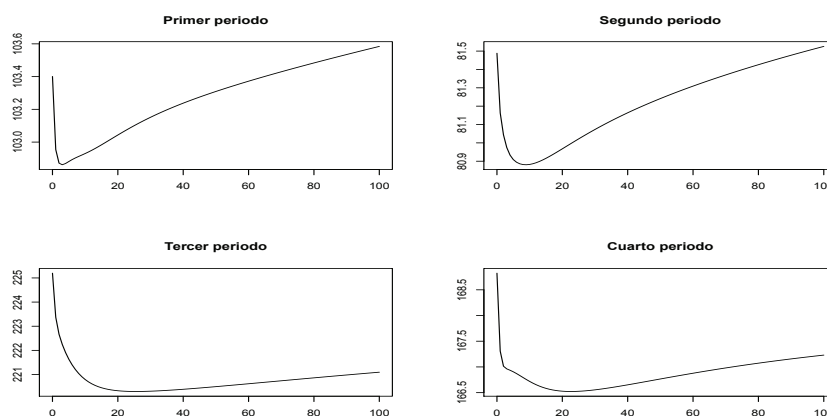


Figura 5.47: Validación cruzada en la estación B2.

5.5 ESTIMACIÓN TIPO NÚCLEO DE LA FUNCIÓN DE REGRESIÓN CIRCULAR-LINEAL55

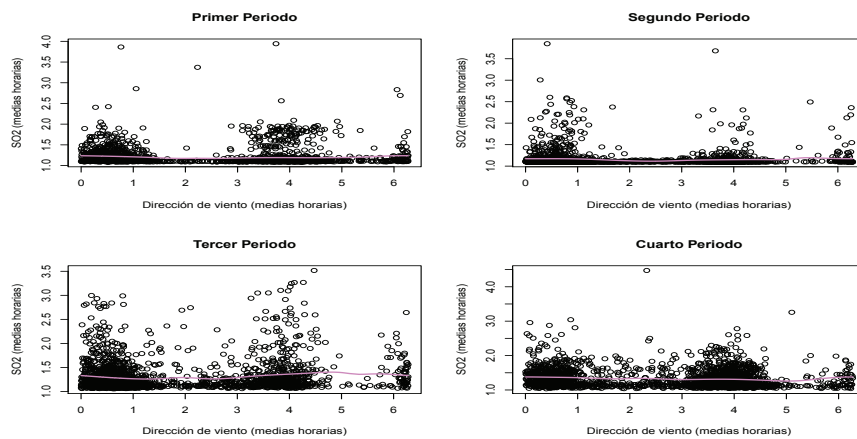


Figura 5.48: Estimación de la función de regresión circular-lineal en B2. Elección del parámetro de suavizado mediante $1/h_s^2$.

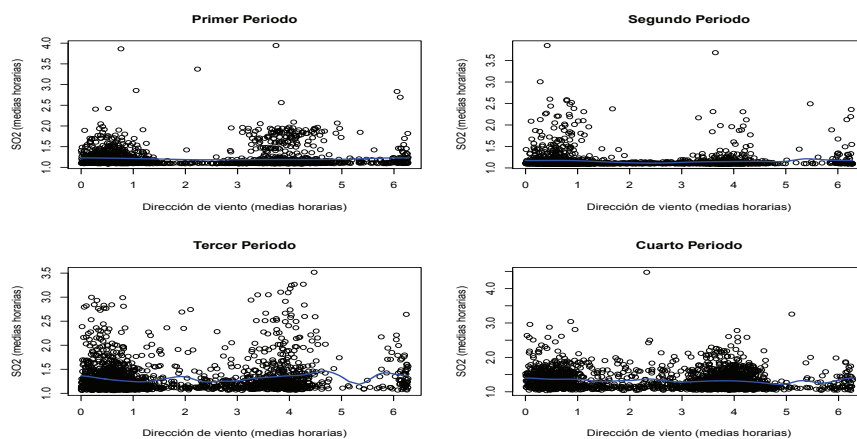


Figura 5.49: Estimación de la función de regresión circular-lineal en B2. Elección del parámetro de suavizado mediante validación cruzada.

5.5.3 Estación C9

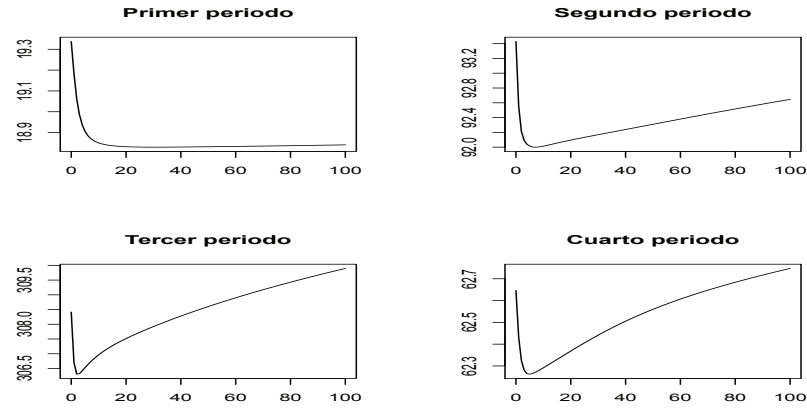


Figura 5.50: Validación cruzada en la estación C9.

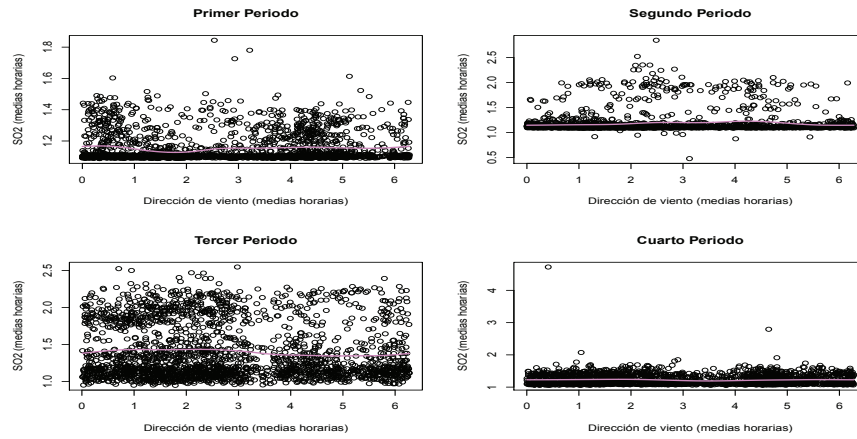


Figura 5.51: Estimación de la función de regresión circular-lineal en C9. Elección del parámetro de suavizado mediante $1/h_s^2$.

5.5 ESTIMACIÓN TIPO NÚCLEO DE LA FUNCIÓN DE REGRESIÓN CIRCULAR-LINEAL⁵⁷

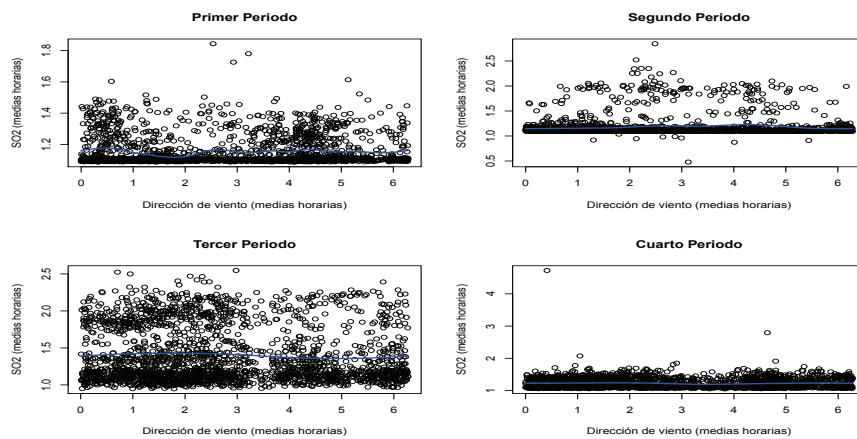


Figura 5.52: Estimación de la función de regresión circular-lineal en C9. Elección del parámetro de suavizado mediante validación cruzada.

5.5.4 Estación F2

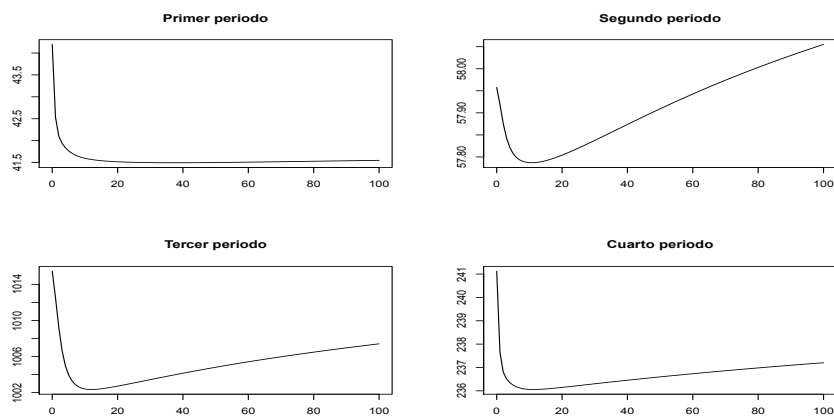


Figura 5.53: Validación cruzada en la estación F2.

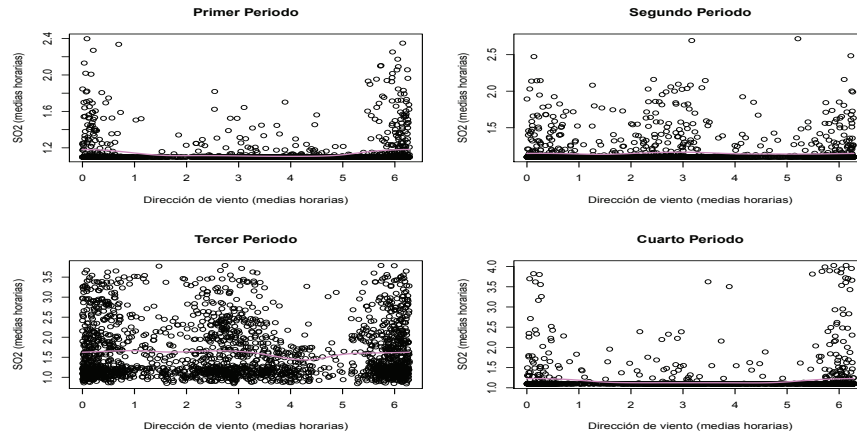


Figura 5.54: Estimación de la función de regresión circular-lineal en F2. Elección del parámetro de suavizado mediante $1/h_s^2$.

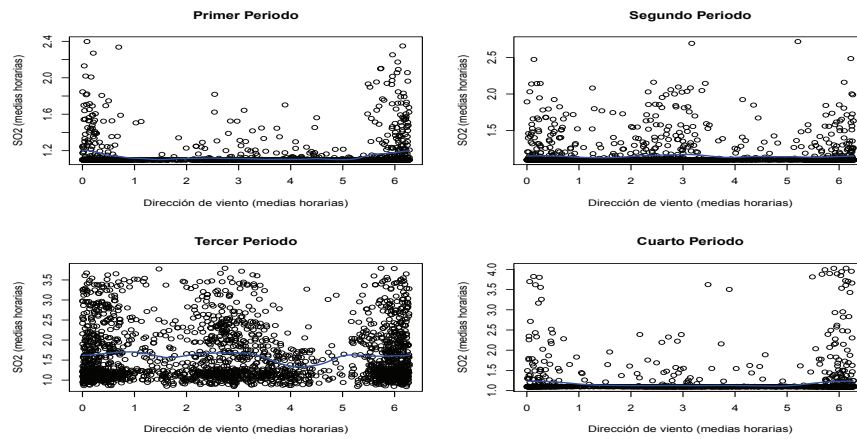


Figura 5.55: Estimación de la función de regresión circular-lineal en F2. Elección del parámetro de suavizado mediante validación cruzada.

5.5.5 Estación G2

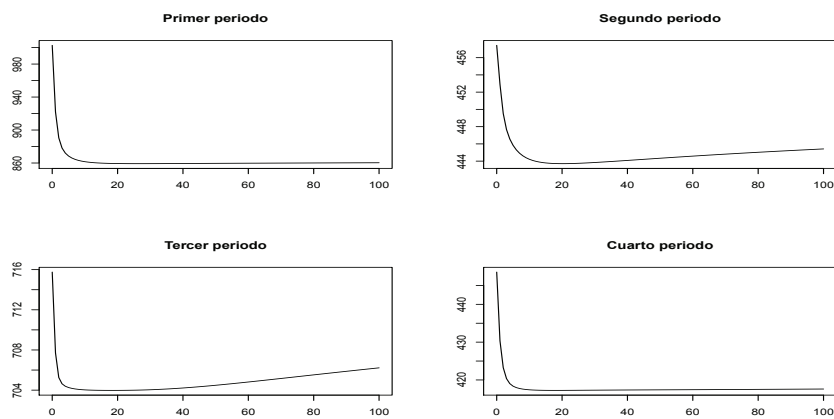
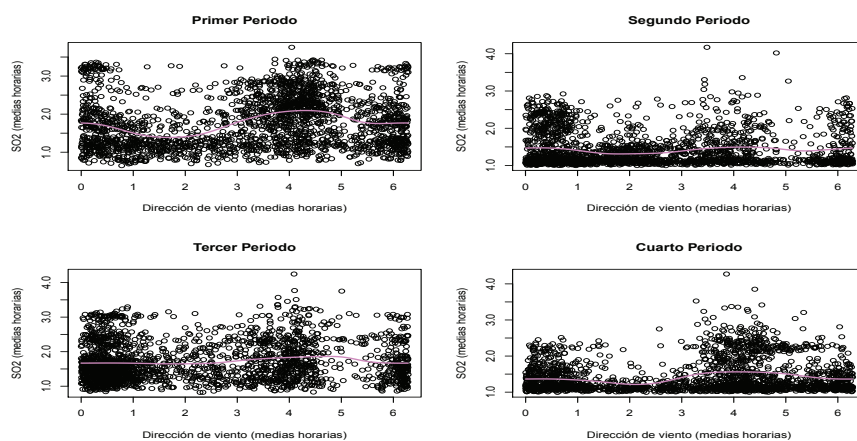


Figura 5.56: Validación cruzada en la estación G2.

Figura 5.57: Estimación de la función de regresión circular-lineal en G2. Elección del parámetro de suavizado mediante $1/h_s^2$.

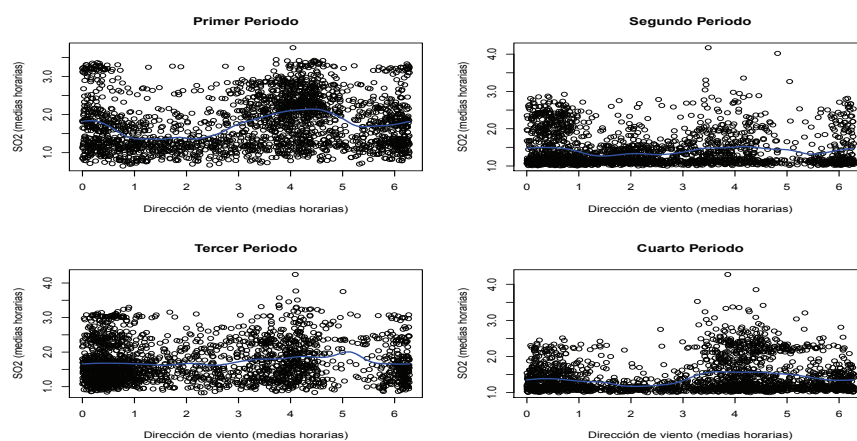


Figura 5.58: Estimación de la función de regresión circular-lineal en G2. Elección del parámetro de suavizado mediante validación cruzada.

Bibliografía

- [1] A. Azzalini. A note on the estimation of a distribution function and quantiles by a kernel method. *Biometrika*, 68:326–328, 1981.
- [2] Panzera A. Di Marzio, M. and C.C. Taylor. Local polynomial regression for circular predictors. *Statistics and Probability Letters*, 79:2066–2075, 2009.
- [3] T. D. Downs and K. V. Mardia. Circular regression. *Biometrika*, 89:683–698, 2002.
- [4] Crujeiras R.M. González Manteiga W. García Portugués, E. Exploring with direction and so2 concentration by circular-linear density estimation. Technical report, Universidade de Santiago de Compostela, Departamento de Estatística e Investigación Operativa., 2011.
- [5] X. Haiyong and F. P. Schoenberg. Kernel regression of directional data with application to wind and wildfire data in los angeles country, california. Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA.
- [6] Watson G.S. Hall, P. and J. Cabrera. Kernel density estimation whitth spherical data. *Biometrika*, 74:751–762, 1987.
- [7] G. Hughes. *Multivariate and Times Series Models for Circular Data with Applications to Protein Conformational Angles*. PhD thesis, The University of Leeds, Department of Statistics., 2007.
- [8] R. Liu and L. Yang. Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, 20:661–667, 2008.
- [9] K.V. Mardia and P.E Jupp. *Directional Statistics*. Wiley; New York, 2000.
- [10] C.C. Taylor. Automatic bandwith selection for circular density estimation. *Computational Statistics and Data Analysis*, 52:3493–3500, 2008.