Análisis Estadístico de Datos Direccionales. Aplicaciones Medioambientales

Proyecto Fin de Máster

Máster en Técnicas Estadísticas

María Leyenda Rodríguez

Índice general

Capítulo 1

Introducción

Endesa Generación S.A. cuenta en el municipio de As Pontes de García Rodríguez, al noroeste de la provincia de A Coruña, con una importante Unidad de Producción Térmica (U.P.T. As Pontes) en la que se encuentran las Centrales Térmica y de Ciclo Combinado de As Pontes.

Las Centrales tienen implantado un Sistema de Control Suplementario de la Contaminación Atmosférica que incluye la adquisición de datos de calidad de aire en tiempo real, su tratamiento y la realización de operaciones específicas que nos ayuden a la reducción de emisiones.

Los datos de calidad de aire en tiempo real se adquieren mediante una Red de Vigilancia de la Calidad Atmosférica constituida por 10 estaciones automáticas, distribuidas en un radio de aproximadamente 30 km alrededor de las centrales. Además, también se dispone de una estación meteorológica central llamada Estación de A Mourela. Estas estaciones, proporcionan medidas en continuo de diversos contaminantes y variables meteorológicas.

Debido al compromiso de Endesa Generación S.A. con el medio ambiente y su interés por mejorar este sistema de predicción, se estableció una fructífera relación entre la Sección de Medio Ambiente de la U.P.T. As Pontes y el Departamento de Estadística e Investigación Operativa de la Universidad de Santiago de Compostela.

De esta relación nace el Sistema de Predicción Estadística de Inmisión (SIPEI) que permite obtener predicciones de los valores de dióxido de azufre y de óxidos de nitrógeno, con media hora de antelación, usando modelos aditivos. Además, este sistema también predice cuál es el origen del episodio de alteración de calidad de aire, ya que este puede ser causado por la Central Térmica, el Ciclo Combinado u otros posibles focos como por ejemplo el tráfico o las actividades agrícolas de la zona.

En este trabajo nos hemos propuesto analizar la dirección de viento medida en la estación meteorólogica A Mourela a 80 metros, considerando la naturaleza angular de esta variable. Por ello, en este trabajo, abordaremos el análisis descriptivo de los datos direccionales así como en la estimación de curvas de este tipo de datos; pues realizaremos la estimación no paramétrica tipo núcleo de la densidad. Además, estudiaremos el modelo de Möebius de series de tiempo ya que las direcciones de viento son datos minutales rocogidos a lo largo del tiempo.

Debido al interés por estudiar la relación entre las concentraciones de dióxido de azufre (SO_2) y la dirección de viento utilizaremos la estimación no paramétrica tipo núcleo de la función de regresión circular-lineal. Hay que destacar que en este apartado se va a trabajar con la dirección de viento recogida en cinco de las estaciones de medida automáticas que forman parte de la Red de Vigilancia de la Calidad Atmosférica de la U.P.T de As Pontes; donde además de medir la dirección de viento, se registran las concentraciones de SO_2. Realizaremos este estudio con el objetivo de que nos ayude a determinar si el episodio tiene como origen la central térmica o no; es decir, se quiere observar si la presencia de SO_2 es debida a las emisiones de la central y para ello vamos a considerar si el viento va en esa dirección o no.

En cambio, a la hora de realizar la estimación no paramétrica de la regresión circular-lineal fue imposible utilizar los datos de dirección de viento ya que en esta estación no se recogen las concentraciones de SO_2. Por ello, nos hemos visto obligados a realizar dicha estimación en las estaciones de medida automáticas que recogen tanto la dirección de viento como las concentraciones de SO_2.

En la Figura 1.1 se puede apreciar la localización de las cinco estaciones de medida automáticas así como la localización de la estación meteorológica. Así pues, tanto las estaciones de medición B1 y F2 denominadas Magdalena y Fraga Redonda, respectivamente, como la estación de A Mourela 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.

El documento se estructura de la siguiente forma en el Capítulo2 se introducen los datos direccionales, en el Capítulo 3 se estudia la estimación no paramétrica tipo núcleo de la densidad circular haciendo un especial hincapié en la selección del

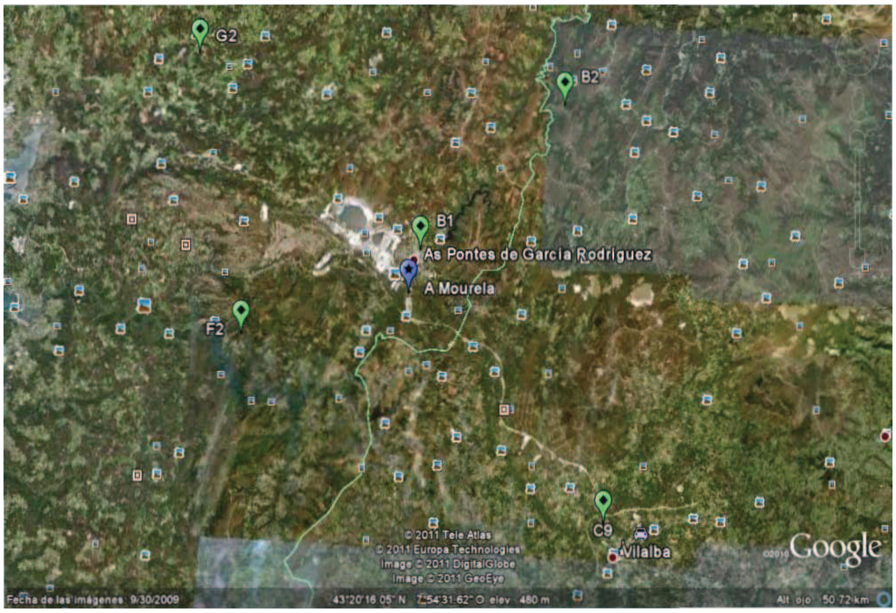

Figura 1.1: Estaciones de medida automáticas (verde). Estación Meteorológica A Mourela (azul)

parámetro de suavizado, en el Capítulo 4 se estudia la estimación no paramétrica tipo núcleo de la función de regresión circular-lineal donde la elección del parámetro de suavizado también será de vital importancia y en el Capítulo 5 se estudia el modelo de Möbius de series de tiempo. Finalmente en el Capítulo 6 se aplican las técnicas anteriores a los datos proporcionados por la Sección de Medio Ambiente de la U.P.T. As Pontes a quienes agradecemos su amabilidad por proporcionarnos los datos y toda la información necesaria para la elaboración de este proyecto.

Capítulo 2

Resumen estadístico de datos direccionales

En diversos campos surgen problemas estadísticos donde los datos son recogidos mediante medidas angulares dando la orientación o ángulos en el plano (datos circulares) o en el espacio (datos esféricos). Los datos circulares constituyen el caso más simple de esta categoría de datos llamada datos direccionales, donde la medida no es escalar, sino que 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, ya que no tienen en cuenta la naturaleza periódica de esta clase de datos.

Los datos circulares se obtienen de diversas 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 se encuentran las direcciones de viento y las direcciones migratorias de los pájaros. Un ejemplo para el reloj son los tiempos de llegada de los pacientes a una unidad de urgencias de un hospital. Otros conjuntos de datos similares surgen al considerar las veces en un año (o veces en un mes) de un cierto evento.

En este capítulo, presentaremos la descriptiva de los datos circulares siguiendo Mardia y Jupp (2000). Comenzaremos estudiando las medidas de localización, concentración y dispersión, Secciones 2.1., 2.2. y finalmente nos centraremos en las principales distribuciones circulares, Sección 2.3.

2.1. Medidas de localización

Las direcciones en el plano se pueden observar como vectores unitarios en el plano o 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 direccción y orientación inicial. Esto sería equivalente a escoger un sistema de coordenadas ortogonal en el plano. Por tanto, cada punto **x** en el círculo unidad puede ser representado por un ángulo θ o por un número complejo unitario z.

$$\mathbf{x} = (\cos\theta, \operatorname{sen}\theta);$$
$$z = e^{i\theta} = \cos\theta + i\operatorname{sen}\theta.$$

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

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

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

Supongamos que $x_1, \ldots, x_n$ es un vector no nulo, en cuyo caso, la longitud media resultante,

$$\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}, \tag{2.1}$$

es mayor que cero, luego $\bar{\theta}$ es la solución de las siguientes ecuaciones

$$\bar{C} = \bar{R}\cos\bar{\theta},\ \ \bar{S} = \bar{R}\operatorname{sen}\bar{\theta}.$$

Por otra parte, la mediana muestral de la dirección $\tilde{\theta}$ de los ángulos $(\theta_1, \ldots, \theta_n)$ es todo 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.

2.2. Medidas de concentración y dispersión

La media de longitud resultante $\bar{R}$ introducida en (2.1) es la medida de dispersión más importante en datos direccionales. Sean $x_1, \ldots, x_n$ vectores unitarios, entonces es claro que $0 \leq \bar{R} \leq 1$; de lo que se deduce que:

- Si las direcciones $\theta_1, \ldots, \theta_n$ están estrechamente agrupadas luego $\bar{R} = 1$.
- Si $\theta_1, \ldots, \theta_n$ 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 $x_1 + \ldots + x_n$.

A veces emplearemos otras medidas de dispersión análogas a las de los datos en la recta real. La medida más simple es la varianza circular muestral,

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

o la desviación circular estándard,

$$v = \sqrt{-2 \log \bar{R}}, \tag{2.2}$$

donde log es el logaritmo neperiano.

También podemos considerar la distancia entre dos ángulos θ y ξ

$$\text{mín}(\theta - \xi, 2\pi - (\theta - \xi)) = \pi - |\pi - |\theta - \xi||$$

y definir la medida de dispersión de los ángulos $\theta_1, \ldots, \theta_n$ sobre un ángulo dado α

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

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

2.3. 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 la de un ángulo aletorio θ y su función de distribución F es definida como la función en toda línea real dada por

$$F(x) = \mathbb{P}(0 < \Theta \leq x),\ 0 \leq x \leq 2\pi$$

y

$$F(x + 2\pi) - F(x) = 1,\ -\infty < x < \infty. \tag{2.3}$$

La ecuación (2.3) solo afirma que todo arco en el círculo unidad de longitud 2π tiene probabilidad uno (este arco es la totalidad de la circumferencia en el círculo). Para $\alpha \leq \beta \leq \alpha + 2\pi$,

$$\mathbb{P}(\alpha < \Theta \leq \beta) = F(\beta) - F(\alpha) = \int_{\alpha}^{\beta} dF(x), \tag{2.4}$$

donde la integral es una integral de Lebesgue-Stieltjes. La función de distribución F es continua por la derecha y por definición

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

Notemos que aunque la función F dependa de la elección inicial, (2.4) muestra que $F(\beta) - F(\alpha)$ es independiente de esta elección, por tanto cambiar la dirección inicial es simplemente añadir una constante a F.
Si la función de distribución F es absolutamente continua tiene como función de densidad a f tal que

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

Una función f es la función de densidad de una distribución absolutamente continua si y sólo 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 $(-\infty, \infty)$.

En definitiva, 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.

2.3.1. Distribución Uniforme

La distribución uniforme es la distribución más básica en el círculo y a menudo se utiliza como modelo nulo. Esta es la única distribución en el círculo que es invariante bajo rotación y reflexión. Su función de densidad es

$$f(\theta) = \frac{1}{2\pi}. \tag{2.5}$$

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

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

es decir, es proporcional a la longitud del arco.

2.3.2. Distribución Cardioide

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

$$f(\theta) = \frac{1}{2\pi}(1 + 2\rho\cos(\theta - \mu)),\ |\rho| < \frac{1}{2}. \tag{2.6}$$

- La media de longitud resultante es ρ.
- La media de la dirección es μ.
- La distribución es simétrica y unimodal en μ (si $\rho > 0$) (ver Figura 2.1).
- Si $\rho = 0$ la distribución se reduce a la distribución uniforme (ver Figura 2.1).

Esta distribución se define para añadir flexibilidad a la distribución uniforme.

Hay que destacar que el conjunto de distribuciones Cardioides es cerrado bajo convolución,

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

es decir, que si dos ángulos independientes siguen una distribución cardioide de parámetros mu_1, μ_2, ρ_1, ρ_2 la suma de estos dos ángulos también seguirá una distribución cardioide de parámetros $\mu_1 + \mu_2$ y $\rho_1\rho_2$.

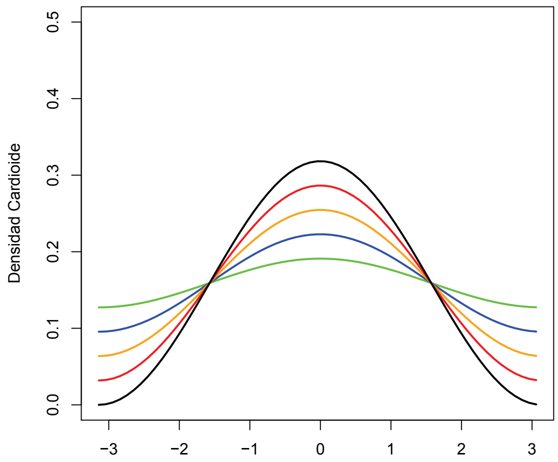

Figura 2.1: Representación de la función de densidad Cardiode C(0,ρ), ρ=0.1 (verde), ρ=0.2 (azul), ρ=0.3 (naranja), ρ=0.4 (rojo), ρ=0.5 (negro)

2.3.3. 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 siguiendo una distribución normal bidimensional con media 0 y matriz de covarianzas Σ, $\mathbb{P}(X = 0) = 0$. Luego $||x||^{-1}x$ es un punto aleatorio sobre el círculo unidad. Luego la función de densidad de la distribución normal proyectada $PN(\mu, \Sigma)$ viene dada por

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

donde $\phi(\cdot; 0, \sum)$ denota la función de densidad de $\mathbf{N}_2(0, \sum)$, ϕ y Φ denotan la función densidad y la función de distribución de N(0,1), $x = (\cos\theta, \operatorname{sen}\theta)^T$,

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

y $\mu \wedge x = \mu_1 \operatorname{sen}\theta - \mu_2\theta \ \mu = (\mu_1, \mu_2)^T$.

2.3.4. Distribución Wrapped

Dada una distribución en la recta real, 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

$$X_w = X \pmod{2\pi}.$$

Si el círculo es identificado con el conjunto de números complejos de módulo la unidad, entonces los puntos de la línea envueltos en el círculo $x \to x_w$ pueden ser escritos como

$$x \to e^{2\pi i x}.$$

Si x tiene como función de distribución F, se deduce que 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 probilidad f, entonces la función de densidad f_w de x_w es

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

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

$$\sigma^2 = -2 \log \rho,$$

es decir,

$$\rho = e^{-\sigma^2/2} \tag{2.7}$$

De (2.7), la función de densidad de $WN(\mu, \rho)$ viene dada por

$$\phi_w(\theta; \mu, \rho) = \frac{1}{2\pi} \left\{ 1 + 2 \sum_{p=1}^{\infty} \rho^{p^2} \cos p(\theta - \mu) \right\}. \tag{2.8}$$

2.3.5. Distribución von Mises

La distribución von Mises es la más utilizada en el círculo. La distribución von Mises $vM(\mu, \kappa)$ tiene como función de densidad

$$g(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta - \mu)}, \tag{2.9}$$

donde

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

$$I_0(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} e^{\kappa \cos \theta} d\theta.$$

- El parámetro μ es la media de las direcciones.
- κ es el parámetro de concentración.

Dado la importancia de la distribución von Mises le reservamos la letra g para denotar su función de densidad.

Como se puede ver en la Figura 2.2 la distribución von Mises es unimodal y es simétrica sobre $\theta = \mu$. Además la moda se encuentra en $\theta = \mu$ y la antimoda en $\theta = \mu + \pi$. El cociente entre el valor de la densidad en la moda y el valor de la densidad en la antimoda viene dado por $e^{2\kappa}$, así que cuanto mayor sea el valor de κ, mayor es el agrupamiento acerca de la moda. En esta Figura 2.2 también se observa el hecho de que cuanto más grande sea el parámetro κ su función de densidad estará más concentrada.

El parámetro κ es una medida de concentración en el caso de que nuestra muestra siga una von Mises. Así que, nos hemos planteado explicar la relación entre diferentes medidas de dispersión como son la media de longitud resultante (2.1), la desviación circular estándard (2.2), la varianza definida de forma análoga al caso lineal, que se definirá a continuación, y el parámetro de concentración de la distribución von Mises κ.

1. Generar una muestra $vM(\pi, \kappa)$ de tamaño 3000 para distintos valores del párametro κ (75 valores).
2. Para cada una de las muestras calculamos:

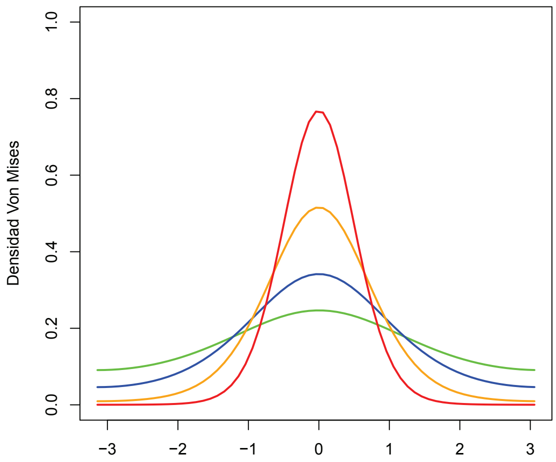

Figura 2.2: Representación de la función de densidad von Mises $vM(0,\kappa)$, κ=0.5 (verde), κ=1 (azul), κ=2 (naranja), κ=4 (rojo).

- La varianza definida de forma análoga al caso lineal,

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

  con respecto a la distancia euclídea y a la distancia definida por la longitud de arco,

$$d^2(\theta_i, \theta) = 2(1 - (\cos(\theta_i - \theta))),$$

  y a la distancia definida por la longitud de arco,

$$d^2(\theta_i, \theta) = \pi - |\pi - |\theta_i - \theta||.$$

- El parámetro de concentración $\bar{R} = \sqrt{\bar{C}^2 + \bar{S}^2}$ y
- La desviación circular estándar $\sqrt{-2\log(\bar{R})}$.

3. Representar el estimador local-lineal para estudiar las relaciones de estas medidas de concentración y dispersión, frente a diferentes valores de κ.

El la Figura 2.3 se puede observar que v, S_n^2 (obtenida mediante cualquiera de las distancias) se comportan de forma similar: disminuyen casi de forma exponencial a

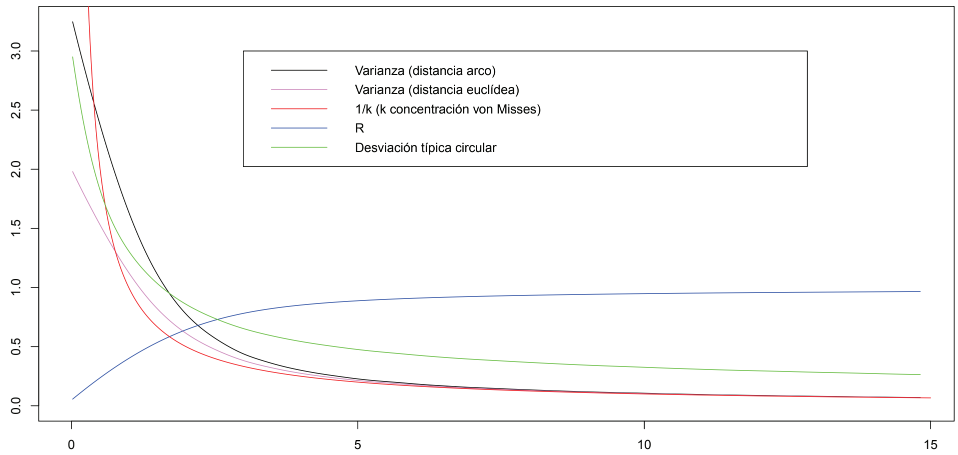

Figura 2.3: Relación entre $\bar{R}$ (azul), v (verde), S_n^2 distancia dada por la longitud de arco (negro), S_n^2 distancia euclídea (rosa) y $1/\kappa$ (rojo)

medida que κ aumenta. Mientras que $\bar{R}$ se comporta de modo diferente: aumenta a medida que lo hace κ.

La distribución von Mises juega un papel crucial en el campo de los datos direccionales. De hecho, juega el mismo papel que la distribución gausiana 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 (2.5). Además para x próximos a 0 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 (2.6). Por tanto, una distribución von Mises con parámetro de concentración pequeño puede ser aproximada por una distribución cardioide con la misma dirección media y mismo parámetro de concentración dividido por 2. Por ello, la distribución cardioide es usada, principalmente, como aproximaciones a las distribuciones von Mises de poca concentración.

- Cuando $\kappa \to \infty$ la distribución $vM(\mu, \kappa)$ está concentrada en el punto $\theta = \mu$.

- Si κ es grande, sea $\theta \sim vM(\mu, \kappa)$ (2.9) y sea $\xi = \kappa^{1/2}(\theta - \mu)$. Luego, de (2.9) se deduce que la función de distribución de ξ es 'proporcional a

$$\exp\left\{-\kappa[1-\cos(\kappa^{-1/2}\xi)]\right\}. \tag{2.10}$$

Para κ grande,

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

así pues, de (2.10), $\xi \sim N(0,1)$. Entonces para grandes valores de κ,

$$\theta \sim vM(\mu,\kappa) \Rightarrow \kappa^{-1/2}(\theta-\mu) \sim N(0,1),\ \kappa \to \infty. \tag{2.11}$$

- De foma más general, cualquier von Mises puede ser aproximada por una distribución wrapped normal (2.8)

$$vM(\mu,\kappa) \cong WN(\mu, A(\kappa)),\ \kappa \to \infty.$$
$$A_\kappa = I_1(\kappa)/I_0(\kappa).$$

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 $vM(\mu_1+\mu_2, A^{-1}(A(\kappa_1)A(\kappa_2)))$; luego

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

Capítulo 3

Estimación de la densidad circular

En la Sección 3.1., descibiremos la estimación no paramétrica tipo núcleo de la densidad circular. 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 cuadrártico medio (L_2) o minimizar la función de pérdida dada por Kullback-Leibler (ver Hall, Watson y Cabrera, 1987)

3.1. Estimación no paramétrica tipo núcleo de la densidad circular

En el caso lineal, si partimos de una muestra aleatoria simple $X_1, \ldots, X_n$, el estimador natural de la probabilidad en cada uno de los intervalos $[x_m, x_{m+1})$, con $x_m = x_0 + hm,\ m \in \mathbb{Z}$, es el histograma (3.1)

$$\hat{f}_{n,H}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\left(X_i \in [x_m, x_{m+1})\right), \tag{3.1}$$

donde $\mathbb{I}$ denota a la función indicadora.

La estimación no paramétrica de la función de densidad dada por el histograma tiene el problema de que se ve muy influenciada por la elección del punto inicial, x_0. Para solucionar este problema modificaremos esta estimación de manera que para cada punto x se construye un intervalo de la forma $(x - h, x + h)$; lo que da lugar a otra estimación no paramétrica de la densidad denominada histograma móvil.

Teniendo en cuenta que la función de densidad de una variable aleatoria X en un punto x viene dada por

$$f(x) = \lim_{x\to 0} \frac{1}{2h} \sum_{i=1}^{n} \mathbb{P}\left(x - h < X < x + h\right). \tag{3.2}$$

Podemos definir ,de manera natural, un estimador tipo núcleo mediante:

$$\hat{f}_{n,N}(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{I}\left(X_i \in (x-h, x+h)\right) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbb{I}\left(x \in (X_i - h, X_i + h)\right).$$

Este estimador es denominado Estimador Naive y se puede escribir como:

$$\hat{f}_{n,N}(x) = \frac{1}{nh} \sum_{i=1}^{n} \omega\left(\frac{x - X_i}{h}\right), \tag{3.3}$$

donde ω es la densidad de la distribución uniforme en (-1, 1). Por tanto la aportación de un dato X_i al Estimador Naive en el punto x viene determinado por el valor de

$$\omega\left(\frac{x - X_i}{h}\right)$$

que vale $\frac{1}{2}$ para todos los puntos en $(x - h, x + h)$ independientemente de su proximidad a x.

Si se reemplaza ω por una densidad K (denominada núcleo) con una única moda en cero y simétrica se obtiene el estimador tipo núcleo; también denominado como el estimador de Parzen-Rosemblat que viene dado por

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{3.4}$$

donde

- 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.

Este estimador no parametrico tipo núcleo de la densidad generaliza al histograma móvil, al promediar pesos variables que penalizan la distancia de los datos al punto x.

La estimación de la densidad tipo núcleo en el caso lineal (3.4) se extiende facilmente a datos circulares, aunque se debe tener cuidado en la selección de la función núcleo.

Cuando usamos datos en el círculo, la distancia entre dos puntos en el círculo viene dada por

$$d_i = ||x - X_i||^2 = 2(1 - x^T X_i) = 2(1 - \cos(\theta - \Theta_i)), \tag{3.5}$$

donde $x^T = (\cos\theta, \operatorname{sen}\theta)$ y $X_i = (\cos\Theta_i, \operatorname{sen}\Theta_i)^T$.

La ecuación (3.4) permite una representación alternativa de la estimación no paramétrica tipo núcleo de la densidad,

$$\hat{f}(\theta, \nu) = \frac{1}{n}\sum_{i=1}^{n} L_\nu \left(1 - x^T X_i\right) = \frac{1}{n}\sum_{i=1}^{n} L_\nu \left(1 - cos(\theta - \Theta_i)\right)$$

Dada una muestra aleatoria de ángulos $\Theta_1, \ldots, \Theta_n \in [0, 2\pi]$, el estimador no paramétrico tipo núcleo de la densidad con función núcleo la función de densidad von Mises (2.9), viene dado por

$$\hat{f}(\theta, \nu) = \frac{1}{2n\pi I_0(\nu)} \sum_{i=1}^{n} \exp(\nu \cos(\theta - \Theta_i)). \tag{3.6}$$

El parámetro de concentración ν ha asumido el papel de la inversa del parámetro de suavizado. Por tanto, 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.

3.2. Selección del parámetro de suavizado.

A continuación nos centramos en la elección del parámetro de suavizado para el estimador tipo núcleo (3.6). Existen varios métodos para seleccionar el parámetro de suavizado como pueden ser el método de validación cruzada o la regla plug-in con densidad de referencia von Mises.

3.2.1. Validación cruzada

A continuación obtendremos las funciones de validación cruzada para minimizar la función pérdida dada por el error cuadrático medio, L_2, y la dada por Kullback-Leibler. Ambas dadas por Hall, Watson y Cabrera, (1987).

$$L_2(\nu) = \int_0^{2\pi} \mathbb{E}\left(\hat{f}(\theta,\nu) - f(\theta)\right)^2 d\theta. \tag{3.7}$$

$$L_{KL}(\nu) = \int_0^{2\pi} f(\theta)\mathbb{E}\left[\log\left\{f(\theta)/\hat{f}(\theta,\nu)\right\}\right] d\theta. \tag{3.8}$$

Comenzamos obteniendo la función de validación cruzada para la función pérdida dada por el error cuadrático medio (3.7). Para ello, comenzamos obteniendo el Error Cuadrático Integrado (ISE). En este caso, se define como

$$ISE_2(\nu) = \int_0^{2\pi} \left(\hat{f}(\theta,\nu) - f(\theta)\right)^2 d\theta$$

y se pude escribir como

$$ISE_2(\nu) = \int_0^{2\pi} \hat{f}^2(\theta,\nu)d\theta - 2\int_0^{2\pi} \hat{f}(\theta,\nu)f(\theta)d\theta + \int_0^{2\pi} f^2(\theta)d\theta. \tag{3.9}$$

Luego, podemos buscar el parámetro de suavizado que haga al ISE más pequeño posible teniendo en cuenta que minimizar la ecuación (3.9) es quivalente a minimizar

$$CV_2(\nu) = \int_0^{2\pi} \hat{f}^2(\theta,\nu)d\theta - 2\int_0^{2\pi} \hat{f}(\theta,\nu)f(\theta)d\theta.$$

Estimamos esta expresión mediante

$$\hat{CV}_2(\nu) = \int_0^{2\pi} \hat{f}^2(\theta,\nu)d\theta - 2n^{-1}\sum_{i=1}^{n} \hat{f}_i(\Theta_i,\nu),$$

donde $\hat{f}_j(\cdot)$ la estimación no paramétrica tipo núcleo de la densidad construída dejando fuera el valor Θ_j de la muestra y viene dada por

$$\hat{f}_j(\theta,\nu) = \frac{1}{2n\pi I_0(\nu)}\sum_{i\neq j} \exp\left(\nu\cos(\theta-\Theta_i)\right). \tag{3.10}$$

Por tanto, el parámetro de suavizado obtenido mediante el método de validación cruzada para minimizar la función de pérdida L_2 viene dado por

$$\upsilon_2 = \arg\min_{\nu\geq 0} \hat{CV}_2(\nu). \tag{3.11}$$

Análogamente se procede con la función de pérdida dada por Kullback-Leibler y se obtiene que otra posible elección para el parámetro de suavizado del siguiente modo

$$v_{KL} = \arg\min_{\nu \geq 0} -CV_{KL}(\nu), \tag{3.12}$$

siendo

$$CV_{KL}(\nu) = n^{-1} \sum_{i=1}^{n} \log\left\{\hat{f}_i(\Theta_i, \nu)\right\},$$

donde la expresión $\hat{f}_i(\cdot)$ viene dada por la ecuación (3.10).

3.2.2. Regla plug-in con densidad de refrencia von Mises

Al igual que se hace para datos lineales es posible seleccionar el parámetro de suavizado como aquel que minimiza una versión asíntótica del Error Cuadrático Medio Integrado (MISE). Bajo la hipótesis de que es una von Mises de con parámetro de concentración ν Taylor obtuvo la expresión del Error Cuadrático Medio Integrado Asintótico (AMISE) que viene dada por

$$AMISE(\nu) = 3\kappa^2 I_2(2\kappa)/\left\{32\pi\nu^2 I_0(\kappa)^2\right\} + \nu^{1/2}/\left(2n\pi^{1/2}\right), \tag{3.13}$$

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

Por tanto, AMISE es de la forma $a\nu^{-2} + b\nu^{1/2}$ el cual puede ser minimizado derivando respecto de ν e igualando a cero. Esto permite desarrollar la regla plug-in con densidad de referencia von Mises para el parámetro de suavizado ν basado en la estimación de κ mediante máxima verosimilitud (ver Taylor, 2008):

$$\nu = [3n\hat{\kappa}^2 I_2(2\hat{\kappa})\left\{4\pi^{1/2} I_0(\hat{\kappa})^2\right\}^{-1}]^{2/5}. \tag{3.14}$$

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

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

Los resultados de simulación de Taylor (ver Taylor, 2008) muestran que la regla plug-in se comporta bien cuando la distribución que genera los datos es una von Mises. Sin embargo, estos resultados no son tan satisfactorios cuando esta hipótesis no es cierta. Por ejemplo, supongamos que la muestra de tamaño 100 está generada por una mixtura de von Mises de parámetros $\mu_1 = \pi/2$, $\kappa_1 = 7$, $\mu_2 = 3\pi/2$ y $\kappa_2 = 7$ con $\rho = 0{,}5$; cuya función de densidad viene dada por

$$f(\theta) = \rho f_1(\theta, \mu_1, \kappa_1) + (1 - \rho) f_2(\theta, \mu_2, \kappa_2) \tag{3.15}$$

y calculamos tres estimaciones no paramétricas tipo núcleo de la densidad, cada una de ellas obtenida mediante uno de los parámetros de suavizado estudiados (3.14, 3.11, 3.12). Se observa en la Figura 3.1 como las estimaciones no paramétricas tipo núcleo de la densidad obtenidas tras estimar el parámetro mediante validacíón cruzada (3.11, 3.12) son mejores que la dada por el parámetro de suavizado obtenido mediante plug-in (3.14).

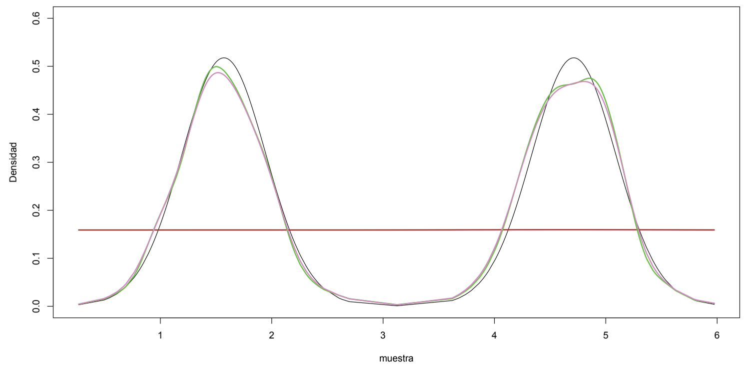

Figura 3.1: Repesentación de la densidad teórica de una mixtura de von Mises (3.15)(negra) y la estimación no paramétrica de la densidad selecionando el parámetro de suavizado mediante las técnicas estudiadas. L_2 (azul), Kullback-Leibler (verde) y plug-in (violeta)

Capítulo 4

Estimación de la regresión circular-lineal

En la Sección 4.1. describiremos la estimación no paramétrica tipo núcleo de la regresión circular-lineal que se puede utilizar para investigar la relación entre una variable explicativa direccional y una variable respuesta lineal. Mediante validación cruzada obtendremos diferentes parámetros de suavizado mediante los métodos de validación cruzada y plug-in (ver Sección 4.2.).

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 dicha relación.

4.1. Estimación no paramétrica tipo núcleo de la función de regresión circular-lineal

Dadas n observaciones de una variable direccional explicativa $\Theta_1, \ldots, \Theta_n$ y una variable respuesta lineal $Y_1, \ldots, Y_n$, suponemos que $Y_i = m(\Theta_i) + \epsilon_i$, donde ϵ_i son variables con media cero, independientes e identicamente distribuídas. Luego, la estimación no paramétrica tipo núcleo de la regresión circular-lineal, $m(\theta)$, viene dada por (ver Haiyong y Schoenberg)

$$\hat{m}(\theta; \nu) = \frac{\sum_{i=1}^{n} Y_i g(\theta - \Theta_i, 0, \nu)}{\sum_{i=1}^{n} g(\theta - \Theta_i, 0, \nu)}, \tag{4.1}$$

donde función núcleo g denota la función de densidad von Mises (2.9).

Dicha estimación (4.1) es análoga al estimador propuesto por Nadaraya-Watson en 1964 para el caso lineal.

4.2. Selección del parámetro de suavizado

Así como ocurría con la estimación no paramétrica de la densidad, la elección del parámetro de suavizado es crucial para la estimación no paramétrica tipo núcleo de la regresión. Existen varios métodos para seleccionar el parámetro de suavizado automáticamente y pueden ser extendidos al caso de la estimación no paramétrica tipo núcleo de la regresión. Nos centraremos en el método plug-in y en el método de validación cruzada (ver Haiyong y Schoenberg).

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 dado por el mínimo de la desviación típica muestral $x_1, \ldots, x_n$ y el rango intercuartílico dividido por 1.34).

Como se vió anteriormente, ambas pueden estar conectadas con el parámetro κ de la distribución de von Mises usando que si θ es distribuída de acuerdo con la distribución von Mises centrada en μ y con parámetro κ, entonces

$$\kappa^{-1/2}(\theta - \mu) \to_D N(0,1), \ \kappa \to \infty.$$

Esto sugiere la elección del parámetro de concentración ν mediante

$$\nu = \frac{1}{h_s^2}. \tag{4.2}$$

Otra opción es seleccionar ν por validación cruzada de mínimos cuadráticos (LCV). Como en el caso lineal, escogemos ν de modo

$$\nu_{opt} = \arg\min_{\nu \geq 0} CV(\nu), \tag{4.3}$$

donde

$$CV(\nu) = n^{-1}\sum_{j=1}^{n}[Y_j - \hat{m}^{-j}(\Theta_j;\nu)]^2,$$

siendo

$$\hat{m}^{-j}(\Theta_j;\nu) = \frac{\sum_{i\neq j}^{n} Y_i g(\Theta_j - \Theta_i, 0, \nu)}{\sum_{i\neq j}^{n} g(\theta - \Theta_i, 0, \nu)},$$

la estimación no paramétrica tipo núcleo de la regresión circular-lineal 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 infrasuavizan, es decir, ν tiende a ser muy grande. Por tanto valores de ν escogidos de acuerdo (4.2) serán preferidos en este caso.

Capítulo 5

El modelo de Möbius de series de tiempo

En este capítulo se adapta un modelo de regresión circular propuesto en Mardia y Downs (2002) al contexto de series de tiempo siguiendo la metodología propuesta en Hughes (2007) (ver Sección 5.1.). La distribución de θ_t dado θ_{t-1} es modelada usando una distribución von Mises particular. La función de máxima verosimilitud (condicionada a la primera observación) es obtenida en la Sección 5.2., así como la estimación de los parámetros.

5.1. El modelo Möbius de series de tiempo

La componente determinística del modelo de regresión estudiado por Downs y Mardia (2002) une a la variable angular dependiente v con la variable angular independiente u mediante la siguiente expresión

$$\tan\frac{1}{2}(v-\beta) = \omega\tan\frac{1}{2}(u-\alpha), \tag{5.1}$$

donde $\omega \in [-1, 1]$ es el parámetro que determina la pendiente y $-\pi \leq \alpha, \beta < \pi$ son parámetros que determinan la localización angular. La ecuación (5.1) nos conduce a

$$v = \beta + 2\tan^{-1}\left\{\omega\tan\frac{1}{2}(u-\alpha)\right\}. \tag{5.2}$$

Por lo que podemos decir que hemos definido una relación uno a uno entre u y v siempre que ω sea distinto de cero (5.2). El lugar de los puntos (u, v) satisfaciendo (5.1) es una curva cerrada y continua que da una sola vuelta alrededor de una superficie toroidal.

Por tanto, aplicaremos la ecuación (5.2) a las series de datos temporales, reemplazando el ángulo v por θ_t y el ángulo u por θ_{t-1}, $t = 2, \ldots, n$. Esta sustitución nos sugiere la existencia de un único parámetro de localización, $\alpha = \beta$, obteniendo así

$$\tan\frac{1}{2}(\theta_t - \alpha) = \omega \tan\frac{1}{2}(\theta_{t-1} - \alpha) \tag{5.3}$$

y

$$\theta_t = \alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\} \tag{5.4}$$

ecuaciones análogas a (5.1) y (5.2).

Para el modelo de serie de tiempo dado en (5.4) se asume que $\theta_t|\theta_{t-1}$ sigue una distribución von Mises

$$\theta_t|\theta_{t-1} \sim vM\left(\alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\}, \kappa\right). \tag{5.5}$$

Por tanto, el modelo de series de tiempo se convierte en

$$\theta_t = \alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\} + \epsilon_t, \tag{5.6}$$

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

$$\mu_t = \alpha + 2\tan^{-1}\left\{\omega \tan\frac{1}{2}(\theta_{t-1} - \alpha)\right\}. \tag{5.7}$$

Notemos que valor de ω se restringe al intervalo [-1,1] así que α es único.

5.2. Función de máxima verosimilitud

Como

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

se obtiene que

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

dónde μ_t es dada por (5.7). Por tanto, la función de masa de probabilidad de $\theta_2, \ldots, \theta_n$ dado θ_1 es

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

Pero en la ecuación (5.5) el valor de θ_t depende solo del valor θ_{t-1}. Además,

$$f(\theta_t|\theta_1, \ldots, \theta_{t-1}) = f(\theta_t|\theta_{t-1}), \ \forall t = 2, \ldots, n.$$

La función de 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 el logaritmo de la verosimilitud condicionada

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

La cual es maximizada respecto los parámetros α y ω desconocidos

$$l_C(\alpha, \omega) = \sum_{t=2}^{n} \cos\left[\theta_t - \alpha - 2\tan^{-1}\left\{\omega \tan \frac{1}{2}(\theta_{t-1} - \alpha)\right\}\right]. \tag{5.8}$$

La maximización de (5.8) la haremos usando la función *nlm* de R. Esta función utiliza un algoritmo tipo Newton para minimizar una función dada. Por tanto hemos minimizado la función $-l_C(\alpha, \omega)$. Una vez que $l_C(\alpha, \omega)$ ha sido máximizada con respecto a α y ω, se obtiene una estimación de κ maximizando

$$l_C(\hat{\alpha}, \hat{\omega}, \kappa) = const. - (n-1)\log I_0(\kappa) + \kappa l_C(\hat{\alpha}, \hat{\omega}), \tag{5.9}$$

respecto de κ. Diferenciando (5.9) respecto de κ obtenemos

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

donde

$$I_1(\hat{\kappa}) = \frac{1}{2\pi}\int_0^{2*\pi} e^{-i(n\theta - \kappa \operatorname{sen}(\theta))} d\theta.$$

Así que κ es la solución de

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

Capítulo 6

Análisis de 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 (ver Sección 6.2.) así como realizaremos la estimación no paramétrica tipo núcleo de la densidad (ver Sección 6.3.) y estimaremos el modelo de Möbius de series de tiempo (ver Sección 6.4.). En la Sección 6.5 estudiaremos qué 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.

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 evitar medidas repetidas y poder llevar a cabo este capítulo fue necesaria realizar una perturbación en los datos, descrita en la Sección 6.1. 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 y G2; sino que vamos a trabajar con los datos horarios perturbados.

6.1. 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 puede ser debido a que estos dispositivos no son lo suficientemente precisos. La aparición de medidas repetidas es un problema a la hora de implementar ciertas técnicas, como la validación cruzada a la hora de seleccionar el parámetro de suavizado. Con el fin de solucionar dicho problema, haremos una perturbación en ambas variables. La perturbación en el caso lineal, motivada por Azzalini (1981), nos lleva a una perturbación en las concentraciones de SO_2 dada por:

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

donde X_i denota los valores observados de SO_2, $b = 1{,}03\hat{\sigma}n^{-1/3}$ y $\epsilon_i, i = 1, \ldots, n$ son variables aleatorias independientes e identicamente distribuídas siguiendo el núcleo de Epanechnikov en $(-\sqrt{5}, \sqrt{5})$; $\hat{\sigma}$ es un estimador robusto de la varianza de X_i, el cual puede ser obtenido usando el rango intercuartílico estandarizado. Azzalini (1981) muestra que esta elección de b para los datos perturbados permite una estimación consistente de la función de distribución, obteniedo 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 perturbada de la dirección del viento viene dada por

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

donde θ_i denota los valores observados de la dirección de viento y $\epsilon_i, i = 1, \ldots, n$ son variables aleatorias independientes generadas de la distribución von Mises con $\mu = 0$ y $\kappa = 1$. García-Portugués, Crujeiras y González-Manteiga (2011) escogieron $d = n^{-1/5}$ basándose en los resultados de Liu y Yang (2008) para la estimación tipo núcleo de la distribución multivariante. Esta perturbación resuelve el problema de los valores repetidos.

6.2. 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 ,ver Figura 1.1, que van a a ser usados en todo el capítulo excepto en la Sección 6.5.

Esta estación recoge datos minutales pero trabajaremos con las medias horarias para reducir la gran cantidad de datos, aunque ésta sigue siendo elevada (8360),así como la dependencia. Con el objetivo de que los resultados sean más fáciles de visualizar, hemos dividido el año 2010 en cuatro periodos de 2090 datos ,ver Tabla 6.1. Estos periodos son de tres meses; el primero corresponde a los meses de enero, febrero, marzo; el segundo a los meses de abril, mayo, junio; el tercero a los meses de julio, agosto, septiembre, el cuarto a los meses de octubre, noviembre, diciembre.

En lo que se refiere a notación hay que decir que la dirección de viento va a tomar valores entre $[0, 2\pi)$ siendo el sentido antihorario el sentido positivo y cada valor representa el ángulo que forma respecto al Este. Por tanto, los valores 0, $\frac{\pi}{2}$, π, $\frac{3\pi}{2}$ 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 (ver Tabla 6.1).

	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

Cuadro 6.1: 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 (ver Figuras 6.1, 6.2, 6.3, 6.4) lo que ya se podía deducir de la Tabla 6.1 donde se muestra que $\bar{R}$ es próxima a 0 en los cuatro periodos en los que hemos divido el año.

A la vista de los diagramas de rosa, ver Figura 6.5, podría parecer que la distribución de la dirección de viento en los cuatro periodos en los que hemos dividido el año 2010 es uniforme. Sin embargo, al realizar el test de uniformidad de Rao (ver Rao, 1976; Rao y SenGupta, 2001) para las medias horarias de la dirección de viento en cada uno de los periodos; se obtiene (con un nivel de significación 0.05) que en ningún periodo las direcciones de viento siguen una distribución uniforme.

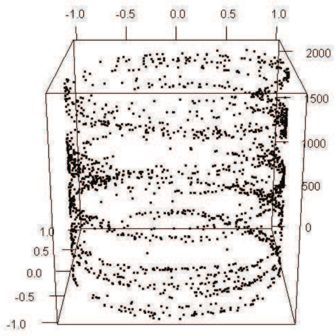

Figura 6.1: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010.

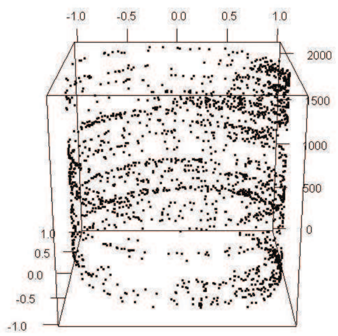

Figura 6.2: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010.

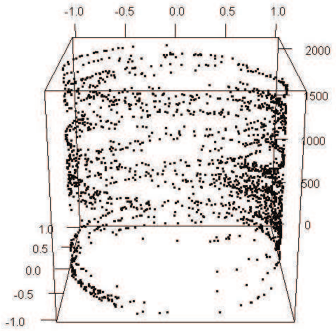

Figura 6.3: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010.

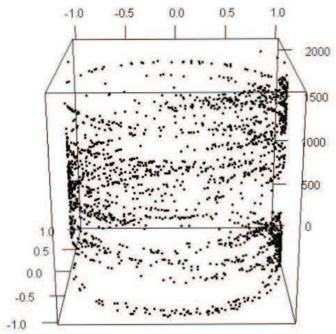

Figura 6.4: Gráfico de dispersión de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010.

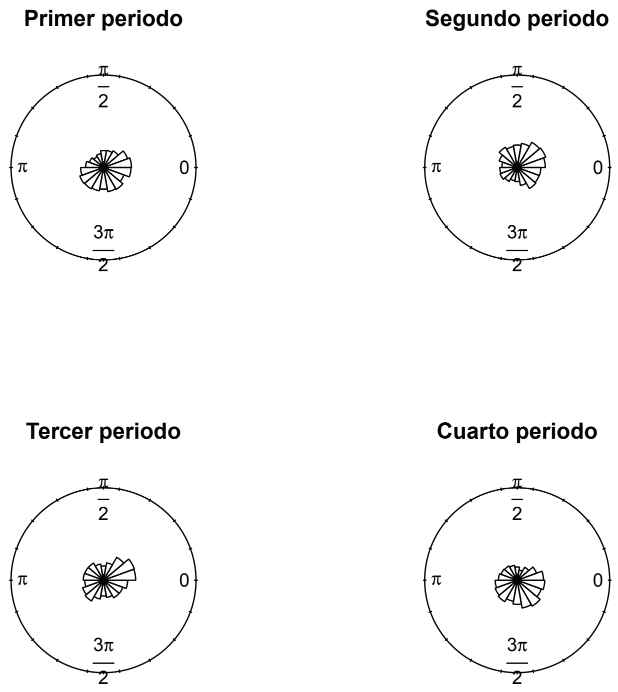

Figura 6.5: Diagrama de rosa de las medias horarias de la dirección de viento.

6.3. Estimación tipo núcleo de la función de densidad circular

En esta sección obtendremos la estimación no paramétrica tipo núcleo de la densidad circular 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; el método plug-in (3.14), el método de validación cruzada dada por la función de pérdida pérdida cuadrática (L_2) (3.11) y mediante el método de validación cruzada dada por la función de pérdida Kullback-Leibler (3.12).

Como se puede ver en la Tabla 6.2 los parámetros de suavizado obtenidos mediante validación cruzada (3.11,3.12) toman valores más altos que los obtenidos mediante el método plug-in (3.14) en todos los periodos. Por eso, las estimaciones no parámetricas tipo núcleo de la densidad circular 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	L_2	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

Cuadro 6.2: 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 (ver Figuras 6.6, 6.7, 6.8, 6.9) como ya apuntaban los valores obtenidos en la Tabla 6.2.

Las estimaciones obtenidadas por los dos métodos de validación cruzada proporcionan prácticamente los mismos resultados para cualquiera de los periodos considerados. Pues, en los cuatro periodos del año de 2010 se obtuvo que tanto la validación cruzada para minimizar la función pérdida dada por error cuadrático medio (L_2) (3.11) como la dada por la función de pérdida de Kullback-Leibler (3.12) alcanzan un mínimo en valores próximos.

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 o la dada por la función de pérdida dada por Kullback-Leibler.

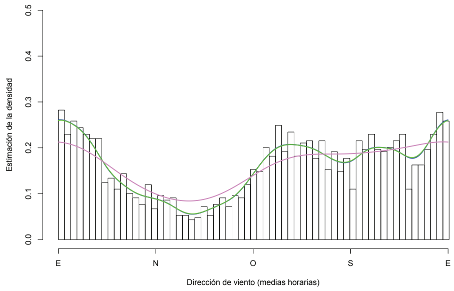

Figura 6.6: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L_2 (azul) y por la función de pérdida Kullback-Leibler (verde).

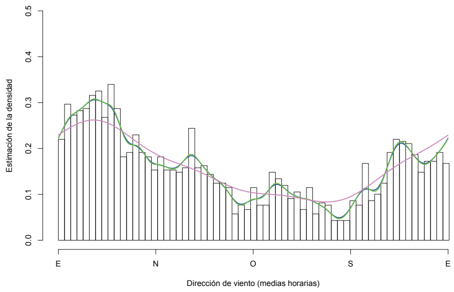

Figura 6.7: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L_2 (azul) y por la función de pérdida Kullback-Leibler (verde).

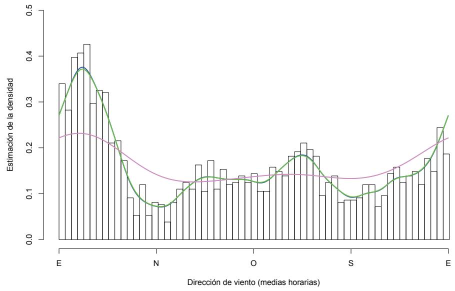

Figura 6.8: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L_2 (azul) y por la función de pérdida Kullback-Leibler (verde).

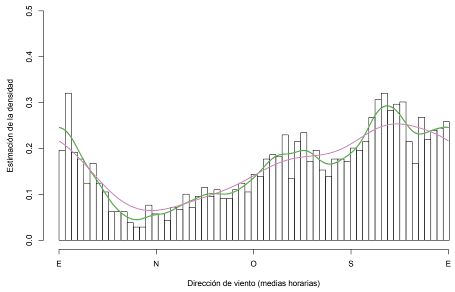

Figura 6.9: Estimación no paramétrica tipo núcleo de la función de densidad circular de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010. Parámetro de suavizado obtenido mediante plug-in (violeta), mediante validación cruzada dada por la función de pérdida L_2 (azul) y por la función de pérdida Kullback-Leibler (verde).

6.4. Estimación del modelo de Möbius de series de tiempo

En esta sección realizaremos la estimación del modelo de Möbius de series de tiempo (5.6) 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).

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

En este caso, (Θ_1, Φ_1) y (Θ_2, Φ_2) son independientes e identicamente distribuídas como (Θ, Φ). El estimador de ρ_T que vamos a utilizar es

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

y constituye la base para el cálculo del coeficiente de autocorrelación circular. Por conveniencia tomamos $\phi_i = \theta_{i+k},\ i = 1, \ldots, n-k$ luego tenemos $(n-k)$ pares

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

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

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

lo que es equivalente a

$$\hat{\rho}_T^k = \frac{\sum_{1\leq i<j\leq n-k}\operatorname{sen}(\theta_i - \theta_j)\operatorname{sen}(\theta_{i+k} - \theta_{j+k})}{\left[\sum_{1\leq i<j\leq n}\operatorname{sen}^2(\theta_i - \theta_j)\sum_{1\leq i<j\leq n-k}\operatorname{sen}^2(\theta_{i+k} - \theta_{j+k})\right]^{1/2}}. \tag{6.1}$$

Por tanto, si representamos $\hat{\rho}_T^k$ frente a k obtendremos la autocorrelación circular para los distintos periodos(ver Figuras 6.10, 6.12, 6.14, 6.16). En estos gráficos se observa que las medias horarias de las direcciones de viento correspondientes primer, segundo y cuarto periodo son muy dependientes. En cambio las medias horarias de las direcciones de viento en el tercer periodo presentan una dependencia no tan destacada.

Para construir el modelo de Möbius de series de tiempo (5.4) hemos tenido que estimar los parámetros α, ω, κ (ver Tabla 6.3) donde $\omega \in [-1, 1]$ y $-\pi\ \alpha < \pi$. El modelo de series de tiempo obtenido en cada uno de los periodos modela bien nuestros datos de dirección de viento (ver Figuras 6.11, 6.13, 6.15, 6.4). De hecho si calculamos, el error cuadrático medio mediante la distancia longitud de arco, que viene dado por

$$\frac{1}{n}\sum_{i=1}^{n} d(\theta_i, \hat{\theta}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(\pi - |\pi - |\theta_i - \hat{\theta}_i||)^2,$$

donde

- $\theta_i,\ i = 1, \ldots, n$ corresponden a las medias horarias de la dirección de viento del periodo correspondiente.
- $\hat{\theta}_i,\ i = 1, \ldots, n$ corresponden a las estimaciones dadas por modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento del periodo correspondiente.
- d es la distancia entre dos ángulos definida por la longitud de arco.

Se obtiene que el error cuadrático medio en cada uno de los periodos es 0.237, 0.249, 0.252 y 0.247. Luego, las desviaciones estandar de las predicciones con respecto a las observaciones está alrededor de $\sqrt{0{,}25}$ radianes, o lo que es lo mismo a 0.5 radianes.

En las Figuras 6.11, 6.13, 6.15, 6.4 se observa que el modelo de Möbius de series de tiempo presenta fuertes subidas y bajadas. Éstas son debidas a la representación empleada ya que se dibujan muy alejados los valores 0 y 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

Cuadro 6.3: Estimaciones de los parámetros que están involucrados en la obtención del modelo de Möbius de series de tiempo

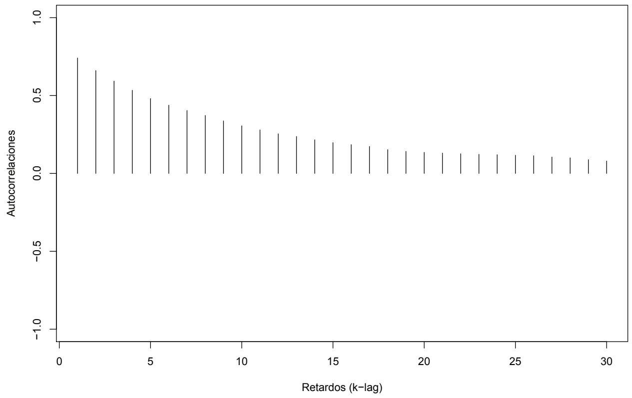

Figura 6.10: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010.

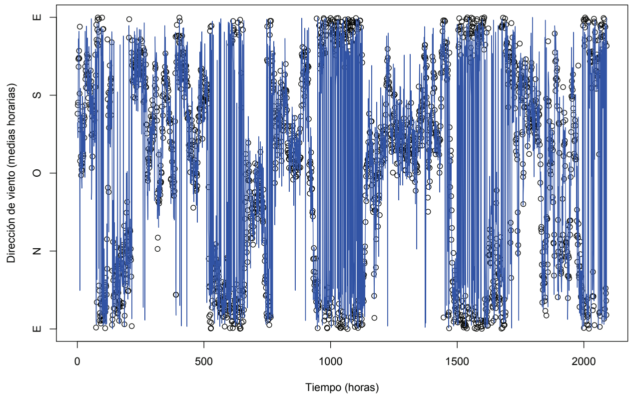

Figura 6.11: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al primer periodo de 2010.

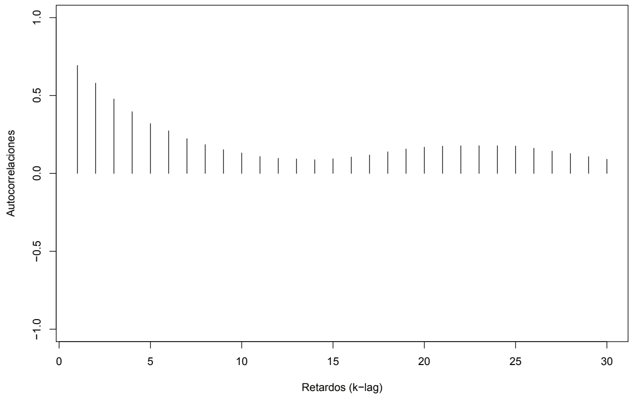

Figura 6.12: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010.

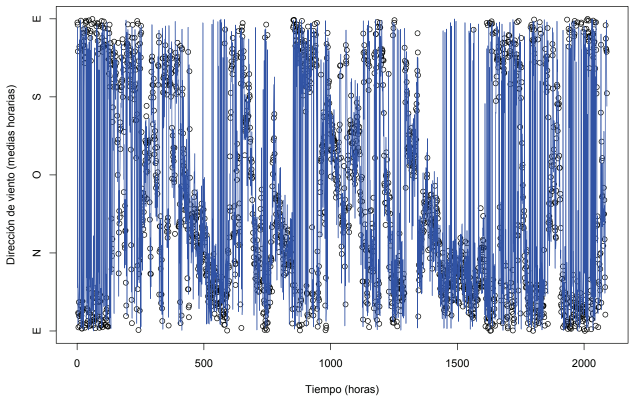

Figura 6.13: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al segundo periodo de 2010.

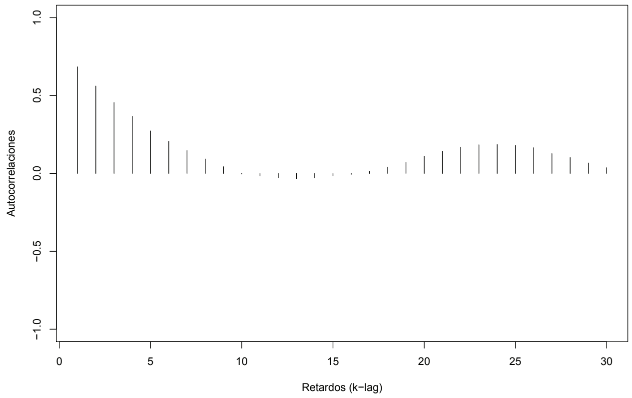

Figura 6.14: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010.

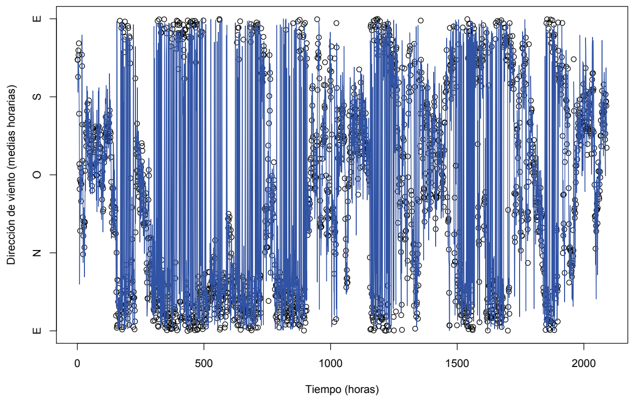

Figura 6.15: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al tercer periodo de 2010.

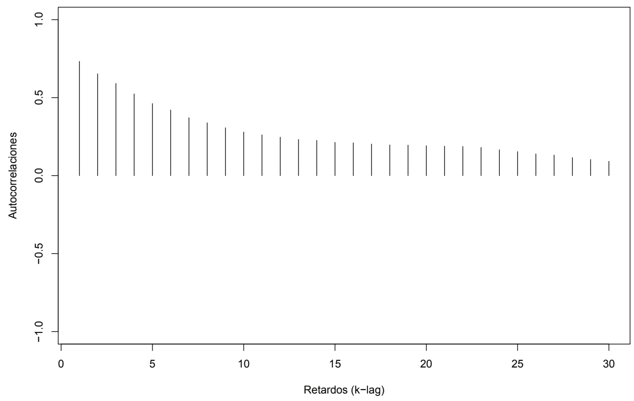

Figura 6.16: Autocorrelaciones circulares de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010.

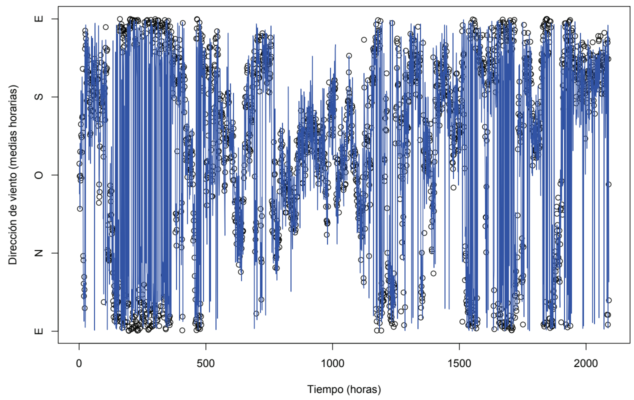

Figura 6.17: Modelo de Möbius de series de tiempo de las medias horarias de la dirección de viento correspondientes al cuarto periodo de 2010

6.5. Estimación tipo núcleo de la función de regresión circular-lineal

En esta sección vamos a realizar la estimación de la regresión circular-lineal de las media horarias transformadas mediante la función logaritmo de las concentraciones de SO_2 frente a las direcciones horarias de viento en las cinco estaciones de medición automática, B1, B2, C9, F2 y G2. Estas cinco estaciones que pretenecen a la Red de Vigilancia y Control de la Calidad Atmosférica de la U.P.T. de As Pontes.

En todas ellas se mide tanto la dirección de viento como las concentraciones de SO_2, utilizaremos estas medidas 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 circular-lineal. Para realizar esta estimación necesitamos que tanto los datos de SO_2 y dirección de viento estén recogidos por la misma estación.

Comencemos analizando los valores que hemos obtenido para el parámetro de suavizado en la estimación de la función de regresión circular-lineal. En la Tabla 6.4) se obtienen los parámetros de suavizado mediante validaión cruzada.

	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

Cuadro 6.4: Estimaciones del parámetro de suavizado obtenidas mediante validación cruzada

Por otra parte hemos obtenido el parámetro de suavizado mediante la expresión dada por (4.2). En este caso los parámetros de suavizado obtenidos son muy variables (ver Tabla 6.5); es decir, obtenemos parámetros pequeños como los obtenidos en el segundo, tercer y cuarto periodo en la estación F2 y demasiado grandes como el obtenido durante el cuarto periodo en B1 o durante el primer periodo en F2. De hecho, no somos capaces de obtener la estimación no paramétrica tipo núcleo de la densidad circular en la estación B1 durante el cuarto periodo (ver Figura 6.23, 6.24) ni en la estación F2 durante el primer periodo (ver Figura 6.29, 6.30) debido a que el parámetro de suavizado es demasiado grande.

Como el parámetro de suavizado seleccionado mediante la técnica plug-in toma valores tan variable se comporta diferente en cada situación; por ejemplo, en la estación G2, Figura 6.32, se observa que la estimación no paramétrica tipo núcleo de la densidad circular es muy suave durante el primer y el cuarto periodo pero durante el segundo y tercer periodo se obtiene una estimación infrasuavizada. Por lo que decimos que escoger el parámetro de suavizado mediante el método plug-in dado por (4.2), no parece ser una buena elección.

	B1	B2	C9	F2	G2
Primer periodo	592.52	1.04	16.48	1329.96	0.76
Segundo periodo	28.55108	108.28	7.32	0.50	71.05
Tercer periodo	44.07	331.62	7.15	0.51	83.39
Cuarto periodo	48020.99	1.21	11.38	0.51	0.72

Cuadro 6.5: 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 las siguientes gráficas (ver Figuras 6.18, 6.19, 6.20, 6.21, 6.22).

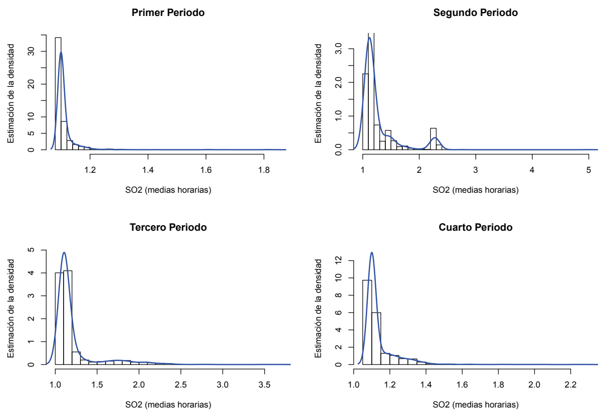

Figura 6.18: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación B1.

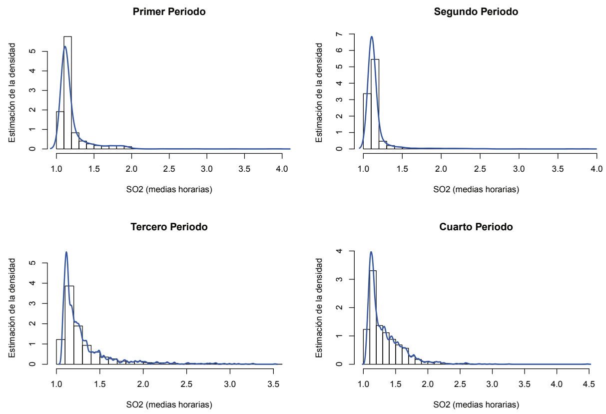

Figura 6.19: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación B2.

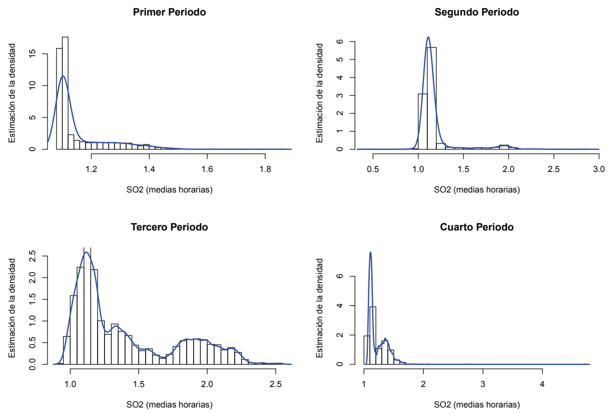

Figura 6.20: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación C9.

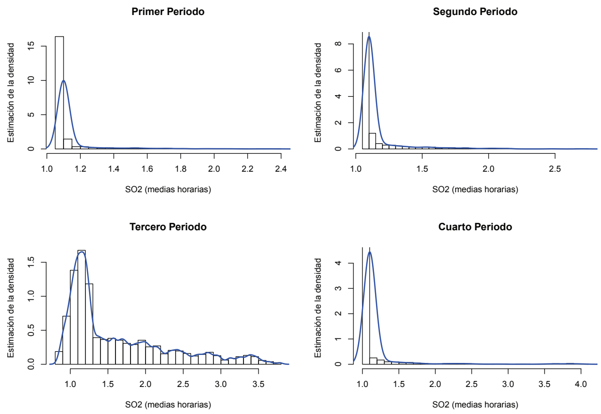

Figura 6.21: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la estación F2.

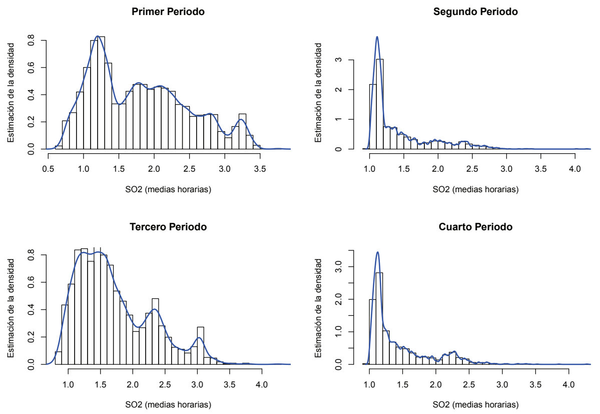

Figura 6.22: Estimación no paramétrica tipo núcleo de la función de densidad de las medias horarias de las concentraciones de SO_2 correspondientes a la G2.

A la vista de las Figuras 6.23, 6.25, 6.27, 6.29, 6.31 podemos decir que cuando se pueden construir ambas estimaciones no parámetricas tipo núcleo de la función regresión circular-lineal, se comportan de modo similar en cuanto a establecer una relación entre las concentraciones de SO_2 y la dirección de viento; puesto que, aparentemente el viento no explica las concentraciones de SO_2. 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 circular-lineal 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.

Sin embargo en la estación G2 durante el primer periodo del año 2010 (ver Figura 6.31) se observa que cuando el viento sopla del suroeste las concentraciones de SO_2 son más elevadas. Por lo que, en este caso, existe relación entre las medias horarias de las concentraciones de SO_2 y las medias horarias de la dirección de viento. Además podemos decir que las concentraciones de SO_2 pueden ser debidas a una fabrica de tableros DF que está situada al suroeste de G2.

Notemos que en este periodo las medias horarias de las concentraciones de SO_2 no están tan concentrados (ver Figura 6.22); las concentraciones las medias horarias transformadas de SO_2 toman valores entre 0.5 y 3.5. Luego las concentraciones las medias horarias de SO_2 toman valores entre 1 y 33, aproximadamente.

Por otra parte a la vista de las Figuras 6.24, 6.26, 6.28, 6.30, 6.32 podemos decir que cuando se puede calcular la estimación no parámetricas tipo núcleo de la función regresión circular-lineal dada por el parámetro de suavizado plug-in esta se comporta de formas diferentes dando situaciones de infra y sobresuavizado. Así pues, en la estación B1 (ver Figura 6.24) se observa en los tres primeros periodos la estimación no paramétrica tipo núcleo de la función de regresión circular-lineal plug-in infrasuaviza; siendo mejor la estimación dada escogiendo el parámetro mediante validación cruzada. En las estaciones B2 y G2 (ver Figura 6.26, 6.32) se observa en el primer y cuarto periodo la estimación no paramétrica tipo núcleo de la función de regresión circular-lineal plug-in nuevamente sobresuaviza mientras que en el segundo y tercer periodo infrasuaviza; siendo mejor la estimación dada escogiendo el parámetro mediante validación cruzada. En las estación F2 (ver Figura 6.30) se observa que en el segundo, tercer, y cuarto periodo la estimación no paramétrica tipo núcleo de la función de regresión circular-lineal plug-in sobresuaviza; siendo mejor la estimación dada escogiendo el parámetro mediante validación cruzada. En cambio en C9 (ver Figura 6.28) se observa que las estimaciones dadas por el parámetro de suavizado escogido mediante la técnica plug-in o mediante validación cruzada son muy similares; llegando prácticamente a coincidir en el segundo periodo.

6.5.1. Estación B1

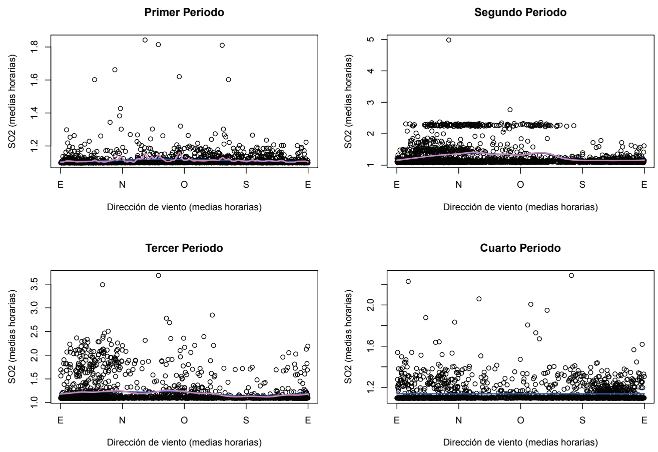

Figura 6.23: Estimación de la función de regresión circular-lineal en B1. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

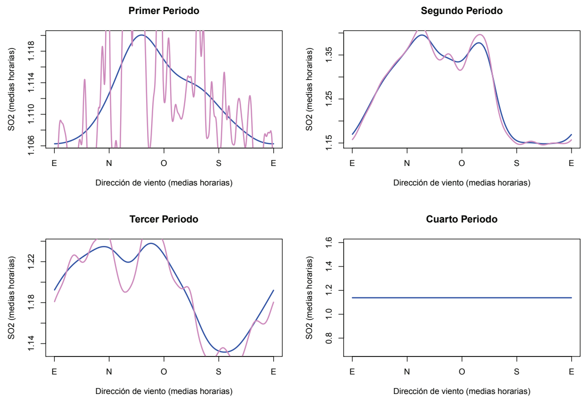

Figura 6.24: Estimación de la función de regresión circular-lineal en B1. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

6.5.2. Estación B2

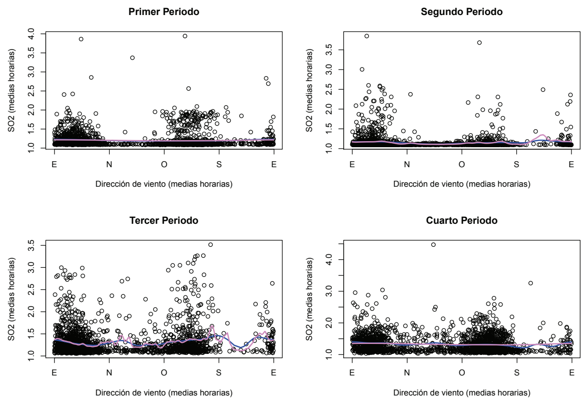

Figura 6.25: Estimación de la función de regresión circular-lineal en B2. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

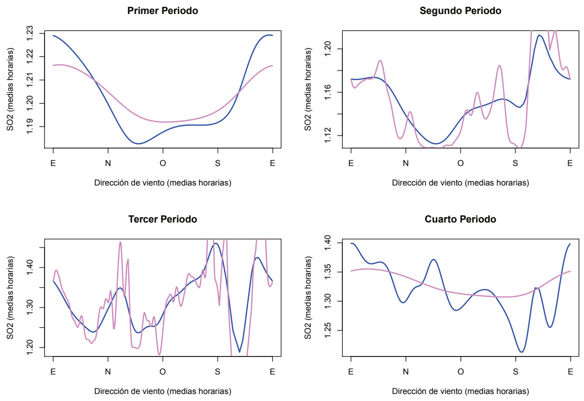

Figura 6.26: Estimación de la función de regresión circular-lineal en B2. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

6.5.3. Estación C9

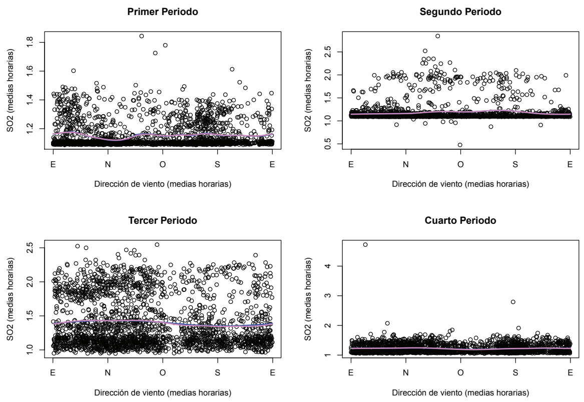

Figura 6.27: Estimación de la función de regresión circular-lineal en C9. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

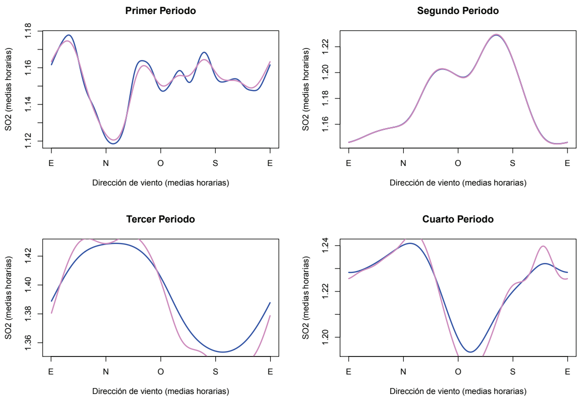

Figura 6.28: Estimación de la función de regresión circular-lineal en C9. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

6.5.4. Estación F2

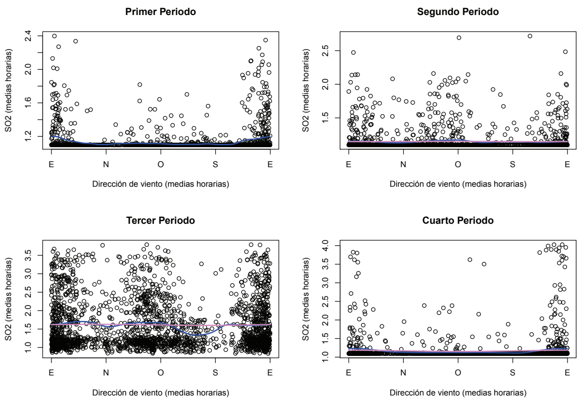

Figura 6.29: Estimación de la función de regresión circular-lineal en F2. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

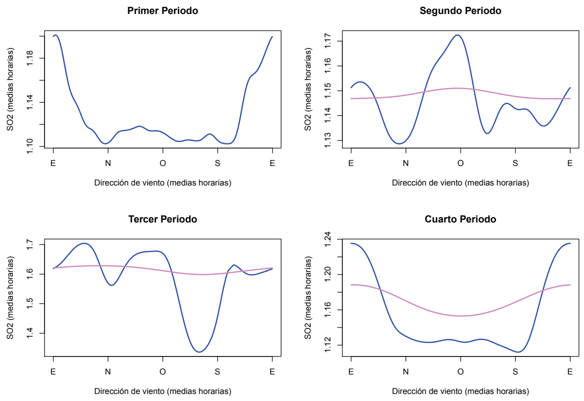

Figura 6.30: Estimación de la función de regresión circular-lineal en F2. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

6.5.5. Estación G2

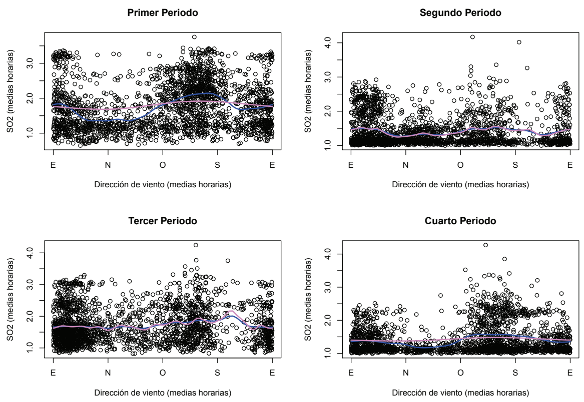

Figura 6.31: Estimación de la función de regresión circular-lineal en G2. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

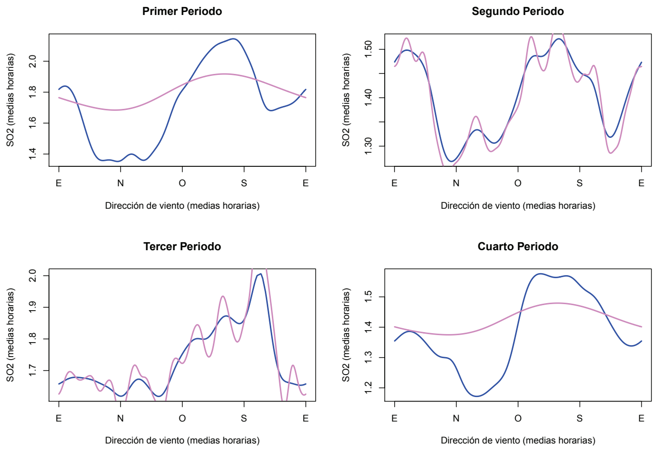

Figura 6.32: Estimación de la función de regresión circular-lineal en G2. Elección del parámetro de suavizado mediante $1/h_s^2$ (violeta). Elección del parámetro de suavizado mediante validación cruzada (azul)

Capítulo 7

Software

- En este trabajo se han utilizado las siguientes librerias del software libre R:
  - circular, CircStats; para realizar cálculos propios de datos circulares como la función de densidad von Mises, estimar el parámetro de concentración von Mises.
  - rgl; para la obtención de gráficos 3D.
- De lo visto en el Capítulo 3 se ha implementado:
  - El estimador no paramétrico tipo núcleo de la densidad circular (3.6).
  - El estimador del parámetro de suavizado mediante el método de validación cruzada dado por la función de pérdida cuadrática (3.11) y el dado por función de pérdida dada por Kullback-Leibler 3.12).
  - Estimación del parámetro de suavizado mediante la regla plug-in con densidad de referencia von Mises (3.14).
- De lo visto en el Capítulo 4 se ha implementado:
  - El estimador no paramétrico tipo núcleo de la función de regresión circular-lineal (4.1).
  - Estimaciones del parámetro de suavizado mediante el método de validación cruzada (4.3).
  - Estimaciones del parámetro de suavizado mediante el método plug-in (4.2).

- De lo visto en el Capítulo 5 se ha implementado:

  - La estimación del modelo de Möbius de series de tiempo (5.6).
  - La estimación de los parámetros mediante maxima verosimilitud (5.8).

  Además se ha implementado la estimación de la autocorrelación circular k-lag (6.1).

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] N. Fisher and A. Lee. A correlation coefficient for circular data. *Biometrika*, 70:327–332, 1983.

[5] Crujeiras R.M. González Manteiga W. García Portugués, E. Exploring with direction and SO_2 concentration by circular-linear density estimation. Technical report, Universidade de Santiago de Compostela, Departamento de Estatística e Investigación Operativa., 2011.

[6] 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.

[7] Watson G.S. Hall, P. and J. Cabrera. Kernel density estimation whith spherical data. *Biometrika*, 74:751–762, 1987.

[8] 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.

[9] S. Rao Jammalamadaka and A. SenGupta. *Topics in Circular Statistics.* World Scientific Press, Singapore., 2001.

[10] R. Liu and L. Yang. Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, 20:661–667, 2008.

[11] K.V. Mardia and P.E Jupp. *Directional Statistics*. Wiley; New York, 2000.

[12] J.S. Rao. Some tests based on arc-lengths for the circle. *Sankhya, The Indian Journal of Statistics, Serial B(4),*, 38:329–338, 1976.

[13] C.C. Taylor. Automatic bandwith selection for circular density estimation. *Computational Statistics and Data Analysis*, 52:3493–3500, 2008.