USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

UNIVERSIDADE DA CORUÑA
Departamento de Matemáticas

Universidade de Vigo

Análisis Estadístico en Geografía Física

María Oliveira Pérez

Julio 2011

Proyecto Fin de Máster

Índice general

Capítulo 1

Introducción

El presente trabajo nace a raíz de la colaboración con el grupo de investigación 40-04 de la facultad de Geografía e Historia de la Universidad de Santiago de Compostela, al que quiero agradecer, especialmente a Augusto Pérez Alberti (investigador principal del grupo), la confianza que han depositado en mí y la gentileza de propocionarme los datos para la realización de este trabajo.

La colaboración consistió en el análisis estadístico de datos recogidos para los proyectos, "*Cartografía y monitorización de formas crionivales en la región sub-antártica: Andes Fueguinos e Isla de los Estados (Tierra del Fuego, Argentina)*" (POL2006-09071) y "*Definición, cartografía y caracterización de las grandes áreas paisajísticas de Galicia*". El primero de ellos se trata de un proyecto enmarcado en el año polar internacional (2006-2009) subvencionado por el Ministerio de Ciencia y Tecnología y el segundo fue financiado por la Consellería de Medio Ambiente, Territorio e Infraestructuras.

En relación con el primer proyecto, se dispone de datos de la temperatura del aire y del suelo en cuatro localizaciones distintas del Monte Alvear situado en Tierra del Fuego (ver Figura 1.1) - Calm 1, Calm 2, Calm 3, Calm 4 - durante los años 2008 y 2009. A partir de las temperaturas registradas en esos puntos se estudia el número de ciclos que se producen anualmente, entendiendo por ciclo el período de tiempo durante el cual la temperatura no cambia de estar por encima de 0 °C a estar por debajo de 0 °C o viceversa. Además de ver simplemente el número de ciclos que se producen, es interesante estudiar a qué hora del día tienen lugar. Se tendrían así un conjunto de datos horarios, esto es, un conjunto de datos circulares o direccionales.

Los datos circulares o direccionales (y en general, los datos esféricos) tienen una serie de características que los hacen distintos de los datos lineales o escalares y por tanto, el análisis

direccional es sustancialmente diferente del análisis estadístico "lineal" estándar ya que la propia naturaleza de los datos obliga a replantear aspectos tan básicos como la medición de distancias. Así surge el problema de cómo estimar la función de densidad a partir de una muestra de datos circulares.

Se tienen también datos de la temperatura del aire, presión, radiación solar, velocidad y dirección del viento recogidos en una estación meteorológica ubicada en el glaciar Vinciguerra (ver Figura 1.1) situado en Ushuaia, en la zona septentrional de Argentina. Al igual que los datos horarios, la dirección del viento también es una variable de tipo circular. Esta variable podría estar relacionada con la temperatura del aire que es lineal, lo que nos llevará a describir técnicas de estimación de la regresión cuando la variable explicativa es circular y la variable respuesta es lineal, es decir, tendríamos una regresión lineal-circular. Cuando ambas variables son lineales podemos estimar no paramétricamente la función de regresión utilizando el estimador local lineal o el estimador de Nadaraya-Watson. Veremos que dichos estimadores se pueden extender al caso en el que la variable explicativa es circular. Además, realizaremos un estudio de simulación para ver si las técnicas adaptadas a datos circulares aportan mejores resultados con respecto a las técnicas lineales estándar.

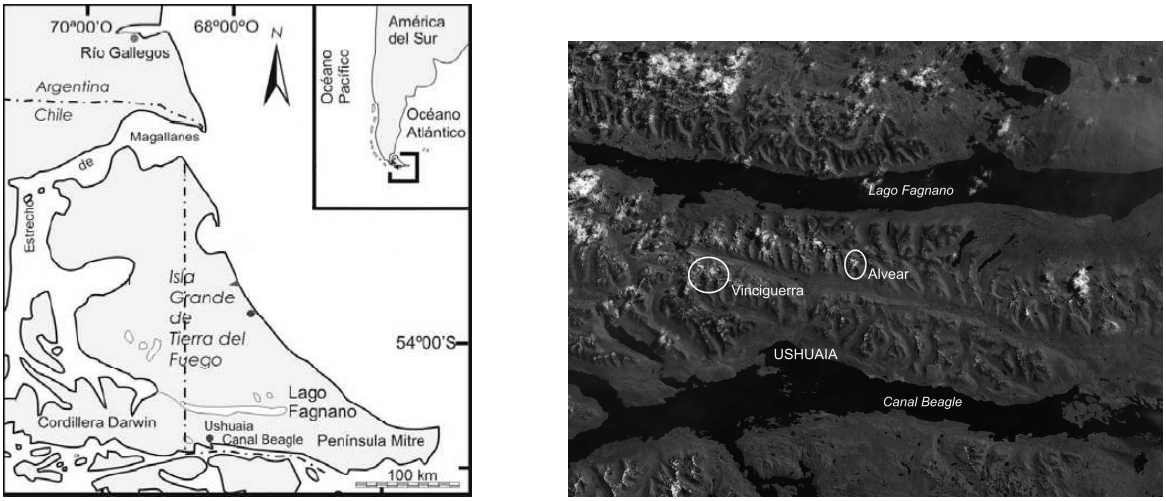


Figura 1.1: Mapa de Isla Grande de Tierra del Fuego (izquierda) y localizaciones donde se tomaron los datos: Monte Alvear y glaciar Vinciguerra (Ushuaia, Argentina) (derecha).

En relación con el segundo proyecto se tienen datos relativos a la dureza de cantos fluviales recogidos en distintos puntos del río Miño entre Ourense y Ribadavia, concretamente en: Puga, Laias, Troncoso, Santa Cruz, A Groba y Prado (ver Figura 1.2). El objetivo es

crear grupos de terrazas fluviales basándose en la dureza de la cuarcita, medida mediante un dispositivo denominado Equotip. Para ello hemos utilizado técnicas de análisis clúster. Estos análisis se presentaron en el congreso *FLAG Biennial Meeting 2010 (Portugal)* y un artículo sobre este estudio ha sido enviado a una revista (ver Pérez *et al.*, 2011).

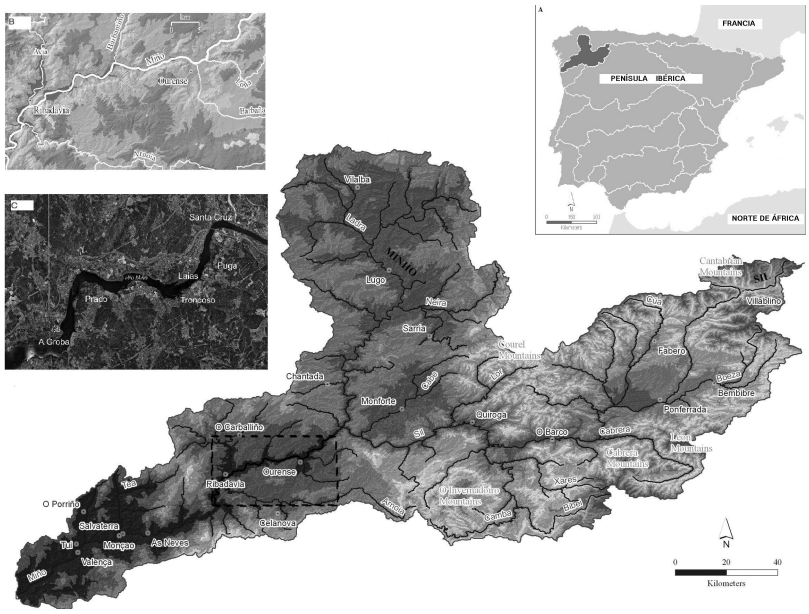


Figura 1.2: Mapa del río Miño, ubicación en la Península Ibérica (A), tramo del río Miño entre Ourense y Ribadavia (B) y localizaciones donde se tomaron los datos (C).

La memoria del proyecto fin de máster se organiza de la siguiente forma. En el Capítulo 2 desarrollaremos las técnicas para estimar no paramétricamente la función de densidad para datos circulares así como la forma de estimar no paramétricamente la función de regresión lineal-circular. En el caso de la función de densidad, generalizaremos el estimador tipo núcleo lineal al contexto circular y abordaremos el problema de selección automática del parámetro de suavizado proponiendo distintos métodos y estudiando su comportamiento en un estudio de simulación. Respecto a la estimación no paramétrica de la función de regresión, extenderemos el estimador local lineal y el estimador de Nadaraya-Watson al caso de que la variable explicativa sea circular y la respuesta lineal. Realizaremos un estudio de simulación en el que compararemos los estimadores adaptados a datos circulares con los estimadores lineales estándar utilizados en este contexto, es decir, ignorando el carácter circular de la variable explicativa.

En el Capítulo 3, presentaremos los datos correspondientes al primer proyecto e ilustraremos

el comportamiento de los estimadores de las funciones de densidad y regresión definidos en el Capítulo 2, con los datos de cambios de ciclo y dirección del viento para la densidad y con los datos de temperatura y dirección del viento para la regresión.

Finalmente, en el Capítulo 4, presentaremos de manera detallada los datos de dureza de cantos fluviales, efectuando un análisis descriptivo de los mismos y aplicando técnicas de formación de grupos para distinguir niveles de terrazas.

Capítulo 2

Estimación no paramétrica de la función de densidad y de regresión para datos circulares

2.1. Introducción

En muchos y diversos campos científicos, las medidas son direcciones. Por ejemplo, un biólogo puede medir la dirección de vuelo de un pájaro, un geólogo puede interesarse por la dirección del polo magnético de la tierra y, en nuestro caso, estamos interesados en la dirección del viento y en las horas a las que se produce un determinado fenómeno, concretamente los cambios de ciclo. Tales direcciones pueden estar en tres dimensiones como el segundo ejemplo o en dos dimensiones como en los demás ejemplos. Este tipo de datos se conocen como datos direccionales.

Las direcciones en dos dimensiones se pueden representar como ángulos medidos con respecto a alguno convenientemente elegido, es decir, una vez elegido un punto de partida y un sentido de rotación (por ejemplo, tomando como dirección positiva el sentido de las agujas del reloj o el sentido antihorario). Dado que una dirección no tiene magnitud, estos datos pueden ser representados como puntos en la circunferencia de un círculo unitario con centro en el origen o como vectores unitarios. Debido a esta representación circular, datos direccionales en dos dimensiones también se llaman datos circulares. De manera análoga, direcciones en tres dimensiones pueden representarse por dos ángulos, como vectores unitarios en tres dimensiones o como puntos en la superficie de una esfera unidad. Los datos direccionales en tres o más dimensiones también se conocen como datos esféricos.

Los datos direccionales tienen características especiales, tanto en términos de modelos como en su tratamiento estadístico. Por ejemplo, la representación numérica de una dirección en dos dimensiones como un ángulo o un vector unidad no es necesariamente única ya que el valor angular depende del punto de partida y del sentido de rotación. Por tanto, es importante asegurarse de que las conclusiones extraídas de los datos en términos descriptivos o inferenciales están en función de las observaciones dadas y no dependen de los valores arbitrarios en los que nos referimos a ellos. Una vez más, debido a esta arbitrariedad, tampoco hay un orden natural para las observaciones, ya que decir que una dirección es "más grande" que otra depende de si es el sentido de las agujas del reloj o el sentido contrario el que se trata como la dirección positiva, así como de donde está el "cero". Finalmente, como el "principio" coincide con el "final", es decir, $0 = 2\pi$ (si medimos los ángulos en radianes) y la medida circular θ es periódica: $\theta = \theta + k \cdot 2\pi$ para cualquier entero k, los métodos para tratar los datos direccionales deben tener esto en cuenta tanto para medir la distancia entre dos observaciones como para proponer modelos. Véase Jammalamadaka y SenGupta (2001) para una introducción más completa sobre datos circulares.

Dentro del contexto de la estadística no paramétrica, existen dos grandes campos de estudio: la estimación no paramétrica de la densidad y la estimación no paramétrica de la regresión. Debido a que el análisis estadístico de los datos circulares difiere de los métodos lineales estándar, las técnicas no paramétricas para datos lineales pueden no proporcionar resultados satisfactorios para datos circulares.

En este capítulo se introducirá la distribución von Mises, uno de los modelos paramétricos más utilizado en datos circulares y que servirá de base para construir el estimador no paramétrico de la densidad que emplearemos. Además nos detendremos en los métodos de selección del parámetro de suavizado. Propondremos diferentes métodos para seleccionarlo y analizaremos su comportamiento mediante un estudio de simulación. Compararemos los resultados con los obtenidos con los métodos estándar para datos lineales. Finalmente, trataremos la estimación no paramétrica de la función de regresión cuando tenemos una variable explicativa circular y una variable respuesta lineal: veremos cómo estimar el parámetro de suavizado en estos casos y compararemos, en un estudio de simulación, las técnicas de estimación no paramétricas para datos circulares con las ya conocidas cuando tanto la variable respuesta como la explicativa son de tipo lineal.

2.2. La distribución von Mises

Una distribución de probabilidad circular es una distribución de probabilidad cuya probabilidad total está concentrada en la circunferencia de un círculo unidad. Como cada punto en la circunferencia representa una dirección, tal distribución es una forma de asignar probabilidades a diferentes direcciones o de definir una distribución direccional. El caso más simple de distribución circular sería aquella que asigna a todas las direcciones la misma probabilidad, esto es, la distribución circular uniforme.

Una de las distribuciones de probabilidad circulares más utilizada es la distribución von Mises, vM(μ, κ). Esta distribución, simétrica y unimodal, es el modelo más habitual de distribución en datos circulares, siendo análoga a la distribución Normal en el caso de variables lineales, de ahí que también sea conocida como distribución Normal Circular.

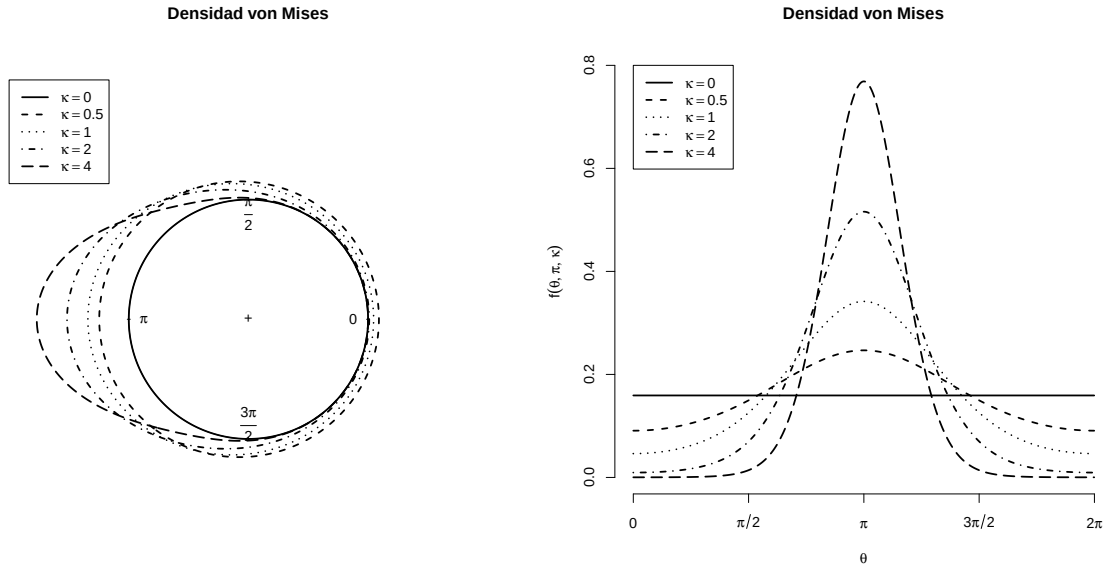


Figura 2.1: Representación circular (izquierda) y lineal (derecha) de una densidad von Mises con $\mu = \pi$ y κ =0, 0.5, 1, 2 y 4. El valor $\kappa = 0$ se corresponde con la circular uniforme.

La distribución von Mises se caracteriza mediante dos parámetros: la dirección media de la distribución, $\mu \in [0, 2\pi)$, y el parámetro de concentración, $\kappa \geq 0$. Su función de densidad es

$$f(\theta; \mu, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\left\{\kappa \cos(\theta - \mu)\right\}, \quad 0 \leq \theta < 2\pi, \tag{2.1}$$

donde $I_0(\kappa)$ denota la función de Bessel modificada de orden 0, que se define como

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

El parámetro μ es punto de simetría de la densidad y donde está localizada la moda. El parámetro de concentración κ mide la variación de la distribución en relación con una distribución circular uniforme. Cuando este parámetro κ tiende a cero, la distribución converge a la distribución circular uniforme; en caso de tender a infinito, la distribución se concentra en la dirección media. En la Figura 2.1 puede verse el efecto de este parámetro para varias distribuciones von Mises con $\mu = \pi$.

Estimación de los parámetros de una distribución von Mises

Sean $\theta_1, \theta_2, \ldots, \theta_n$ un conjunto de observaciones de una muestra aleatoria simple de una distribución von Mises con parámetros μ y κ. A partir de dichas observaciones, podemos plantearnos estimar los parámetros que caracterizan nuestro modelo vM(μ, κ). Para ello, introduciremos el método de los momentos y el método de máxima verosimilitud. Se trata de dos técnicas clásicas, pero que detallaremos en el contexto de los datos circulares.

Estimación por el método de los momentos

Los momentos de la distribución von Mises generalmente se calculan como los momentos de la variable $z = e^{i\theta}$, es decir, representando los datos como puntos en el plano. A estos momentos se les denomina "momentos circulares".

Si definimos la función de Bessel modificada de orden p como

$$I_p(\kappa) = \frac{1}{2\pi} \int_0^{2\pi} \cos(p\theta) \exp(\kappa \cos\theta) d\theta, \quad p = 0, 1, 2, \ldots$$

y tenemos en cuenta la relación:

$$\frac{1}{2\pi} \int_0^{2\pi} \text{sen}(p\theta) \exp(\kappa \cos\theta) d\theta = 0,$$

el momento de orden p de la densidad von Mises está dado por

$$\frac{1}{2\pi I_0(\kappa)}\int_0^{2\pi} e^{ip\theta}\exp(\kappa\cos(\theta-\mu))d\theta = \frac{1}{2\pi I_0(\kappa)}\int_0^{2\pi} e^{ip(\omega+\mu)}\exp(\kappa\cos\omega)d\omega =$$

$$= \frac{e^{ip\mu}}{2\pi I_0(\kappa)}\int_0^{2\pi}[\cos(p\omega)+i\,\mathrm{sen}(p\omega)]\exp(\kappa\cos\omega)d\omega = \frac{I_p(\kappa)}{I_0(\kappa)}e^{ip\mu} =$$

$$= \frac{I_p(\kappa)}{I_0(\kappa)}\left[\cos(p\mu)+i\,\mathrm{sen}(p\mu)\right].$$

El momento muestral de orden p de la variable $z = e^{i\theta}$ viene dado por:

$$\frac{1}{n}\sum_{j=1}^{n}\left(e^{i\theta_j}\right)^p = \frac{1}{n}\sum_{j=1}^{n}\cos(p\theta_j) + i\frac{1}{n}\sum_{j=1}^{n}\mathrm{sen}(p\theta_j).$$

Igualando los momentos poblacionales a los momentos muestrales para $p = 1$ se obtienen las ecuaciones que definen los estimadores $\hat{\mu}$ y $\hat{\kappa}$ por el método de los momentos para μ y κ, respectivamente:

$$A(\hat{\kappa})\cos\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\cos\theta_i, \tag{2.2}$$

$$A(\hat{\kappa})\,\mathrm{sen}\,\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{sen}\,\theta_i, \tag{2.3}$$

donde $A(\hat{\kappa}) = I_1(\hat{\kappa})/I_0(\hat{\kappa})$.

Supuesto que $\sum_{i=1}^{n}\cos\theta_i \neq 0$, haciendo el cociente de las ecuaciones (2.2) y (2.3) resulta:

$$\hat{\mu} = \tan^{-1}\left(\frac{\sum_{i=1}^{n}\mathrm{sen}\,\theta_i}{\sum_{i=1}^{n}\cos\theta_i}\right), \tag{2.4}$$

donde $\tan^{-1}$ denota la función arco tangente.

El estimador por el método de los momentos de μ es la dirección de la media muestral

$$\frac{1}{n}\sum_{j=1}^{n}\cos\theta_j + i\frac{1}{n}\sum_{j=1}^{n}\mathrm{sen}\,\theta_j.$$

Multiplicando (2.2) por $\cos\hat{\mu}$ y (2.3) por $\operatorname{sen}\hat{\mu}$ y sumando se obtiene que $\hat{\kappa}$ es solución de

$$A(\hat{\kappa}) = \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i - \hat{\mu}). \tag{2.5}$$

Estimación por máxima verosimilitud

Para obtener los estimadores de μ y κ por máxima verosimilitud partimos de la función de verosimilitud que viene dada por

$$L(\mu,\kappa|\theta_1,\theta_2,\ldots,\theta_n) = \frac{1}{[2\pi I_0(\kappa)]^n}\exp\left\{\sum_{i=1}^{n}\kappa\cos(\theta_i-\mu)\right\}.$$

Su logaritmo viene dado por

$$l = \log L = -n\log(2\pi I_0(\kappa)) + \kappa\sum_{i=1}^{n}\cos(\theta_i-\mu).$$

De la expresión anterior, calculando las derivadas parciales respecto a μ y κ e igualando a cero, se obtienen las ecuaciones que definen los estimadores $\hat{\mu}$ y $\hat{\kappa}$ de máxima verosimilitud para μ y κ, respectivamente:

$$\hat{\kappa}\sum_{i=1}^{n}\operatorname{sen}(\theta_i-\hat{\mu}) = 0,$$

$$-n\frac{I_1(\hat{\kappa})}{I_0(\hat{\kappa})} + \sum_{i=1}^{n}\cos(\theta_i-\mu) = 0$$

teniendo en cuenta que $\frac{dI_0(\kappa)}{d\kappa} = I_1(\kappa)$, la función de Bessel modificada de orden 1. Equivalentemente,

$$\hat{\kappa}\left[\cos\hat{\mu}\sum_{i=1}^{n}\operatorname{sen}\theta_i - \operatorname{sen}\hat{\mu}\sum_{i=1}^{n}\cos\theta_i\right] = 0, \tag{2.6}$$

$$A(\hat{\kappa}) = \frac{1}{n}\sum_{i=1}^{n}\cos(\theta_i-\hat{\mu}). \tag{2.7}$$

Si $\sum_{i=1}^{n} \cos\theta_i \neq 0$, de la ecuación (2.6) obtenemos el estimador de máxima verosimilitud para μ:

$$\hat{\mu} = \tan^{-1}\left(\frac{\sum_{i=1}^{n} \operatorname{sen}\theta_i}{\sum_{i=1}^{n} \cos\theta_i}\right).$$

En el caso de que $\hat{\kappa}$ se estime como cero, no existe un valor único para la estimación de μ.

El estimador de máxima verosimilitud de κ se obtiene como la solución de (2.7).

Por tanto los estimadores de máxima verosimilitud coinciden con los de estimadores obtenidos por el método de los momentos dados en (2.4) y (2.5).

Mixtura de von Mises

Al igual que en el contexto de datos lineales, las mixturas finitas de distribuciones von Mises, vM(μ_i, κ_i), con parámetros de mezcla p_i, $i = 1, \ldots, M$ ($\sum_{i=1}^{M} p_i = 1$), proporcionan una clase mucho más rica de modelos circulares.

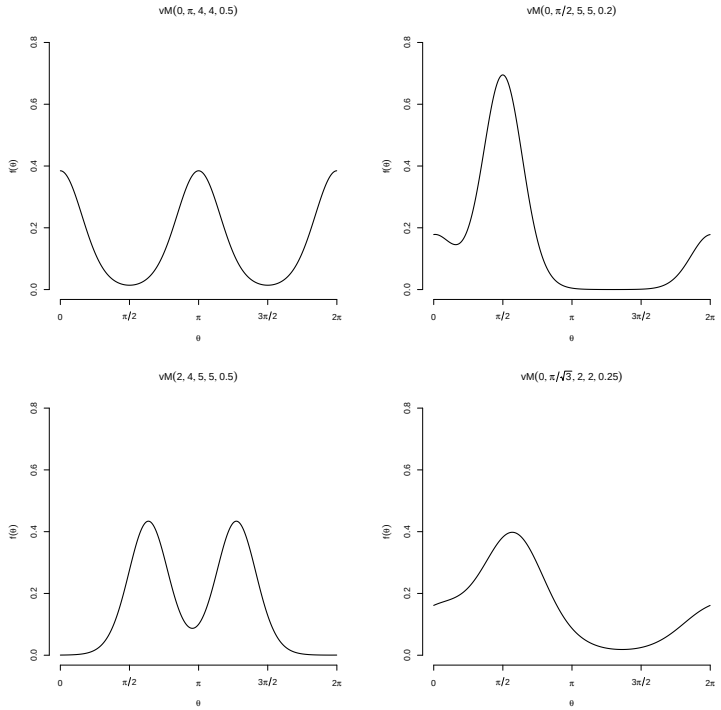


Figura 2.2: Representación lineal de las mixturas de von Mises: vM$(0, \pi, 4, 4, 0.5)$, vM$(0, \pi/2, 5, 5, 0.2)$, vM$(2, 4, 5, 5, 0.5)$ y vM$(0, \pi/\sqrt{3}, 2, 2, 0.25)$.

Su función de densidad es:

$$f(\theta) = \sum_{i=1}^{M} p_i \frac{\exp\{\kappa_i \cos(\theta - \mu_i)\}}{2\pi I_0(\kappa_i)}, \quad 0 \leq \theta < 2\pi. \tag{2.8}$$

Utilizaremos la notación vM(μ_1, μ_2, κ_1, κ_2, p) para denotar la mixtura de dos von Mises, vM(μ_1, κ_1) y vM(μ_2, κ_2), con proporciones de mixtura $p_1 = p$ y $p_2 = 1 - p$.

En la Figura 2.2, se representan gráficamente cuatro mixturas de von Mises: vM$(0, \pi, 4, 4, 0.5)$, vM$(0, \pi/2, 5, 5, 0.2)$, vM$(2, 4, 5, 5, 0.5)$ y vM$(0, \pi/\sqrt{3}, 2, 2, 0.25)$. Nótese que incluso con dos von Mises conseguimos modelos bastante flexibles, que permiten la presencia de bimodalidad y asimetría.

2.3. Estimación tipo núcleo de la densidad circular

Un problema fundamental de la estadística es la estimación de la función de densidad de una variable o vector aleatorio a partir de la información proporcionada por una muestra. Dicho problema se puede enfocar de dos formas. Un enfoque consiste en considerar que la función de densidad que deseamos estimar pertenece a una determinada familia paramétrica, por ejemplo: Normal, Exponencial, etc. Bajo este enfoque la estimación se reduce a determinar el valor de los parámetros del modelo a partir de la muestra. Este tipo de estimación se denomina estimación paramétrica de la densidad. Una alternativa es no predeterminar a priori ningún modelo para la distribución de probabilidad de la variable y dejar que la función de densidad pueda adoptar cualquier forma, sin más límites que los impuestos por las propiedades que se exigen a las funciones de densidad para ser consideradas como tales. Este enfoque se denomina estimación no paramétrica de la densidad y es en el que nos centraremos.

Entre los estimadores no paramétricos de la densidad consideraremos los estimadores tipo núcleo. Para datos lineales, dada una muestra aleatoria simple $X_1, X_2, \ldots, X_n$ de la variable de interés, el estimador tipo núcleo de la función de densidad viene dado por:

$$\hat{g}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} L\left(\frac{x - X_i}{h}\right), \tag{2.9}$$

donde $h > 0$ es el parámetro de suavizado, llamado comúnmente ventana, y L es una función real, llamada núcleo. Generalmente L es una función de densidad unimodal y simétrica alrededor del cero.

El estimador tipo núcleo definido en (2.9) se puede escribir como una mixtura de densidades centradas en cada dato X_i donde h controla su nivel de concentración. De forma más precisa, el estimador tipo núcleo se puede expresar como:

$$\hat{g}(x;h) = \sum_{i=1}^{n} \frac{1}{n} L_{ih}(x),$$

donde $L_{ih}(x) = \frac{1}{h} L\left(\frac{x-X_i}{h}\right)$ es la densidad de la variable $(X_i + hZ)$, siendo Z una variable aleatoria con densidad L. En el caso de que el núcleo sea el gaussiano, L_{ih} es la densidad de la normal con media X_i y varianza h^2. Habitualmente, h tenderá a cero cuando el tamaño muestral tiende a infinito, concentrando cada vez más la densidad L_{ih} alrededor del dato X_i.

Dada $\theta_1, \theta_2, \ldots, \theta_n \in [0, 2\pi)$, muestra aleatoria simple de ángulos, se define el estimador tipo núcleo de la densidad circular f como,

$$\hat{f}(\theta;\nu) = \sum_{i=1}^{n} \frac{1}{n} K_{i\nu}(\theta) = \frac{1}{n(2\pi)I_0(\nu)} \sum_{i=1}^{n} \exp\{\nu \cos(\theta - \theta_i)\}. \tag{2.10}$$

Este estimador generaliza el estimador (2.9) a datos circulares, tomando como $K_{i\nu}$ la von Mises con media θ_i y parámetro de concentración ν, donde ahora ν tenderá a infinito para que $K_{i\nu}$ esté cada vez más concentrada alrededor del dato θ_i. Es decir, $K_{i\nu}$ es una vM(θ_i,ν).

En el estimador (2.10), los núcleos son vM(θ_i, ν), jugando ν el papel del parámetro de suavizado.

2.3.1. Selección del parámetro de suavizado

El problema principal del estimador tipo núcleo, tanto en el caso lineal como en el circular, es la elección del parámetro ventana.

En el caso lineal, una ventana pequeña conduce a que el estimador utilice pocas observaciones en la ponderación, disminuyendo el sesgo, pero haciendo que las estimaciones dependan demasiado de la variabilidad muestral. Esta situación recibe el nombre de infrasuavización. Por el contrario, una ventana grande produce un aumento del sesgo y una reducción de la variabilidad, puesto que hace uso de muchas observaciones en el promedio, algunas muy alejadas del punto de interés, lo que se conoce como sobresuavización.

En el caso circular, el parámetro de concentración del núcleo ν juega un papel análogo a la

ventana h pero en sentido contrario, ya que valores grandes de ν proporcionarán menos suavizado y valores pequeños de ν proporcionarán más suavizado. Es decir, el comportamiento de ν en (2.10) es el inverso de h en (2.9) (ver Figura 2.3).

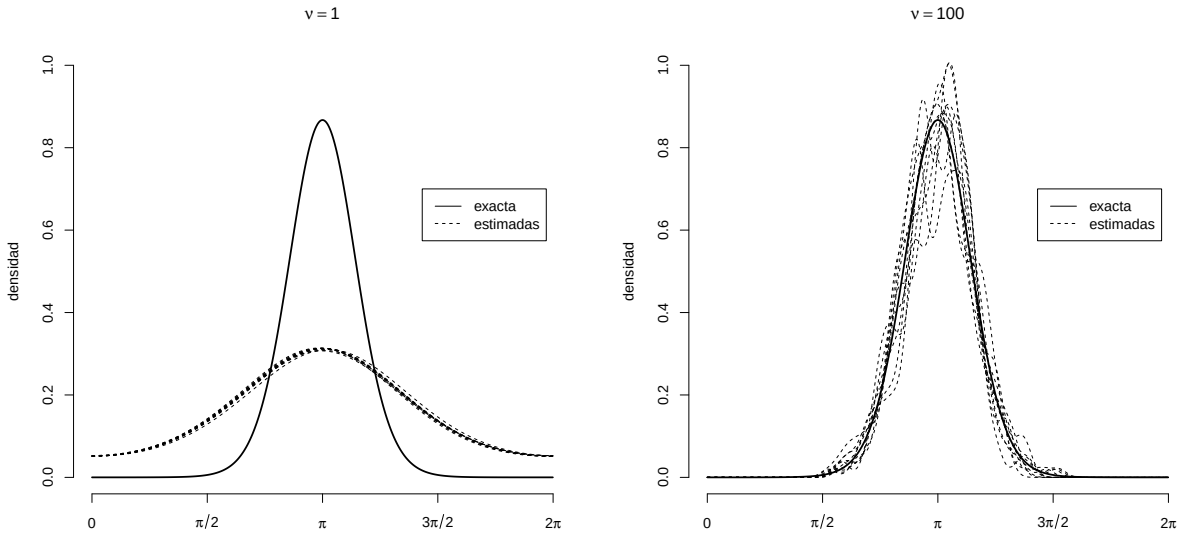


Figura 2.3: Estimación tipo núcleo de la densidad circular construida a partir de 100 muestras diferentes de tamaño 200 de una vM$(\pi, 5)$ con $\nu = 1$ (izquierda), sobresuavizado y $\nu = 100$ (derecha), infrasuavizado.

La selección del parámetro de suavizado o ventana es, por tanto un problema crucial, y el estudio de procedimientos que permitan elegir la ventana óptima según algún criterio constituye uno de los problemas más importantes en la estadística no paramétrica. En todos los procedimientos se escoge el parámetro de suavizado de manera que minimice alguna medida de distancia entre la función de densidad verdadera y su estimador. Sin embargo, no existe ningún método que proporcione una expresión para la ventana óptima de manera que en su cálculo sólo se utilice la información proporcionada por la muestra, puesto que la función de discrepancia depende de la curva teórica.

Existe una gran variedad de técnicas de selección automática de la ventana: métodos plug-in, de validación cruzada (*cross-validation*), métodos basados en el remuestreo bootstrap, etc. Además, en función de la medida de distancia, el selector puede ser de dos tipos: global o puntual, según el interés se centre en la estimación de la curva completa o en un punto particular.

Para un punto θ fijo, $\hat{f}(\theta, \nu)$ es una variable aleatoria. Para medir su calidad como estimador

de $f(\theta)$ se puede utilizar el *Error Cuadrático Medio* (MSE, *Mean Squared Error*):

$$MSE\,(\theta,\nu) = \mathbb{E}\left(\hat{f}(\theta;\nu) - f(\theta)\right)^2.$$

Este criterio de error afronta el problema de la estimación no paramétrica de una forma puntual. Sin embargo, el interés de la estimación no paramétrica radica en obtener una estimación y representación de la densidad completa, por tanto se hace necesario recurrir a criterios de error globales como puede ser el *Error Cuadrático Integrado* (ISE, *Integrated Squared Error*):

$$ISE\,(\nu;\theta_1,\ldots,\theta_n) = \int \left(\hat{f}(\theta;\nu) - f(\theta)\right)^2 d\theta.$$

El ISE es una variable aleatoria que depende de la verdadera (y desconocida) densidad y del parámetro de concentración ν. Incluso con estas dos cantidades fijadas, el ISE es una función de una realización particular de n puntos. Por tando, resulta más realista plantearse como criterio de error un promedio del ISE sobre las diversas realizaciones

$$MISE\,(\nu) = \mathbb{E}\left[\int \left(\hat{f}(\theta;\nu) - f(\theta)\right)^2 d\theta\right]$$

que, intercambiando la esperanza con la integral, no es más que un promedio de los errores cuadráticos medios en cada punto

$$MISE\,(\nu) = \int MSE(\theta,\nu)d\theta.$$

El *Error Cuadrático Medio Integrado* ($MISE$, *Mean Integrated Squared Error*) es una medida de distancia global entre la función de densidad verdadera y su estimador, probablemente la más estudiada y usada en la práctica.

Regla Plug-in

En el espacio Euclídeo, Silverman (1986) propone un selector de la ventana óptima minimizando una versión asintótica del $MISE$ ($AMISE$, *Asymptotic Mean Integrated Squared Error*). Esa expresión asintótica depende de la función núcleo y de la integral de la derivada segunda al cuadrado de la densidad desconocida f, haciendo esto último que la ventana no sea calculable directamente. Silverman (1986) resuelve este problema suponiendo que f es

una función de densidad Normal con media μ y desviación típica σ. Tomando como función núcleo la densidad gaussiana, proporciona el selector plug-in:

$$h_{PI} = 1.06\sigma n^{-1/5}. \tag{2.11}$$

El valor de σ se estima a partir de los datos observados mediante

$$\hat{\sigma} = \text{mín}\left(S, \frac{RIC}{\Phi^{-1}(0.75) - \Phi^{-1}(0.25)}\right), \tag{2.12}$$

donde S^2 es la varianza muestral de los datos, RIC es su rango intercuartílico y Φ^{-1} es la función cuantil de la normal estándar.

A continuación veremos como obtener una regla plug-in similar para la estimación de la densidad en el círculo. En este caso tomaremos como referencia la distribución von Mises.

Bajo la suposición de que los datos siguen una distribución von Mises con parámetro de concentración κ y, sin pérdida de generalidad, dirección media $\mu = 0$ y tomando como estimador el definido por la ecuación (2.10), en Taylor (2008) se obtiene la expresión asintótica del $MISE$, supuesto $\nu \to \infty$ y $n\nu^{-1/2} \to \infty$ cuando $n \to \infty$:

$$MISE\,(\nu) = \frac{3\kappa^2 I_2(2\kappa)}{32\pi\nu^2 I_0(\kappa)^2} + \frac{\nu^{1/2}}{2n\pi^{1/2}} + O(\nu^{-4}) + o\left(\frac{\nu^{1/2}}{n}\right).$$

Por tanto, se define el $AMISE$ como:

$$AMISE\,(\nu) = \frac{3\kappa^2 I_2(2\kappa)}{32\pi\nu^2 I_0(\kappa)^2} + \frac{\nu^{1/2}}{2n\pi^{1/2}}. \tag{2.13}$$

La expresión anterior puede minimizarse derivando con respecto a ν e igualando a cero. Esto conduce a la regla plug-in basada en la distribución von Mises para el parámetro de suavizado ν del núcleo:

$$\hat{\nu} = \left[\frac{3n\hat{\kappa}^2 I_2(2\hat{\kappa})}{4\pi^{1/2} I_0(\hat{\kappa})^2}\right]^{2/5}, \tag{2.14}$$

donde $\hat{\kappa}$ es una estimación del parámetro κ de la densidad a estimar. Esta estimación de κ puede obtenerse por el método de los momentos o por máxima verosimilitud, que tal como hemos visto coinciden, o por otro método robusto como el sugerido por Silverman en (2.12) para estimar σ.

En este punto se presentan dos cuestiones: ¿cómo es de buena en la práctica la versión asintótica del $MISE$ que aparece en (2.13)?; ¿qué sucede si los datos no proceden de la densidad de referencia (von Mises)?

Para dar respuesta a la primera de las preguntas, para la distribución von Mises con media 0 y parámetro de concentración $\kappa = 1$ y para tamaños muestrales $n = 50$ y $n = 500$, hemos comparado el $MISE$ obtenido mediante una aproximación de Montecarlo con 500 muestras con el $AMISE$ dado por la expresión (2.13). Los resultados se muestran en la Figura 2.4 y de ella se deduce que la aproximación es bastante buena, mejorando a medida que aumenta n.

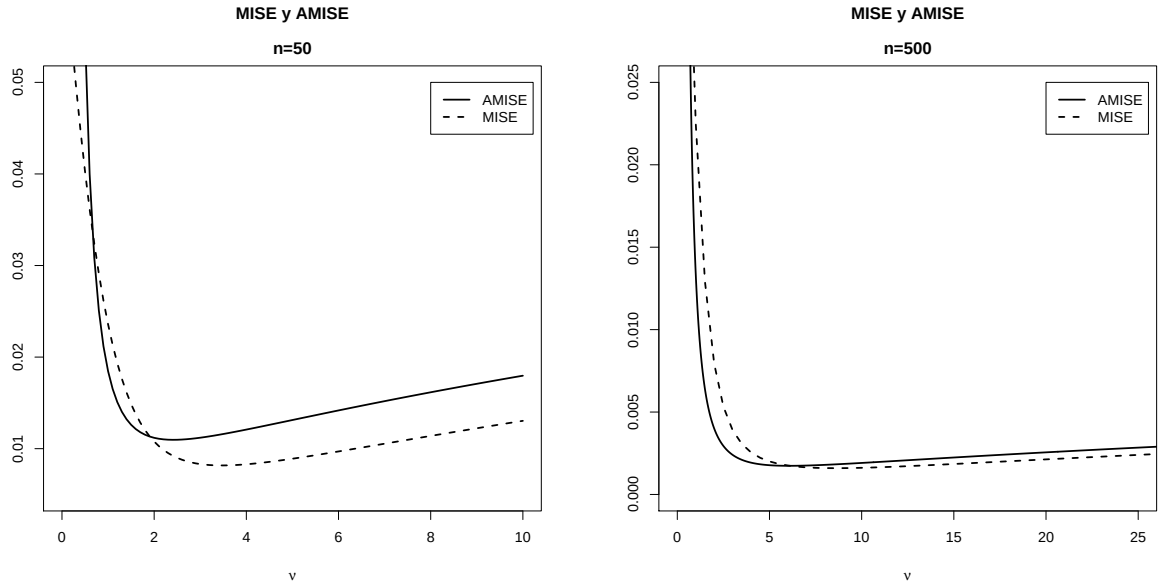


Figura 2.4: Aproximación de Montecarlo del $MISE$ (línea discontinua) y $AMISE$ (línea continua) para 500 simulaciones de tamaño $n = 50$ (izquierda) y $n = 500$ (derecha) de una $\mathrm{vM}(0, 1)$.

Cuando los datos son unimodales, la regla de selección dada anteriormente funciona razonablemente bien, como puede verse en la Figura 2.5. Sin embargo, para datos bimodales, la estimación de κ por máxima verosimilitud o por el método de los momentos (ambos métodos coinciden para la estimación de los parámetros de una von Mises), puede ser prácticamente inútil. En el caso más extremo de una mixtura de dos von Mises con igual probabilidad, una concentrada en torno a ϕ con otra concentrada en torno a $\phi + \pi$, proporcionará una estimación de κ próxima a cero. Cuando $\hat{\kappa} = 0$ entonces, por (2.14), se tendrá $\hat{\nu} = 0$, con lo cual $\hat{f}(\theta) \equiv 1/(2\pi)$, y así este método nos lleva a estimaciones erróneas de la densidad, como muestra la Figura 2.6.

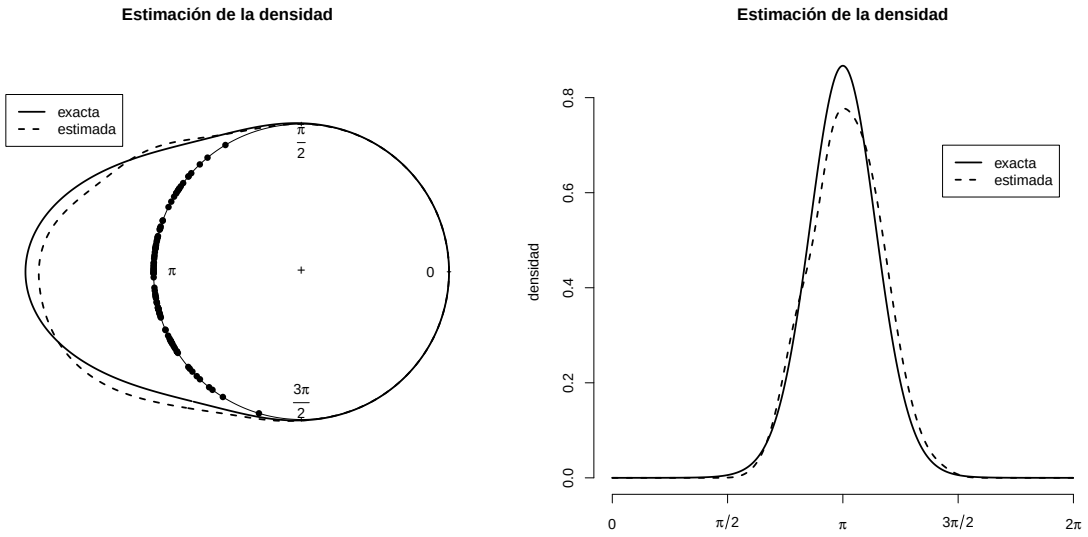


Figura 2.5: Representación circular (izquierda) y representación lineal (derecha) de la estimación tipo núcleo de la densidad de una $\text{vM}(\pi, 5)$ a partir de una muestra de tamaño 100, con parámetro de suavizado calculado con la regla plug-in.

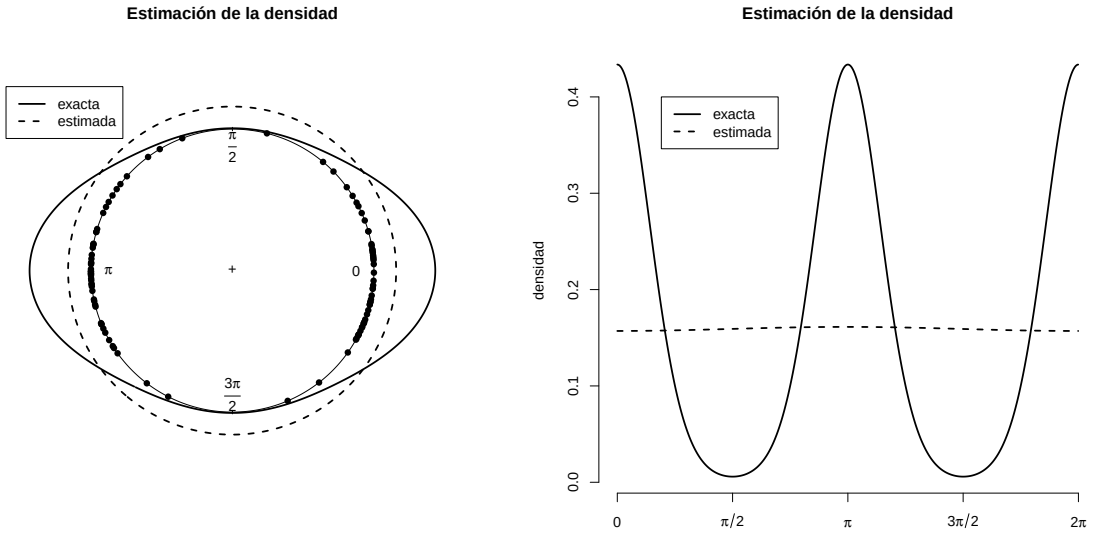


Figura 2.6: Representación circular (izquierda) y representación lineal (derecha) de la estimación tipo núcleo de la densidad de una mixtura, $\text{vM}(0, \pi, 5, 5, 0.5)$, a partir de una muestra de tamaño 100, con parámetro de suavizado calculado con la regla plug-in.

Como ya mencionamos anteriormente, existen otros métodos de selección del parámetro de suavizado, como son los métodos de validación cruzada, que introducimos a continuación.

Reglas de validación cruzada

Validación cruzada por mínimos cuadrados

La validación cruzada por mínimos cuadrados (*LSCV*, *Least Squares Cross-Validation*) es un método automático para seleccionar el parámetro de ventana sugerido por Rudemo (1982) y Bowman (1984). Se basa en escribir el $MISE$ de la forma siguiente:

$$\begin{aligned} MISE(\nu) &= \mathbb{E}\left[\int \left(\hat{f}(\theta;\nu) - f(\theta)\right)^2 d\theta\right] \\ &= \mathbb{E}\left[\int \hat{f}^2(\theta;\nu)d\theta\right] - 2\mathbb{E}\left[\int \hat{f}(\theta;\nu)f(\theta)d\theta\right] + \mathbb{E}\left[\int f^2(\theta;\nu)d\theta\right]. \end{aligned}$$

El último término no depende de la estimación $\hat{f}$, por tanto la elección de ν para minimizar el $MISE$ equivale a la minimización de

$$\mathbb{E}\left[\int \hat{f}^2(\theta;\nu)d\theta - 2\int \hat{f}(\theta;\nu)f(\theta)d\theta\right]. \tag{2.15}$$

Un estimador de $\int \hat{f}(\theta;\nu)f(\theta)d\theta$ viene dado por

$$\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(\theta_i;\nu), \tag{2.16}$$

donde $\hat{f}_{-i}(\theta;\nu)$ es la densidad estimada a partir de los datos extrayendo de la muestra el dato θ_i. Si estimamos la densidad con el estimador tipo núcleo dado en (2.10), entonces

$$\hat{f}_{-i}(\theta;\nu) = \frac{1}{(n-1)(2\pi)I_0(\nu)}\sum_{j\neq i}^{n} \exp\{\nu\cos(\theta-\theta_j)\}. \tag{2.17}$$

El estimador (2.16) es un estimador insesgado para $\mathbb{E}\left[\int \hat{f}(\theta,\nu)f(\theta)\right]$. Por tanto un estimador insesgado de (2.15) viene dado por

$$LSCV(\nu) = \int_0^{2\pi} \hat{f}^2(\theta;\nu)d\theta - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(\theta_i;\nu). \tag{2.18}$$

El estimador de validación cruzada por mínimos cuadrados del parámetro de suavizado, ν_{LSCV}, será el valor que minimice la expresión anterior.

Validación cruzada por máxima verosimilitud

La validación cruzada por máxima verosimilitud (*LCV*, *Likelihood Cross-Validation*), introducida por Habema, Hermans y Van den Broek (1974), se basa en considerar a ν como un parámetro que es estimado por máxima verosimilitud. La función de verosimilitud, dado que la densidad verdadera es desconocida, es estimada por

$$LCV(\nu) = \prod_i \hat{f}_{-i}(\theta_i;\nu), \tag{2.19}$$

donde $\hat{f}_{-i}$ es el estimador definido en (2.17).

El estimador de validación cruzada por máxima verosimilitud del parámetro de suavizado, ν_{LCV}, será el valor que maximice la expresión anterior.

2.3.2. Estudio de simulación

Hemos analizado el comportamiento de la regla plug-in y de los métodos de validación cruzada, cuando los datos provienen de una distribución von Mises o de una mixtura de dos distribuciones von Mises. Como modelos de prueba se han considerado las densidades de las siguientes distribuciones: vM$(0,1)$, vM$(0,0.1)$, vM$(\pi,1)$, vM$(0,\pi,4,4,\frac{1}{2})$, vM$(0,\frac{\pi}{2},5,5,\frac{1}{5})$, vM$(2,4,5,5,\frac{1}{2})$ y vM$(0,\frac{\pi}{\sqrt{3}},2,2,\frac{1}{4})$ (los cuatro últimos modelos están representados en la Figura 2.2).

Para llevar a cabo este estudio de simulación, se han generado $B = 500$ muestras de tamaño $n = 50$ y $n = 500$ de cada una de las densidades.

Para cada modelo, mediante una aproximación de Montecarlo con $B = 500$ muestras hemos calculado el $MISE(\nu)$ para una rejilla de ventanas (usando una rejilla de 500 puntos para evaluar las integrales numéricamente), y hemos obtenido el valor de ν que minimiza $MISE(\nu)$ y que denotamos por ν_0.

También, para cada modelo, se ha calculado el promedio de los ISE cuando ν se obtiene,

para cada conjunto de datos, a través de cada uno de los métodos de selección de ventana propuestos: plug-in (ν_{PI}), validación cruzada cuadrática (ν_{LSCV}) y validación cruzada por máxima verosimilitud (ν_{LCV}). Cabe notar que los valores que se muestran para ν_{PI} se han obtenido estimando κ por máxima verosimilitud.

Además de los métodos mencionados anteriormente se incluyen los resultados de una adaptación del método propuesto en Taylor (2008) basado en la estimación del rango intercuartílico, similar a la idea de Silverman para estimar la ventana para datos lineales (ver en (2.12) la expresión de $\hat{\sigma}$), y que consiste en:

1. Calcular cualquiera de los arcos de menor longitud que contiene el 50 % de los datos.
2. Calcular $\hat{\kappa}$ de forma que la probabilidad de una von Mises centrada en el punto medio de dicho arco sea 0.5.
3. Con el valor de $\hat{\kappa}$ obtenido en el paso anterior, utilizar (2.14) para obtener el parámetro de suavizado.

A la ventana obtenida mediante este procedimiento la denotamos por ν_{RIC}.

Para poder comparar los diferentes métodos se ha calculado la siguiente medida relativa del error

$$Efi(\nu_\bullet) = \left(\frac{MISE(\nu_\bullet)}{MISE(\nu_0)} - 1\right) 100\,\%, \tag{2.20}$$

donde $\nu_\bullet$ denota el valor del parámetro de suavizado obtenido por el método de selección considerado (PI, $LSCV$, LCV o RIC). Dicho porcentaje informa acerca de la eficiencia de cada método. Así, cuánto más pequeño sea el valor de $Efi(\nu_\bullet)$ mejor será el método de selección de la ventana.

Con el objetivo de comparar los métodos no paramétricos de estimación de la densidad circulares con los lineales también se proporcionan los resultados de estimar la función de densidad con el estimador lineal definido en (2.9), donde se ha considerado el núcleo gaussiano y el parámetro h se ha seleccionado utilizando el selector plug-in (h_{PI}) para datos lineales definido en (2.11) y mediante el método de Seather y Jones en dos etapas (ver Seather y Jones, 1991). En este caso, en la definición (2.20) tendríamos que reemplazar $\nu_\bullet$ por $h_\bullet$ (h_{PI}, h_{SJ}).

La Tabla 2.1 muestra los resultados de calcular la eficiencia, definida en (2.20), para las distribuciones von Mises (no mixturas). Solamente ν_{PI} da resultados razonables para las tres

distribuciones. La validación cruzada por mínimos cuadrados sólo da buenos resultados para el modelo vM$(\pi, 1)$. La validación cruzada por máxima verosimilitud da malos resultados cuando la distribución está próxima a una uniforme. El método basado en la estimación del rango intercuartílico funciona razonablemente bien excepto para el modelo vM$(0, 0.1)$ y tamaño muestral $n = 50$. Los métodos lineales dan buenos resultados cuando la distribución von Mises no está concentrada en los extremos del intervalo $(0, 2\pi]$, como es el caso de la vM$(\pi, 1)$.

Modelo	vM$(0,1)$		vM$(0,0.1)$		vM$(\pi,1)$	
n	50	500	50	500	50	500
ν_{PI}	35.55	15.46	71.79	63.41	37.48	15.45
ν_{LSCV}	239.05	1317.23	693.31	221.70	54.10	18.43
ν_{LCV}	46.25	21.69	941.09	159.42	45.80	19.64
ν_{RIC}	29.02	12.02	221.89	41.43	26.64	12.37
h_{PI}	643.50	1935.68	1461.15	1277.33	9.32	9.07
h_{SJ}	282.88	551.47	1428.12	988.36	12.14	11.56

Tabla 2.1: Porcentaje $Efi(\nu_\bullet)$ para cada método de selección de la ventana para tres modelos de distribuciones von Mises.

Modelo	vM$(0,\pi,4,4,\frac{1}{2})$		vM$(0,\frac{\pi}{2},5,5,\frac{1}{5})$		vM$(2,4,5,5,\frac{1}{2})$		vM$(0,\frac{\pi}{\sqrt{3}},2,2,\frac{1}{4})$	
n	50	500	50	500	50	500	50	500
ν_{PI}	471.62	3118.44	29.47	49.14	144.81	332.47	25.78	22.20
ν_{LSCV}	126.54	777.58	29.49	40.99	20.25	6.52	25.13	7.35
ν_{LCV}	19.35	7.60	22.11	16.81	19.43	8.99	-19.20	17.49
ν_{RIC}	151.85	1728.24	16.55	11.91	93.09	627.74	19.25	16.41
h_{PI}	431.95	1642.44	25.20	55.03	94.26	205.62	47.30	112.07
h_{SJ}	202.25	287.81	42.28	54.54	11.61	3.61	34.26	88.24

Tabla 2.2: Porcentaje $Efi(\nu_\bullet)$ para cada método de selección de la ventana para cuatro modelos de mixturas de von Mises.

En la Tabla 2.2, se muestran los resultados para las mixturas de von Mises. Para las mixturas, la regla plug-in, ν_{PI}, puede dar estimaciones poco satisfactorias. El método de validación cruzada por máxima verosimilitud se comporta bien en estos casos, mientras que el método de validación cruzada mínimo cuadrática no proporciona buenos resultados para una de las mixturas. La ventana ν_{RIC} proporciona buenos resultados para dos de las mixturas, aquéllas que tienen proporción de mixtura $p \neq 1/2$. Este compartamiento se ha observado

con más modelos lo que nos permite decir que la elección del parámetro de suavizado por este método proporciona buenos resultados cuando $p \neq 1/2$. Igual que antes, para las mixturas, los métodos lineales se comportan bien cuando la distribución no está concentrada en los extremos del intervalo $[0, 2\pi)$. Nótese que aunque en la distribución $\text{vM}(0, \pi/2, 5, 5, 1/5)$ interviene la distribución $\text{vM}(0, 5)$ (concentrada en 0), los métodos lineales proporcionan resultados buenos porque la proporción en la que se encuentra la $\text{vM}(0, 5)$ es pequeña ($p = 0.2$).

A la vista de los resultados de las simulaciones, ningún método de selección del parámetro de suavizado proporciona buenos resultados en todos los casos. En general, para distribuciones unimodales la regla plug-in es la que mejor se comporta, mientras que para las mixturas parece mejor el criterio de validación cruzada por máxima verosimilitud. Debido a qué ningún método destaca claramente sobre los demás, antes de estimar la función de densidad de un conjunto de datos podría ser útil hacer un análisis descriptivo previo de los mismos y probar con varios métodos de selección del parámetro de suavizado.

2.4. Estimación local lineal de la función de regresión con covariable circular y respuesta lineal

En muchas situaciones, se tienen datos de dos o más variables, donde alguna o todas ellas son angulares. Por ejemplo, en el estudio de la migración de aves, podría observarse la dirección de donde vienen las aves, así como la dirección de su regreso, dando lugar a observaciones sobre un toro, o puede registrarse, tanto la dirección del viento como la dirección del vuelo de aves migratorias. En estos casos, se podría estar interesado en cuestiones de correlación entre variables tales como la regresión con el objetivo de predecir una variable dadas otras. En esta sección veremos como estimar no paramétricamente la función de regresión cuando se tiene una variable respuesta lineal y la variable explicativa es circular.

2.4.1. Estimador local lineal circular

Consideremos el conjunto de datos $\{(\theta_i, Y_i), i = 1, \ldots, n\}$, donde θ_i e Y_i son observaciones de variables aleatorias absolutamente continuas tomando valores en $[0, 2\pi)$ y $\mathbb{R}$, respectivamente. Desde este momento asumiremos que

$$Y_i = m(\theta_i) + \epsilon_i, \ i = 1, \ldots, n \tag{2.21}$$

donde ϵ_i son variables aleatorias con media cero y varianza σ^2. El objetivo es construir un estimador de $m(\theta)$, esto es, una función del conjunto de datos cuando θ_i y ϵ_i son ambos independientes e idénticamente distribuidos.

Cuando ambas variables son lineales el estimador local lineal consiste en estimar la función de regresión ajustando localmente una recta por mínimos cuadrados ponderados. La idea es la siguiente: supongamos que existe la primera derivada de $m(\cdot)$. En un entorno de un punto x, y mediante el desarrollo de Taylor en x, se tiene que

$$m(t) \approx m(x) + m'(x)(t-x) = \beta_0 + \beta_1(t-x), \tag{2.22}$$

para t próximo a x, donde $m(x) = \beta_0$ y $m'(x) = \beta_1$.

Entonces, dada una muestra $\{(X_i, Y_i), i = 1, \ldots, n\}$, donde X_i e Y_i son observaciones de variables aleatorias absolutamente continuas tomando valores en $\mathbb{R}$, (2.22) sugiere la estimación de $m(\cdot)$ encontrando los parámetros β_0 y β_1 que minimizan la suma de los residuos al cuadrado de un modelo de regresión localmente lineal (un polinomio local de grado $p = 1$). Por tanto se buscarán $\hat{\beta}_0(x)$ y $\hat{\beta}_1(x)$ que minimicen

$$\hat{\beta} = \left(\hat{\beta}_0(x), \hat{\beta}_1(x)\right) = \arg\min_{(a,b)} \sum_{i=1}^{n} L_{ih}(x - X_i) \left[Y_i - (a + b(X_i - x))\right]^2,$$

donde $L_h(u) = h^{-1}L(u/h)$. Así, el estimador local lineal de la función de regresión m se define como $\hat{m}_{LL}(x) = \hat{\beta}_0(x)$.

Si en vez de ajustar una recta ajustásemos una constante tendríamos el estimador de Nadaraya-Watson.

El estimador local lineal se puede generalizar al caso de que la variable explicativa sea circular, ajustando localmente la función de regresión por el polinomio trigonométrico de grado uno

$$m(\omega) \approx \beta_0 + \beta_1 \operatorname{sen}(\omega - \theta),$$

ver Di Marzio, Panzera y Taylor (2009). Dado que $\operatorname{sen}(\theta) \approx \theta$ para valores muy pequeños de θ, los parámetros $\hat{\beta}_0(\theta)$ y $\hat{\beta}_1(\theta)$ se determinan de manera análoga a como se hace para datos lineales, pero ahora tomando como función núcleo, por ejemplo, la densidad von Mises:

$$\hat{\beta} = \left(\hat{\beta}_0(\theta), \hat{\beta}_1(\theta)\right) = \arg\min_{(a,b)} \sum_{i=1}^{n} K_\nu(\theta - \theta_i)[Y_i - (a + b\,\mathrm{sen}(\theta_i - \theta))]^2,$$

donde por K_ν estamos denotando la densidad von Mises con media 0 y parámetro de concentración ν.

Utilizando la notación matricial:

$$\Theta := \begin{pmatrix} 1 & \mathrm{sen}(\theta_1 - \theta) \\ \vdots & \vdots \\ 1 & \mathrm{sen}(\theta_n - \theta) \end{pmatrix},$$

$Y := (Y_1, \ldots, Y_n)^T$ y $W := diag\{K_\nu(\theta - \theta_1), \ldots, K_\nu(\theta - \theta_n)\}$, asumiendo la no singularidad de $\Theta^T W \Theta$, la teoría de mínimos cuadrados ponderados proporciona $\hat{\beta} = (\Theta^T W \Theta)^{-1} \Theta^T W Y$, y por tanto, el estimador tipo núcleo local lineal de $m(\theta)$ está dado por la primera entrada de dicho vector

$$\hat{m}_{LL_{cir}}(\theta) = e_1^T (\Theta^T W \Theta)^{-1} \Theta^T W Y, \tag{2.23}$$

donde $e_1 = (1, 0)^T$.

Si en lugar de ajustar un polinomio de grado uno, ajustáramos un polinomio de grado 0 tendríamos el estimador de Nadaraya-Watson cuando la variable explicativa es circular:

$$\hat{m}_{NW_{cir}}(\theta) = \frac{\sum_{i=1}^{n} Y_i K_\nu(\theta - \theta_i)}{\sum_{i=1}^{n} K_\nu(\theta - \theta_i)}, \tag{2.24}$$

que es el análogo al estimador de Nadaraya-Watson para datos lineales.

2.4.2. Selección del parámetro de suavizado

Al igual que en la estimación de la densidad la elección del parámetro de suavizado es de crucial importancia en la estimación de la función de regresión como se puede ver en la Figura 2.7. Valores de ν grandes infrasuavizan la función de regresión mientras que valores pequeños de este parámetro proporcionan una estimación sobresuavizada.

Una opción muy común en regresión no paramétrica es seleccionar ν por validación cruzada mínimo cuadrática. Para ello, consideremos n observaciones de una variable explicativa

circular $\theta_1, \ldots, \theta_n$ y una respuesta lineal $Y_1, \ldots, Y_n$. Como en el caso lineal, se puede elegir ν de forma que se minimice la función

$$CV(\nu) = \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \hat{m}^{-i}(\theta_i) \right]^2, \tag{2.25}$$

donde $\hat{m}^{-i}$, $i = 1, \ldots, n$ denota el estimador de la función de regresión construido a partir de la muestra original después de eliminar el par (θ_i, Y_i).

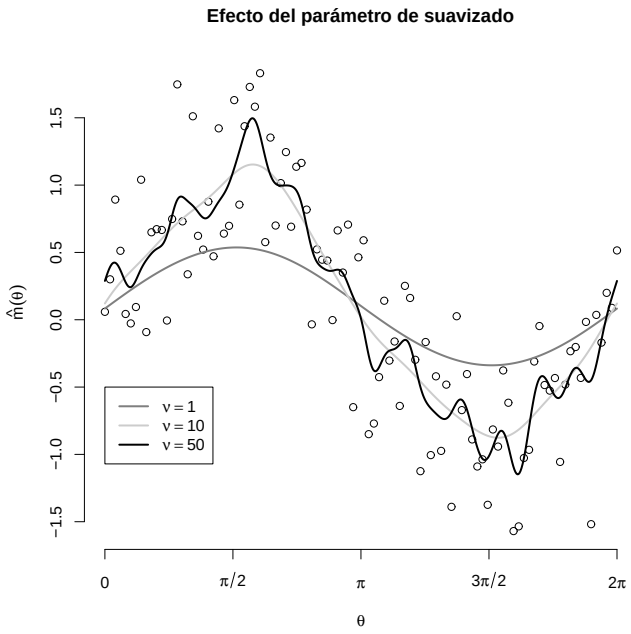


Figura 2.7: Efecto del parámetro de suavizado en la estimación no paramétrica de la función de regresión con $\nu = 1$, $\nu = 10$ y $\nu = 50$.

2.4.3. Estudio de simulación

En esta sección se presentan los resultados de un estudio de simulación con cuatro modelos diferentes, de acuerdo a distintas formas de la función $m(\cdot)$, para comparar el comportamiento en la práctica de los estimadores local-lineal circular (LL_{cir}) y Nadaraya-Watson circular (NW_{cir}). Además hemos comparado estos estimadores con los análogos para datos lineales (LL y NW respectivamente), con el objetivo de ver si los estimadores adaptados a datos circulares proporcionan una mejora substancial con respecto a los estimadores estándar.

Como criterio de error se ha utilizado

$$ISE_f(\nu;\theta_1,\ldots,\theta_n) = \int (\hat{m}(\theta) - m(\theta))^2 f(\theta) d\theta, \tag{2.26}$$

donde las diferencias entre la función de regresión teórica y la estimada están ponderadas por la función de densidad f que genera el diseño de la covariable.

Se han considerado cuatro modelos distintos, de diversa complejidad (ver Figura 2.8):

- Modelo 1. $m_1(\theta) = \text{sen}(\theta)$.
- Modelo 2. $m_2(\theta) = \cos(\theta)$.
- Modelo 3. $m_3(\theta) = \text{sen}\left(\frac{3}{2}\left(\theta - \frac{\pi}{2}\right)\right) + \frac{2\sqrt{2}}{3}\cos\left(\frac{\theta}{3}\right)$.
- Modelo 4. $m_4(\theta) = 2 + \text{sen}(\theta - 1.2\pi) + 3\exp\left\{-10\left(15\frac{(\theta-\pi)}{2\pi}\right)^2\right\}$.

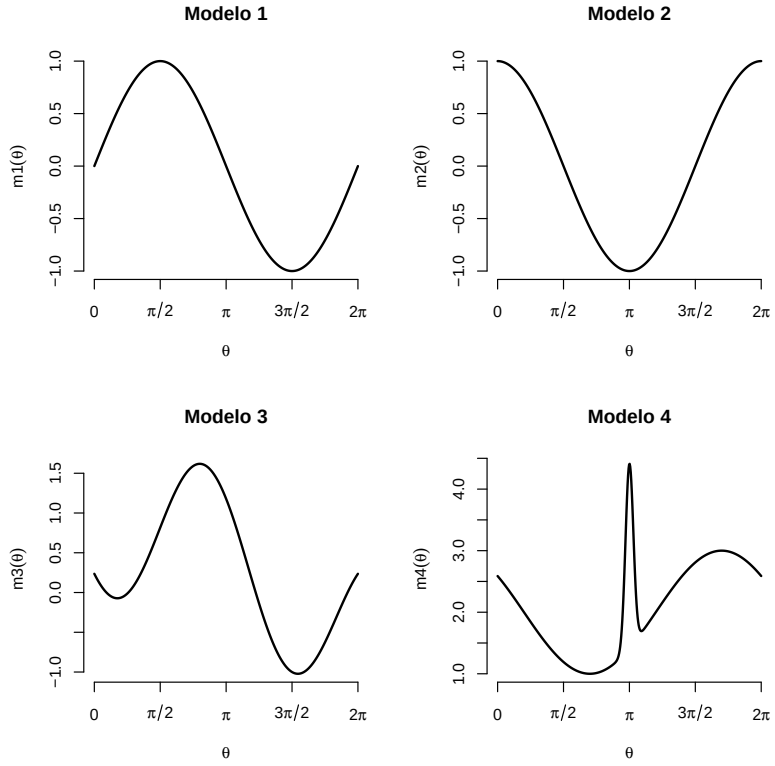


Figura 2.8: Modelos de regresión considerados en el estudio de simulación.

Para la variable explicativa circular, se han considerado dos diseños diferentes:

- Diseño 1. Valores equiespaciados $\theta_1, \ldots, \theta_n$ en $[0, 2\pi)$.
- Diseño 2. Los cuantiles de una distribución vM$(0, 2)$.

Los errores ϵ_i del modelo (2.21) se han tomado normales con media cero y varianza 0.5.

Para cada modelo y diseño se han generado $B = 500$ muestras de tamaño $n = 50$ y $n = 100$ y se han evaluado los cuatro estimadores analizados, seleccionando en todos los casos el parámetro de suavizado ν por validación cruzada. Las Tablas 2.3 y 2.4 muestran los promedios del ISE_f, la primera para el diseño uniforme (Diseño 1) y la segunda para el diseño por cuantiles (Diseño 2).

Modelo	n	NW	LL	NW_{cir}	LL_{cir}
1	50	0.065	0.065	0.061	0.060
	100	0.037	0.036	0.032	0.036
2	50	0.063	0.063	0.065	0.065
	100	0.033	0.035	0.035	0.036
3	50	0.079	0.077	0.077	0.077
	100	0.043	0.044	0.042	0.042
4	50	0.210	0.213	0.202	0.200
	100	0.127	0.132	0.123	0.123

Tabla 2.3: Promedio de los ISE_f de los cuatro estimadores de la función de regresión para los modelos 1-4 con tamaños muestrales $n = 50$ y $n = 100$ cuando la variable explicativa toma valores equiespaciado en el intervalo $[0, 2\pi)$.

Modelo	n	NW	LL	NW_{cir}	LL_{cir}
1	50	0.078	0.057	0.059	0.044
	100	0.041	0.049	0.036	0.026
2	50	0.074	0.066	0.062	0.056
	100	0.040	0.046	0.035	0.037
3	50	0.085	0.071	0.069	0.064
	100	0.052	0.067	0.047	0.043
4	50	0.209	0.438	0.197	0.348
	100	0.128	0.195	0.141	0.167

Tabla 2.4: Promedio de los ISE_f de los cuatro estimadores de la función de regresión para los modelos 1-4 con tamaños muestrales $n = 50$ y $n = 100$ cuando la variable explicativa toma los valores de los cuantiles de una vM$(0, 2)$.

Como era de esperar, el error disminuye al aumentar el tamaño muestral, en ambos casos.

Los resultados de la Tabla 2.3 muestran un comportamiento similar de los métodos lineales y circulares para los cuatro modelos. Los resultados de la Tabla 2.4 muestran un comportamiento mejor de los métodos circulares frente a los lineales, obteniendo un ISE_f menor en los cuatro modelos considerados.

Capítulo 3

Análisis de datos de glaciares

3.1. Introducción

En este capítulo se ilustran las técnicas de estimación no paramétricas de la función de densidad circular y de la regresión lineal-circular expuestas en el capítulo anterior, mediante su aplicación a un conjunto de datos reales procedentes del campo de la Geografía.

Los datos analizados son los relativos al proyecto "*Cartografía y monitorización de formas crionivales en la región sub-antártica: Andes Fueguinos e Isla de los Estados (Tierra del Fuego, Argentina)*".

3.2. Datos de glaciares

Por una parte, se tienen datos de la temperatura a nivel de la superficie y a diferentes niveles de profundidad en varias localizaciones del monte Alvear - Calm 1, Calm 2, Calm 3 y Calm 4 - durante los años 2008 y 2009. Por otra parte se tienen datos de la estación meteorológica Vinciguerra (Ushuaia, Argentina) también durante los años 2008 y 2009, donde se registraron medidas de varias variables atmosféricas, entre ellas, la temperatura del aire medida en grados centígrados (ºC) y la dirección del viento medida en grados tomando como "cero" el Norte.

Antes de comenzar con el análisis de los datos, cabe comentar algunos detalles sobre ellos. Los datos de temperatura recogidos en las diferentes localizaciones del monte Alvear fueron tomados por el propio equipo de investigación, que instaló los dispositivos de medición de los datos en el mes de Febrero de 2008, con lo cual no se tienen datos anteriores a la

fecha de instalación, es decir, los correspondientes a Enero de 2008. A finales de Enero del año siguiente volvieron al lugar para descargar los datos recogidos por el aparato hasta el momento y volvieron a ponerlo en su sitio para seguir registrando datos. Es por esto que existe un pequeño período de tiempo, entre finales de Enero y principios de Febrero de 2009, en el que no se han registrado datos o en el que algunos de los datos registrados no son reales ya que se ha manipulado el aparato.

Además, aunque generalmente el aparato se programaba para que tomase las medidas de forma horaria, existe un caso en el que los datos fueron recogidos cada dos horas y un segundo. Se trata de los datos recogidos en Calm 1 a partir de Febrero de 2009. En este caso se aplicó una técnica de suavizado a dichos datos y luego se tomó la temperatura estimada a cada hora.

Los datos de temperatura del aire y dirección del viento recogidos en la estación meteorológica Vinciguerra no fueron tomados por el equipo de investigación, de ahí que no podamos explicar el motivo de los datos faltantes.

Cabe mencionar que todos los datos que analizamos en esta memoria fueron depurados y se les ha realizado un análisis descriptivo previo que no se incluye en este documento.

A continuación mostraremos un breve resumen de los datos recogidos en cada localización.

CALM 1

En esta ubicación se recogen las temperaturas a cuatro niveles, en la superficie (0 cm) y a 10, 20 y 60 cm de profundidad, durante los años 2008 y 2009.

En la Figura 3.1 se representan la evolución de las temperaturas en cada nivel en el que fue medida la temperatura para los años 2008 y 2009, respectivamente.

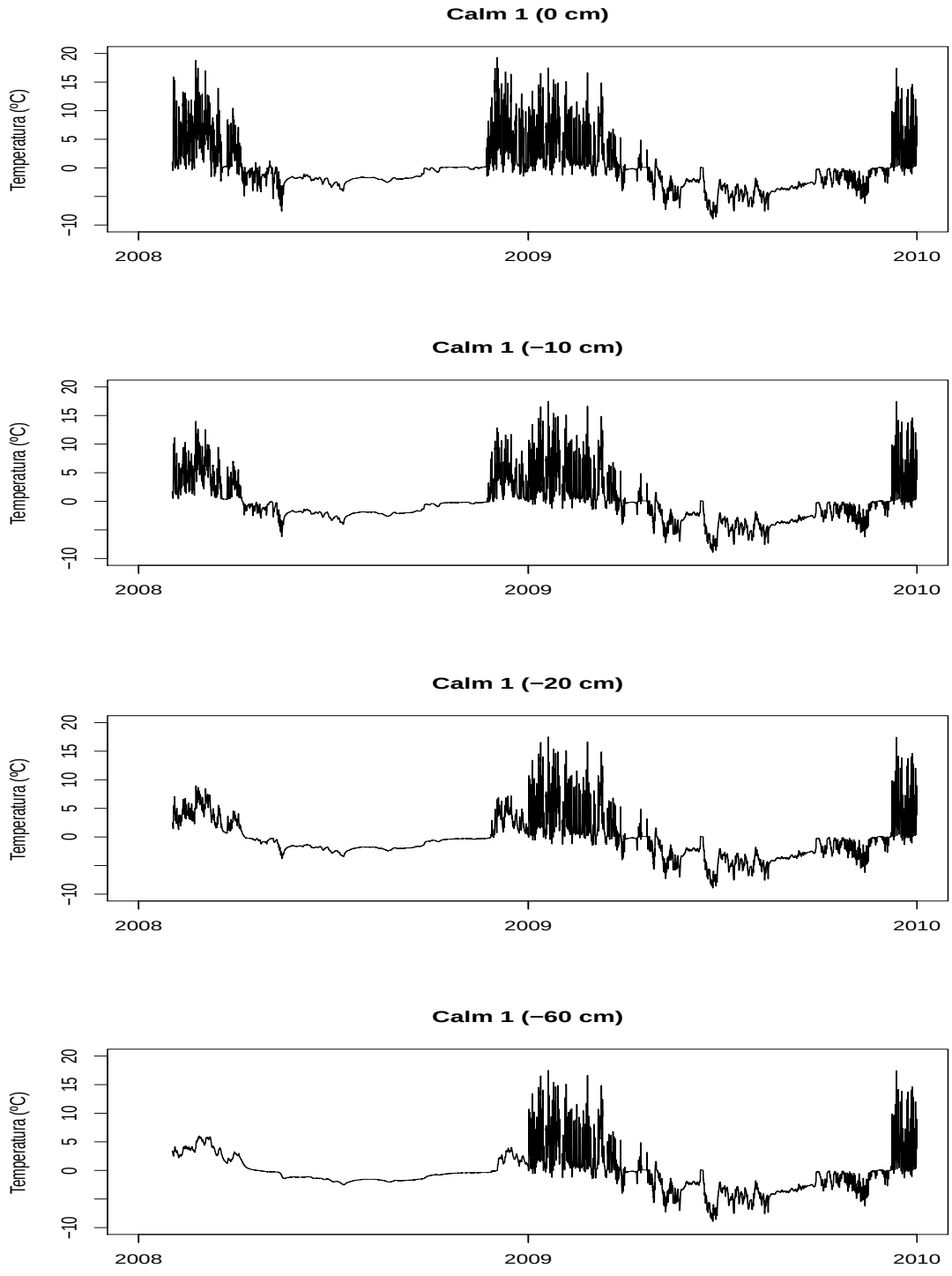


Figura 3.1: De arriba abajo, evolución de las temperaturas en las distintas posiciones del sensor (superficie, 10 cm de profundidad, 20 cm de profundidad y 60 cm de profundidad) en Calm 1 durante los años 2008-2009.

CALM 2

Para esta ubicación se recogen las temperaturas a dos niveles, en la superficie (0 cm) y a 50 cm de profundidad.

En la Figura 3.2 se muestra gráficamente la evolución de las temperaturas para cada nivel de profundidad:

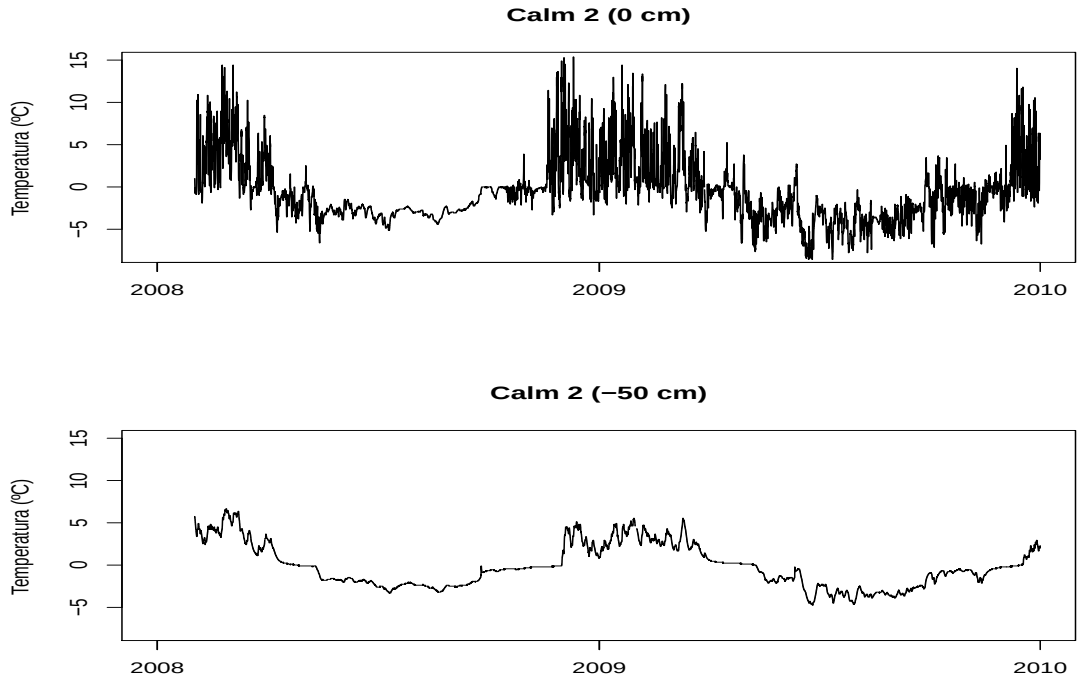


Figura 3.2: Evolución de las temperaturas medidas a nivel de la superficie (panel superior) y a 50 cm de profundidad (panel inferior) en Calm 2 durante los años 2008-2009.

CALM 3

Para esta ubicación también se recogen las temperaturas a dos niveles, en la superficie (0 cm) y a 50 cm de profundidad. En el año 2009, el aparato situado a nivel de la superficie falla y en consecuencia no se tienen datos a partir del 28 de Noviembre, ver Figura 3.3.

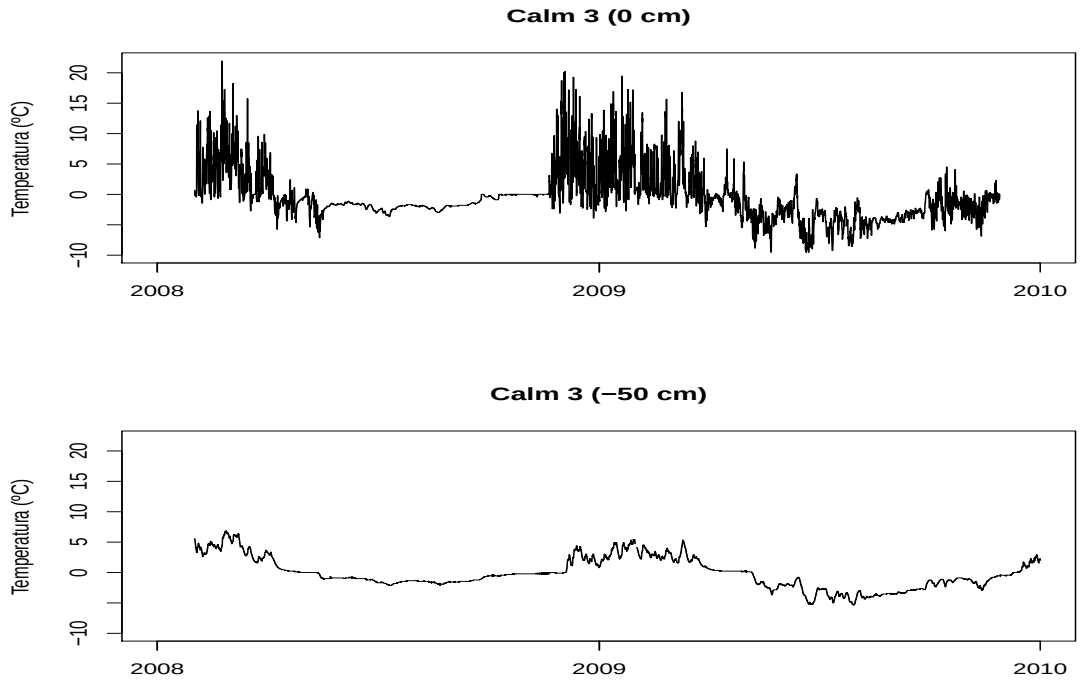


Figura 3.3: Evolución de las temperaturas medidas a nivel de la superficie (panel superior) y a 50 cm de profundidad (panel inferior) en Calm 3 durante los años 2008-2009.

CALM 4

Al igual que en Calm 2 y Calm 3, en Calm 4 se registra la temperatura en la superficie y a 50 cm de profundidad. El aparato instalado a 50 cm de profundidad en el año 2008 falla, registrando sólo veinte datos, con lo cual no se tienen datos de la temperatura a esta profundidad para el año 2008 y tampoco de Enero del año 2009 (ver Figura 3.4).

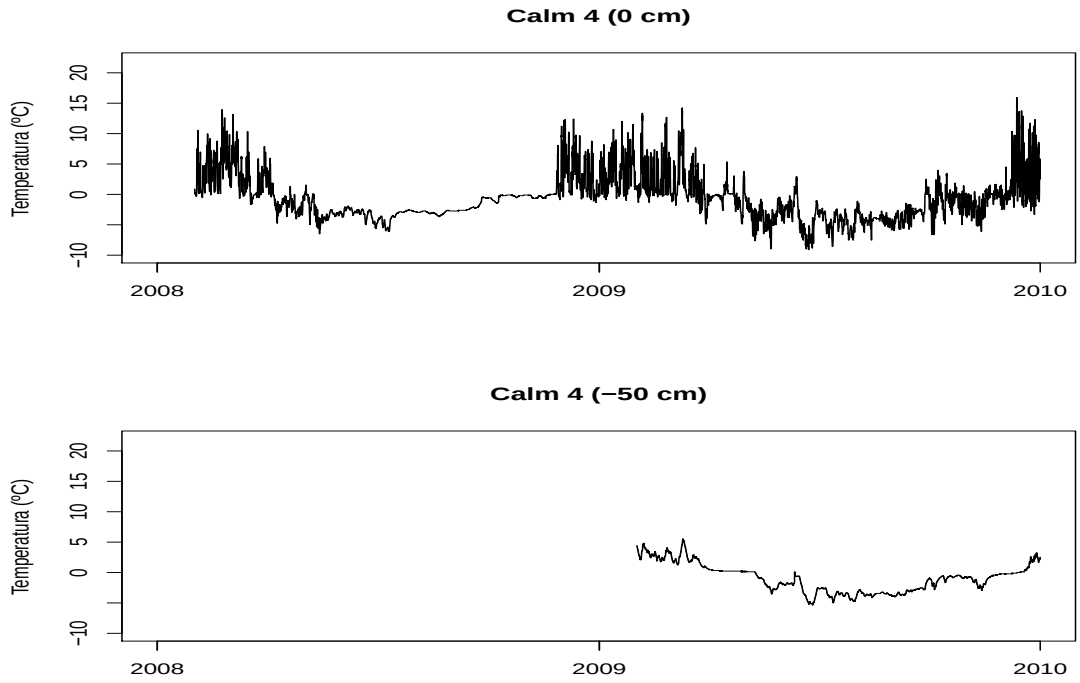


Figura 3.4: Evolución de las temperaturas medidas a nivel de la superficie (panel superior) y a 50 cm de profundidad (panel inferior) en Calm 4 durante los años 2008-2009.

Estación meteorológica Vinciguerra

Como ya se ha comentado, en esta estación meteorológica situada en el glaciar Vinciguerra (Ushuaia, Argentina) se registran datos de varias variables atmosféricas durantes los años 2008 y 2009. En este trabajo haremos uso solamente de los datos de la dirección del viento y de la temperatura del aire. En ambos casos, se trata de observaciones horarias.

3.3. Estimación de la función de densidad

3.3.1. Análisis de datos horarios

Uno de los objetivos del estudio de la temperatura en el monte Alvear era analizar la movilidad de la capa superficial de la tierra a través de los cambios de ciclo que se producen en profundidad. Con ese propósito, estudiamos el número de cambios de ciclo que tienen lugar a lo largo del día durante los años 2008-2009 y calculamos las horas a las que producen dichos cambios. De esta forma, para cada ubicación y para cada nivel, tenemos una muestra de datos horarios a los que podemos aplicarles la técnicas de estimación no paramétrica de la densidad circular estudiada en el Capítulo 2.

En las Figuras 3.1-3.4 se puede ver como las temperaturas son más estables a medida que aumenta la profundidad y en consecuencia el número de ciclos que se producen en profundidad es mucho menor que en superficie (ver Tablas 3.1 y 3.2). Por lo tanto, sólo trataremos la estimación de la densidad de las horas a las que se producen los cambios de ciclo a nivel de la superficie, que es cuando tenemos una muestra mayor de cambios de ciclo.

Además distinguiremos los cambios de ciclo que se producen por congelación de los que se producen por descongelación.

	CALM 1			
	0 cm	-10 cm	-20 cm	-60 cm
Congelación	97 ciclos	13 ciclos	3 ciclos	5 ciclos
Descongelación	97 ciclos	13 ciclos	3 ciclos	5 ciclos

Tabla 3.1: Número de ciclos por congelación y descongelación según el nivel en el que fue tomada la temperatura en la localización Calm 1.

	CALM 2		CALM 3		CALM 4	
	0 cm	-50 cm	0 cm	-50 cm	0 cm	-50 cm
Congelación	212 ciclos	8 ciclos	175 ciclos	15 ciclos	166 ciclos	2 ciclos
Descongelación	212 ciclos	8 ciclos	175 ciclos	15 ciclos	166 ciclos	2 ciclos

Tabla 3.2: Número de ciclos por congelación y descongelación según el nivel en el que fue tomada la temperatura para las localizaciones Calm 2, Calm 3 y Calm 4.

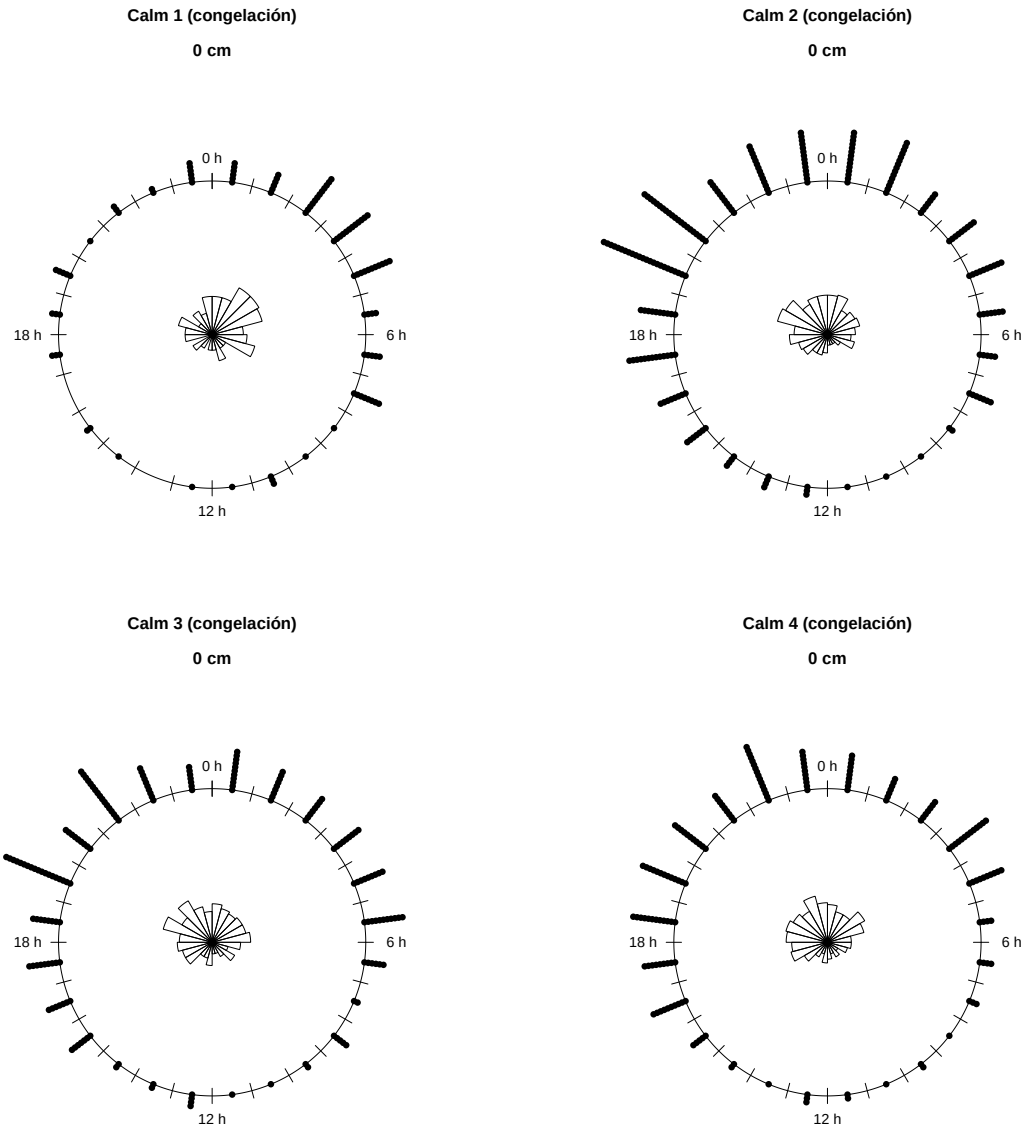


Figura 3.5: Representación de las horas en las que se producen los cambios de ciclo por congelación para las temperaturas de Calm 1, 2, 3 y 4 durante los años 2008 y 2009.

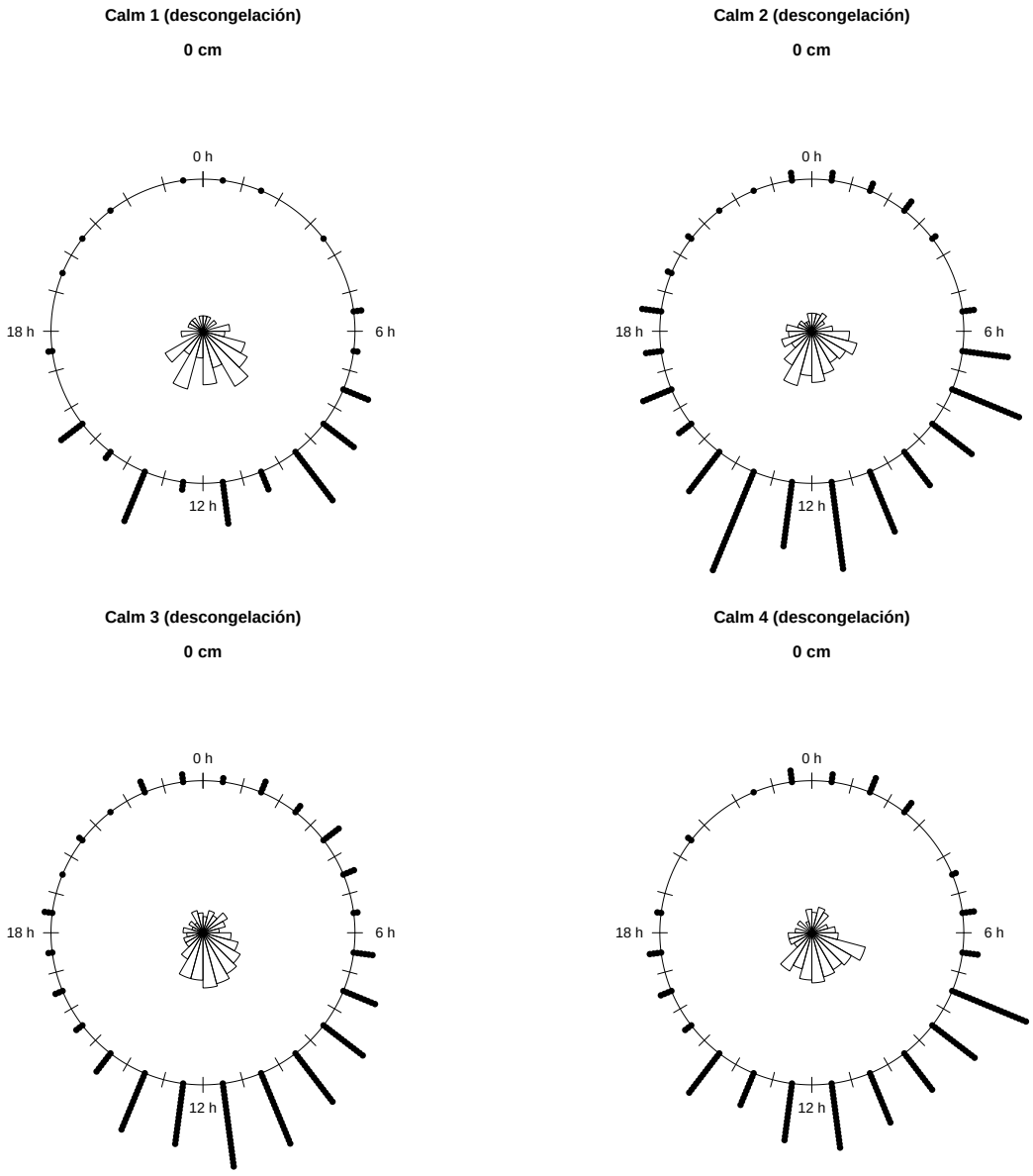


Figura 3.6: Representación de las horas en las que se producen los cambios de ciclo por descongelación para las temperaturas de Calm 1, 2, 3 y 4 durante los años 2008 y 2009.

En las Figuras 3.5 y 3.6 podemos ver la representación de las horas a las que se producen los cambios de ciclo por congelación y descongelación para cada ubicación. La circunferencia se divide en 24 grupos iguales, representando las 24 horas del día, y las horas a las que se producen los cambio de ciclo se representan por puntos apilados dentro del intervalo horario correspondiente. En el mismo gráfico está representado un diagrama de rosa de los datos, en el cual cada grupo se muestra como un sector. El radio de cada sector es proporcional

a la raíz cuadrada de las frecuencias relativas de las observaciones en cada grupo lo que asegura que el área del sector es proporcional a la frecuencia del grupo.

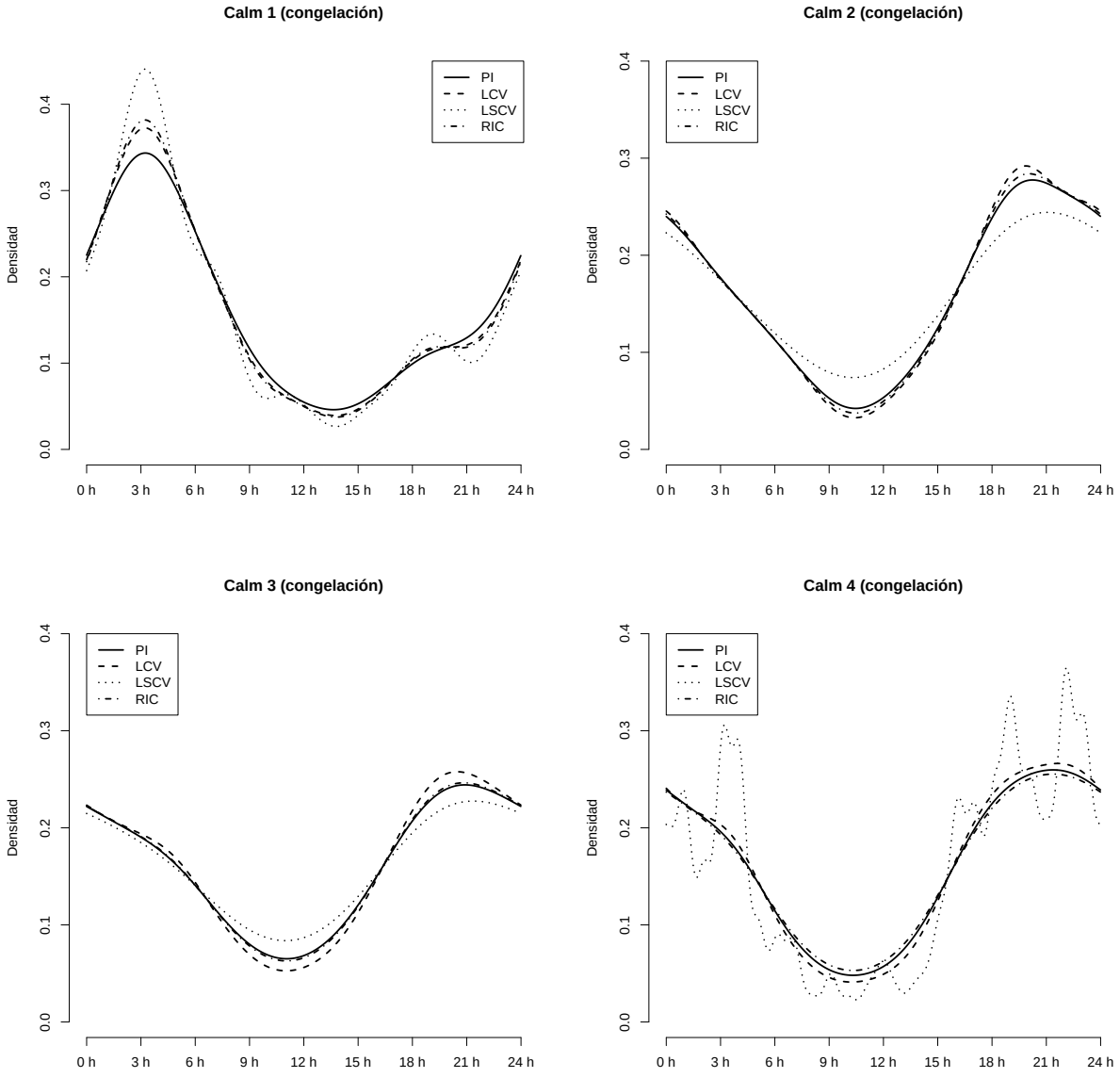


Figura 3.7: Estimación tipo núcleo de la densidad circular de las horas a las que se producen los cambios de ciclo por congelación en cada Calm durante los años 2008 y 2009 utilizando los parámetros de suavizado obtenidos mediante la regla plug-in, $LSCV$, LCV y el método basado en el rango intercuartílico (RIC), considerando núcleo von Mises.

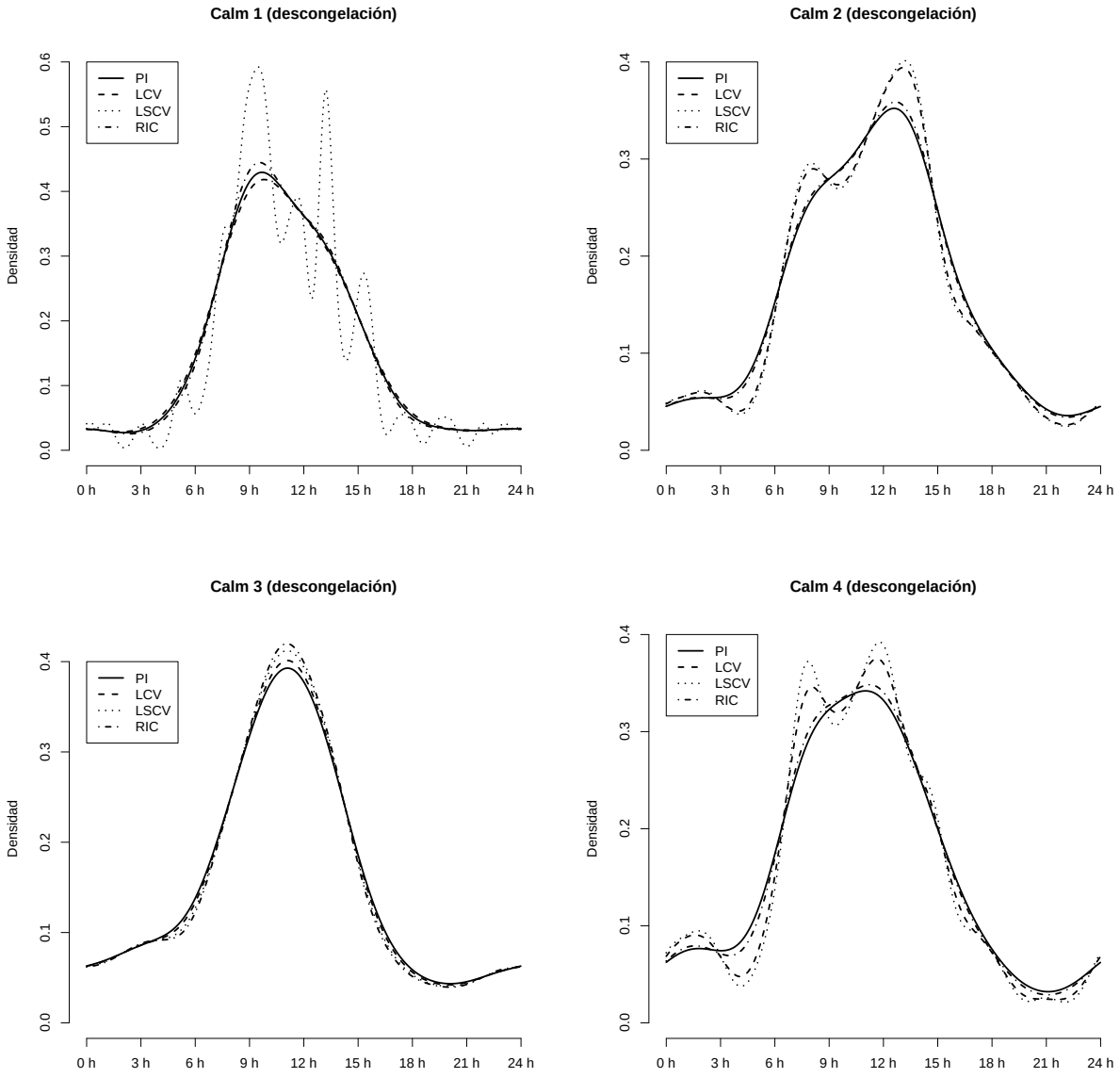


Figura 3.8: Estimación tipo núcleo de la densidad circular de las horas a las que se producen los cambios de ciclo por descongelación en cada Calm durante los años 2008 y 2009 utilizando los parámetros de suavizado obtenidos mediante la regla plug-in, $LSCV$, LCV y el método basado en el rango intercuartílico (RIC), considerando núcleo von Mises.

A partir de las representaciones de las Figuras 3.5 y 3.6 parece que los datos tienen una distribución unimodal, produciéndose los cambios de ciclo por congelación mayoritariamente de noche y de madrugada y los cambios de ciclo por descongelación en torno al mediodía, con lo cual el método plug-in para la selección de la ventana (basado en la distribución

von Mises) debería aportar buenas estimaciones de la función de densidad. Aún así, también veremos qué estimaciones se obtienen cuando el parámetro de suavizado se selecciona utilizando los métodos de validación cruzada introducidos en (2.18) y (2.19) y el método basado en la estimación del rango intercuartílico.

Como vemos en la Figuras 3.7 y 3.8, los cuatro métodos de selección de la ventana propuestos en el capítulo anterior proporcionan una estimación similar, excepto la regla de validación cruzada por mínimos cuadrados que en dos casos da lugar a una estimación infrasuavizada de la función de densidad.

En el estudio de simulación realizado en el Capítulo 2 vimos que, utilizando el estimador lineal de la densidad y seleccionando el parámetro de suavizado por el método propuesto por Seather y Jones se obtenían buenos resultados cuando los datos no estaban muy concentrados en los extremos del intervalo $[0, 2\pi)$. Esa situación se dá en los cambios de ciclo por descongelación. En dichos casos el método de selección lineal del parámetro de suavizado proporciona buenos resultados.

3.3.2. Análisis de datos de la dirección del viento

La dirección del viento es una variable circular, por lo tanto aplicamos también la estimación de la densidad circular a los datos de la dirección del viento recogidos por la estación meterológica Vinciguerra (ver Figura 3.9). En este caso se muestra únicamente la estimación que se obtiene utilizando la ventana plug-in ya que con los otros métodos se obtiene prácticamente la misma estimación, al tratarse de datos unimodales.

De la Figura 3.9 se concluye que, en esta ubicación, predomina el viento que sopla del sureste.

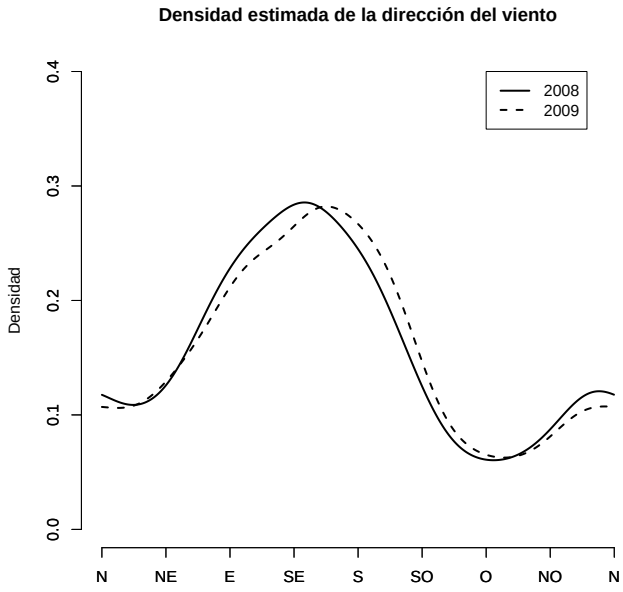


Figura 3.9: Estimación tipo núcleo de la densidad para los datos de la dirección del viento para los años 2008 (línea continua) y 2009 (línea discontinua), utilizando la ventana plug-in.

3.4. Estimación de la función de regresión

En esta sección, aplicaremos la estimación no paramétrica de la regresión vista en el Capítulo 2 para estudiar la relación entre la temperatura del aire (lineal) y la dirección del viento (circular).

La Figura 3.10 muestra el diagrama de dispersión y la correspondiente estimación local lineal de la función de regresión entre la dirección del viento y la temperatura del aire para los años 2008 y 2009 con selección del parámetro de suavizado por validación cruzada. Dichas figuras muestran una estimación claramente infrasuavizada, posiblemente debida a la dependencia existente entre las observaciones. No obstante, no conocemos ninguna técnica que permita el cálculo del parámetro de suavizado con datos dependientes en el contexto de la regresión no paramétrica circular-lineal.

Las representaciones de las funciones de regresión estimadas, tanto para el año 2008 como para el año 2009, parecen indicar que no hay relación entre la temperatura y la dirección del viento. Sin embargo, las representaciones lineales dejan ver como las temperaturas bajan con los vientos del suroeste.

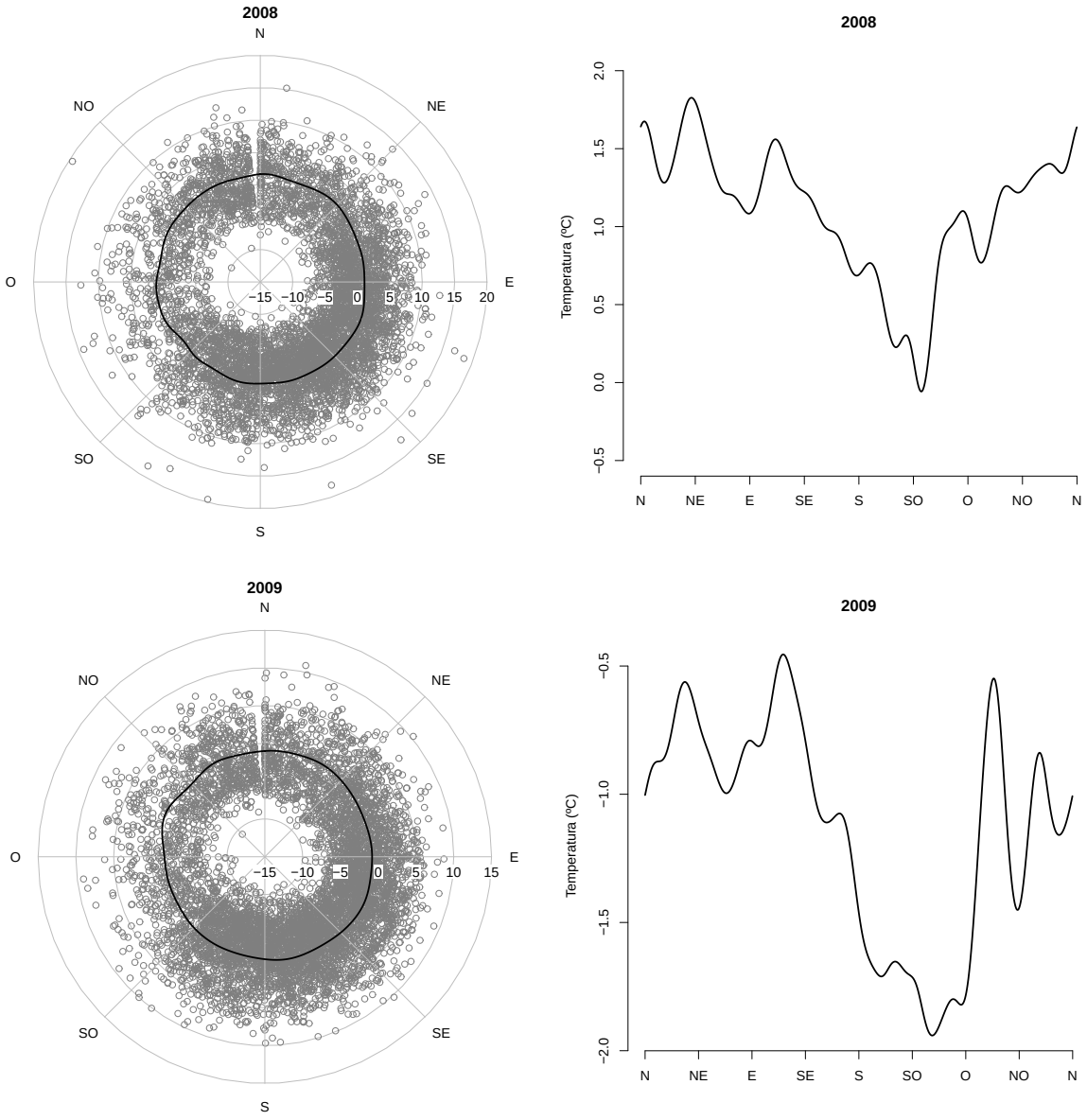


Figura 3.10: Representación circular (izquierda) y representación lineal (derecha) de la regresión tipo núcleo de la temperatura frente a la dirección del viento para los años 2008 y 2009 utilizando el estimador local-lineal circular.

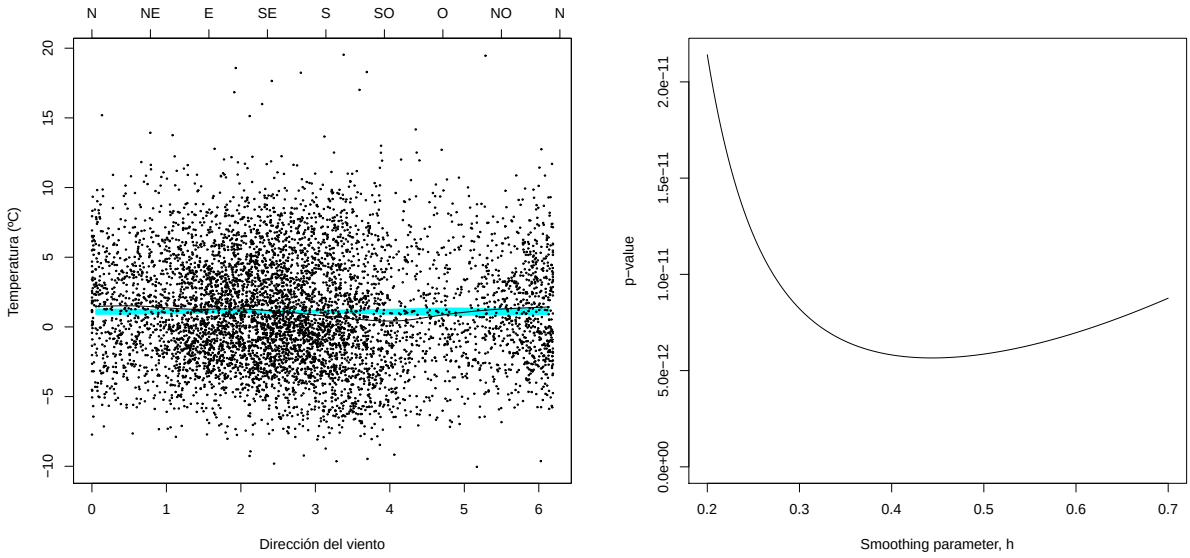


Figura 3.11: Banda de referencia al 95 % (izquierda) y p-valores para un rango de ventanas lineales (derecha) para el contraste de no efecto propuesto por Bowman y Azzalini (1997).

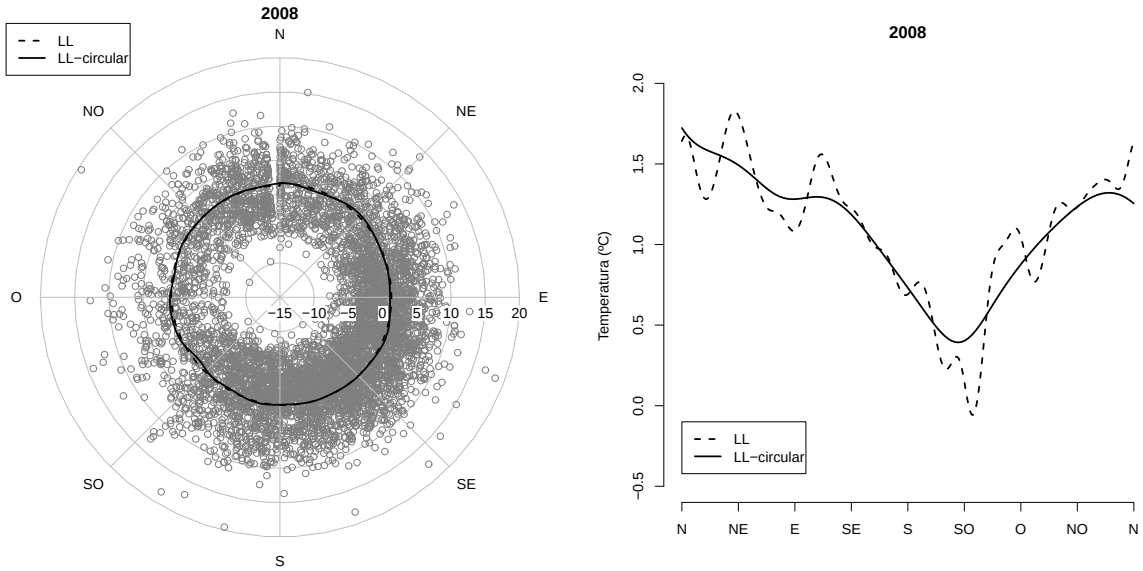


Figura 3.12: Representación circular (izquierda) y representación lineal (derecha) de la regresión tipo núcleo de la temperatura frente a la dirección del viento para el año 2008 utilizando el estimador local lineal (línea discontinua) y local lineal circular (línea continua).

Para comprobar si realmente la dirección del viento tiene un efecto significativo sobre la temperatura del aire, realizamos el test propuesto en Bowman y Azzalini (1997) que contrasta si una covariable tiene algún efecto sobre el valor medio de una variable respuesta. Este contraste está adaptado para variables periódicas. En la Figura 3.11 (izquierda) está representada la banda de referencia calculada bajo la hipótesis nula de no efecto (al 95 %) y en la representación de la derecha, los p-valores del contraste como una función del parámetro de suavizado para los datos del 2008. Dichas representaciones muestran la existencia de un efecto significativo de la dirección del viento en la temperatura del aire. Resultados similares se obtienen para el año 2009.

Cabe destacar que aunque el test de no efecto utiliza el estimador local lineal de datos lineales los resultados que proporciona son válidos dado que, como hemos visto en el capítulo anterior y como se ve en la Figura 3.12, el estimador local lineal para datos lineales y para datos circulares proporcionan resultados parecidos, siendo más suave la curva obtenida con el estimador para datos lineales.

Consideremos ahora la regresión por estaciones. Por simplicidad, consideraremos verano los meses de Enero, Febrero y Marzo, otoño los meses de Abril, Mayo y Junio, invierno los meses de Julio, Agosto y Septiembre, y primavera los meses de Octubre, Noviembre y Diciembre. En la Figura 3.13 se muestran los resultados para el año 2008.

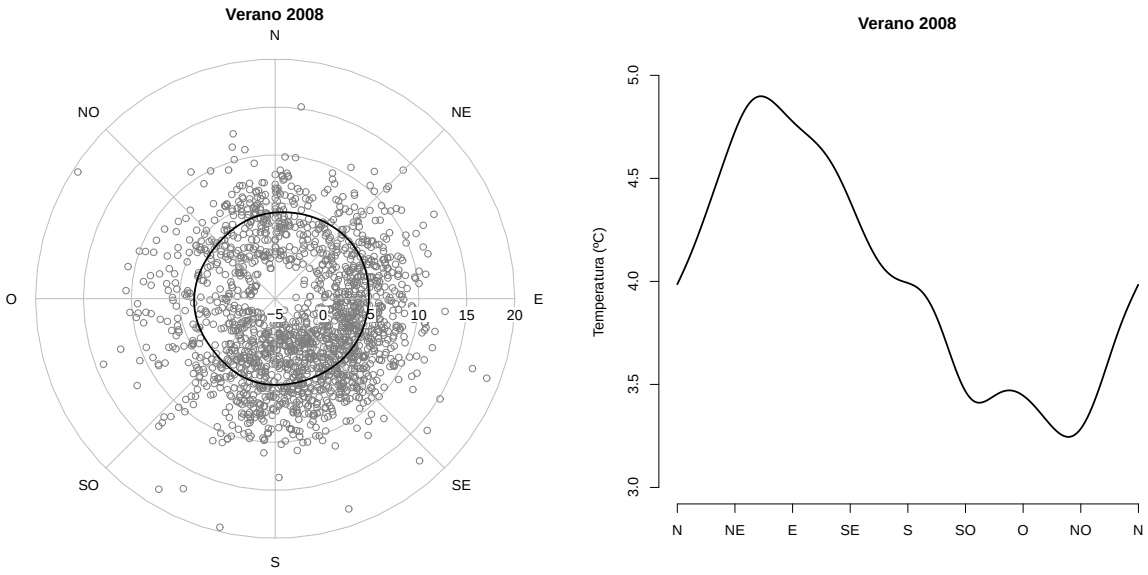
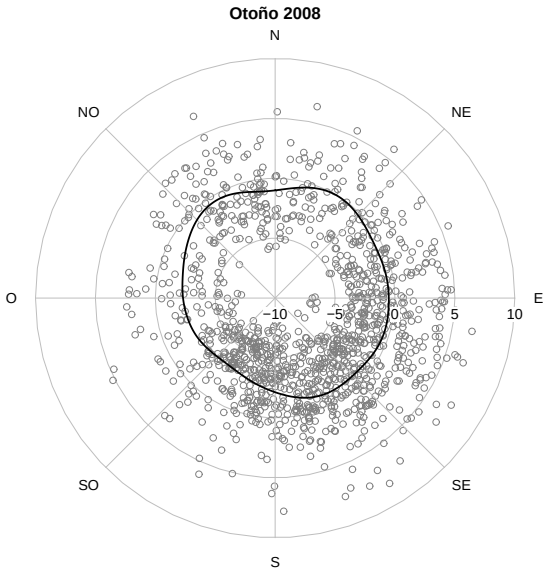
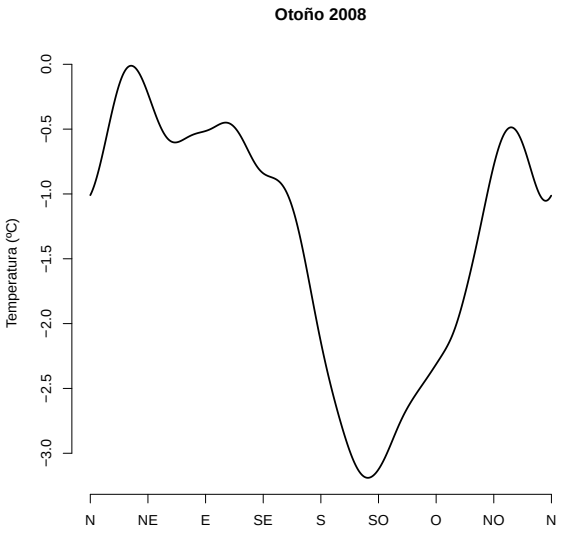
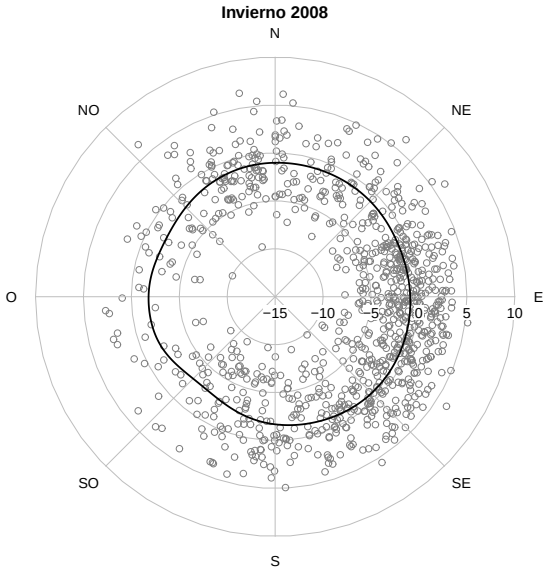
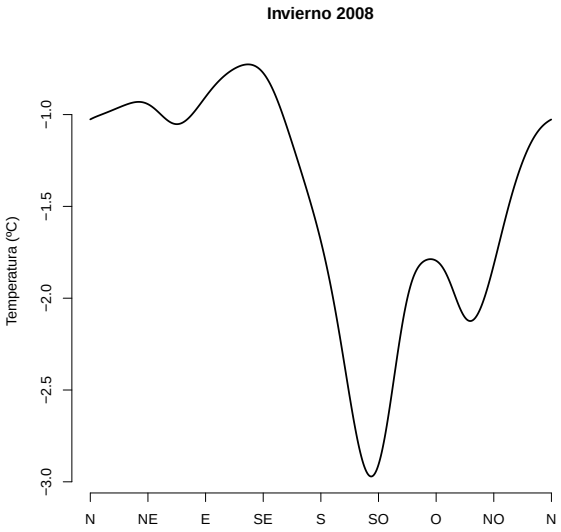


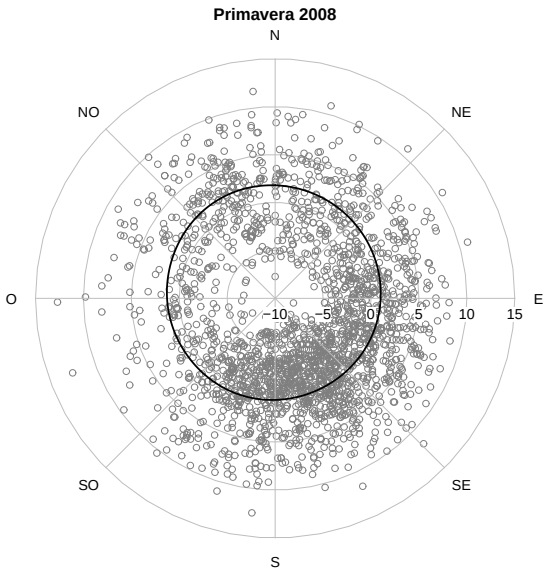
Figura 3.13: Regresión local lineal de la temperatura frente a la dirección del viento por estaciones para el año 2008 (continúa en la página siguiente).

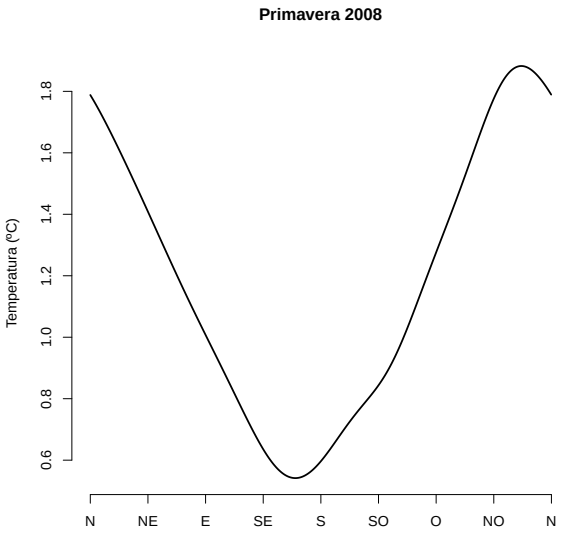

Otoño 2008
N
NE
E
SE
S
SO
O
NO
−10
−5
0
5
10


Otoño 2008
Temperatura (ºC)
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
N
NE
E
SE
S
SO
O
NO
N


Invierno 2008
N
NE
E
SE
S
SO
O
NO
−15
−10
−5
0
5
10


Invierno 2008
Temperatura (ºC)
−1.0
−1.5
−2.0
−2.5
−3.0
N
NE
E
SE
S
SO
O
NO
N


Primavera 2008
N
NE
E
SE
S
SO
O
NO
−10
−5
0
5
10
15


Primavera 2008
Temperatura (ºC)
1.8
1.6
1.4
1.2
1.0
0.8
0.6
N
NE
E
SE
S
SO
O
NO
N

Al igual que pasaba tomando el año completo, el test de no-efecto confirma la existencia de una relación entre la dirección del viento y la temperatura del aire para las cuatro estaciones. Salvo en primavera, cuando las temperaturas son más altas cuando el viento sopla del noroeste (NO), las temperaturas aumentan con vientos de componente este.

Capítulo 4

Análisis de datos de dureza

4.1. Introducción

Dentro del proyecto de investigación "*Definición, cartografía y caracterización de las grandes áreas paisajísticas de Galicia*", se enmarca un estudio sobre terrazas fluviales en la zona del bajo Miño, entre Orense y Ribadavia. Dicho estudio describe una nueva metodología para distinguir niveles de terrazas fluviales basándose en el grado de dureza de cantos de cuarcita utilizando un dispositivo denominado Equotip (`http://www.equotip.com`).

El grupo de investigación parte de la hipótesis de que los cantos con menor dureza son los que presentan un desgaste mayor y por lo tanto son más antiguos que los que tienen un valor de dureza más alto. Los estudios realizados tienen por objetivo contrastar dicha hipótesis.

4.2. Datos de dureza

Se dispone de mediciones tomadas en cantos de cuarcita. También se disponía de cantos de cuarzo pero no se midieron porque superan el máximo valor de dureza recomendado por el fabricante y porque producen uniformemente valores altos de rebote.

La dureza se midió con el durómetro Equotip que se muestra en la Figura 4.1. El Equotip es un instrumento diseñado originalmente para medir la dureza en metales, pero en los últimos años se ha utilizado para medir la dureza de la roca. Está compuesto por un dispositivo de impacto, concretamente una bola de carburo de tungsteno con un diámetro de 3 mm, y un dispositivo electrónico que mide la velocidad con la que la bola rebota sobre la superficie

del cuerpo impactado. La bola, lanzada contra la superficie del material a medir, al rebotar, genera una corriente en una pequeña bobina siendo la medida de su voltaje la que indica la velocidad de rebote. El valor de dureza se expresa como el número de Leeb (el valor L) o la dureza Leeb (HL del inglés *Hardness Leeb*) y es el cociente entre la velocidad de rebote y la velocidad de impacto multiplicado por 1000 (Kompatscher, 2004),

$$\mathrm{L} = \frac{\text{Velocidad de rebote}}{\text{Velocidad de impacto}} \times 1000.$$

Es un coeficiente adimensional y toma valores entre 0 y 1000. Cuanto mayor sea la dureza del material medido, mayor será la velocidad de rebote y por tanto mayor será el valor L. De esta forma, el valor L del Equotip representa un valor de la dureza y como tal puede ser utilizado para comparar diversos materiales sometidos al mismo test.

Además, el Equotip dispone de distintos dispositivos de impacto en función de la superficie sobre la que se vaya a trabajar. Los datos con los que se trabaja aquí se han obtenido utilizando el dispositivo de impacto D. En este caso la unidad de dureza es HLD.

Figura 4.1: Dispositivo de medición de dureza de la cuarcita: Equotip.

Los datos constan de 25 medidas de dureza de cantos fluviales en cada uno de los 24 puntos de muestreo en las zonas de Puga, Laias, Troncoso, Santa Cruz, A Groba y Prado. (Ver Figura 4.2)

La Figura 4.3 muestra las medidas de dureza en los distintos puntos de muestreo.

4.3. Análisis estadístico

Realizaremos un análisis de dichos datos. El objetivo de este análisis es doble: primero, describir las medidas de dureza obtenidas en cada punto y segundo, clasificar los puntos

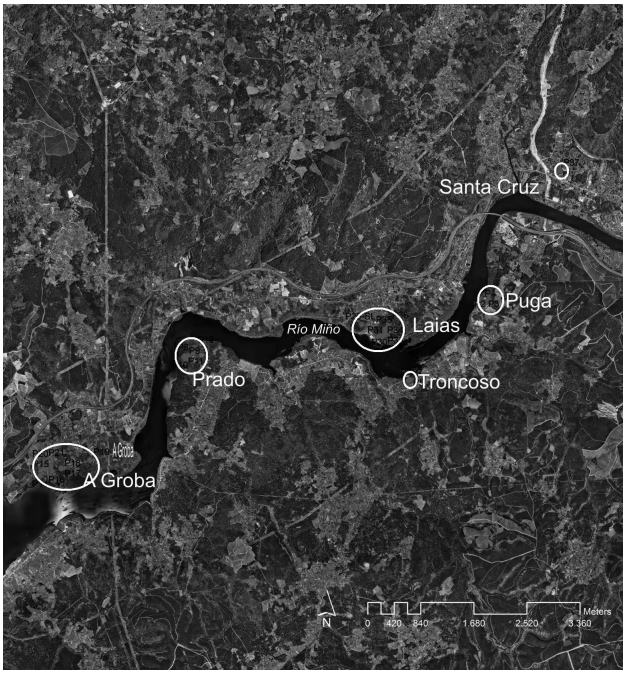


Figura 4.2: Mapa con las localizaciones en donde se toman los datos.

donde se han tomado las medidas en grupos con valores de dureza similares.

Como medidas estadísticas descriptivas, se calculan medidas de localización de tendencia central como la media y la mediana. Grandes diferencias entre la media y la mediana pueden indicar asimetría en la distribución de los datos, o la presencia de datos atípicos (outliers). Para la descripción de la tendencia no central, se calcula el primer y el tercer cuartil. Como medida de dispersión se proporciona la desviación estándar y el rango de variación de los datos (valores mínimo y máximo).

Para el segundo de los objetivos de este análisis haremos uso de las técnicas de formación de grupos. Estos métodos se clasifican en métodos jerárquicos y métodos de particionamiento. Los primeros parten de una matriz de distancia entre individuos, y en base a ella pretenden un agrupamiento de los individuos a distintos niveles. En el nivel más bajo cada grupo estaría formado por individuos, mientras que los grupos a niveles superiores serían el resultado de agregar grupos de niveles inferiores. Los métodos de particionamiento, en cambio, pretenden formar una partición de los individuos de la muestra en k grupos, en base a los valores de las variables observadas en cada individuo. Por supuesto, los grupos se formarán por proximidad en el espacio d-dimensional, siendo d el número de variables.

De acuerdo a esta clasificación y al objetivo de nuestro análisis lo adecuado es utilizar un método de particionamiento. Concretamente, usaremos el denomidado algoritmo de las k-

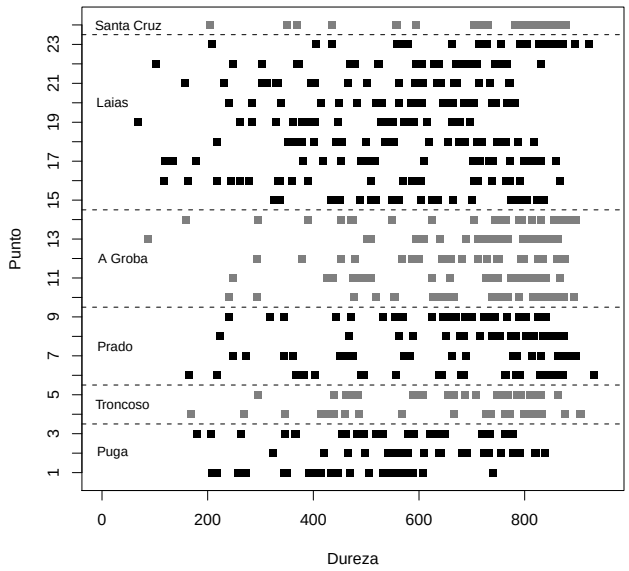


Figura 4.3: Gráfico de puntos para las medidas de dureza en cada punto de muestreo, clasificados por lugares: Puga, Troncoso, Prado, A Groba, Laias y Santa Cruz.

medias (ver Everitt, 2005). Este algoritmo proporciona una forma sencilla de clasificar un conjunto de datos dado en un cierto número de clústers (k fijo), definiendo el centroide de cada grupo tan lejos como sea posible de los demás. Cada observación se asigna al clúster cuyo valor medio sea más próximo, de una forma iterativa, que puede describirse como sigue:

- **Paso 1.** Crear una partición inicial en k grupos.
- **Paso 2.** Asignar cada observación al grupo cuyo valor medio quede más próximo y que será el centroide del grupo.
- **Paso 3.** Recalcular los centroides, esto es, la media del grupo, en base a los nuevos datos.
- **Paso 4.** Repetir los pasos 2 y 3 hasta que no haya más reasignaciones.

En la Tabla 4.1, se muestran los estadísticos resumen, medidas de tendencia y dispersión, para cada uno de los puntos. Los gráficos de caja para las medidas de dureza en cada punto se muestran en la Figura 4.4. En ambas representaciones (Figura 4.3 y Figura 4.4) se identifican las localizaciones, en concreto, Puga, Troncoso, Prado, A Groba, Laias y Santa

Cruz. A partir de la Tabla 4.1 y de la Figura 4.4 se puede ver que existen diferencias en las medidas de dureza en cada punto de muestreo, aún dentro de la misma localización.

En la Figura 4.4 se observa la existencia de datos atípicos. Cabe notar en este momento que, para el análisis estadístico que se ha realizado se ha trabajado teniendo en cuenta estos datos ya que el grupo que ha tomado los mismos está seguro de que el material medido es cuarcita y que por tanto los datos que aparecen como atípicos no se deben a que se haya medido la dureza de un canto que no fuese cuarcita.

		Mín.	1er cuartil	Mediana	Media	3er cuartil	Máx.	Desv. típica
Puga	Punto 1	208	391	445	445.8	540	740	128.3
	Punto 2	324	556	642	647.7	778	838	140.2
	Punto 3	180	454	529	524.2	645	778	169.3
Troncoso	Punto 4	169	460	763	656.3	815	906	215.4
	Punto 5	296	596	708	670.8	785	863	148.1
Prado	Punto 6	164	495	684	649.8	829	932	213.2
	Punto 7	248	475	662	652.8	864	897	212.5
	Punto 8	223	685	774	729.6	818	874	146.2
	Punto 9	240	554	670	635.0	743	840	164.5
A Groba	Punto 10	240	626	740	691.7	820	893	172.8
	Punto 11	248	496	750	669.7	804	867	175.3
	Punto 12	293	600	661	668.8	793	877	150.3
	Punto 13	88	689	745	709.9	811	864	163.7
	Punto 14	159	549	790	693.6	856	898	207.0
Laias	Punto 15	325	488	625	608.8	774	836	159.5
	Punto 16	118	334	510	508.6	712	866	224.3
	Punto 17	120	490	732	616.8	798	859	227.4
	Punto 18	217	442	549	565.4	709	817	164.8
	Punto 19	68	378	528	472.6	587	697	158.6
	Punto 20	241	518	606	585.7	693	781	148.2
	Punto 21	158	396	593	523.1	667	772	177.6
	Punto 22	103	478	637	577.8	695	830	180.9
	Punto 23	209	662	805	734.1	855	922	180.0
Santa Cruz	Punto 24	204	595	791	706.2	833	879	189.8

Tabla 4.1: Estadísticos descriptivos para los datos de dureza en cada punto.

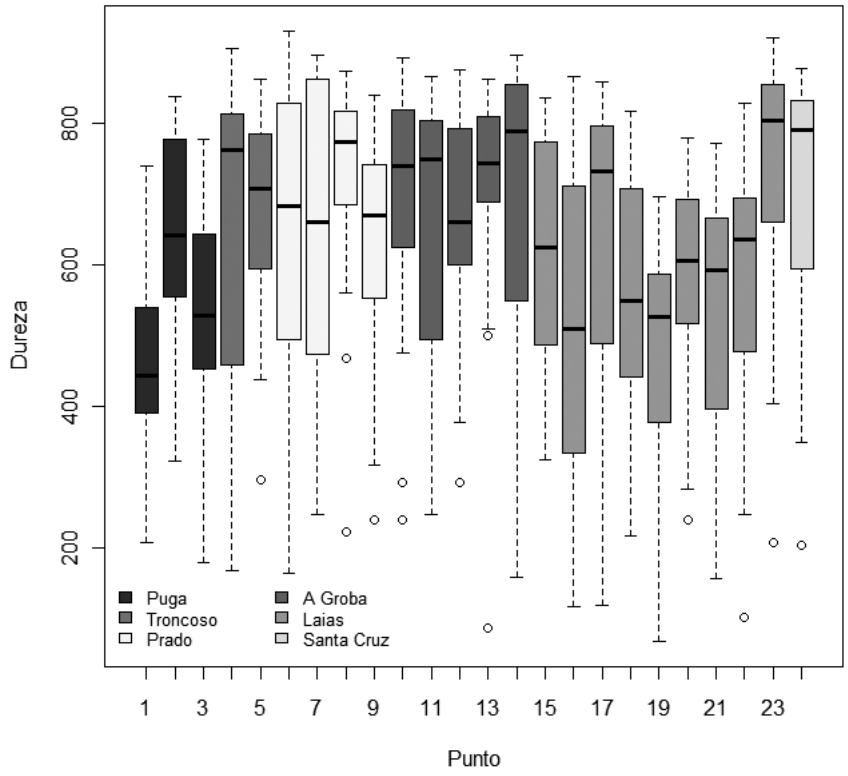


Figura 4.4: Diagramas de caja para las medidas de dureza en cada punto de muestreo, clasificados por lugares: Puga, Troncoso, Prado, A Groba, Laias y Santa Cruz.

Para aplicar el algoritmo de las k-medias para clasificar los puntos en grupos, la información de cada punto de muestreo se resume por su dureza media. Por tanto, se calculará la media de las 25 mediciones de cada punto. Se construirán grupos o clústers de forma que la suma residual de cuadrados dentro de cada grupo sea lo más pequeña posible. La Figura 4.5 muestra la evolución de la suma residual de cuadrados en función del número de grupos. Así, cuatro grupos parece ser un número adecuado.

Con el algoritmo de las k-medias, los puntos 1, 19, 16, 21 y 3 pertenecen al grupo con la dureza más baja y por tanto, con más antigüedad. El siguiente grupo está formado por los puntos 18, 22, 20, 15 y 17, con dureza media-baja. Los puntos 9, 2, 6, 7, 4, 12, 11 y 5 pertenecen al grupo de dureza media-alta. Finalmente, los puntos con la dureza más alta son 10, 14, 24, 13, 8 y 23. Estos resultados se muestran en la Figura 4.6.

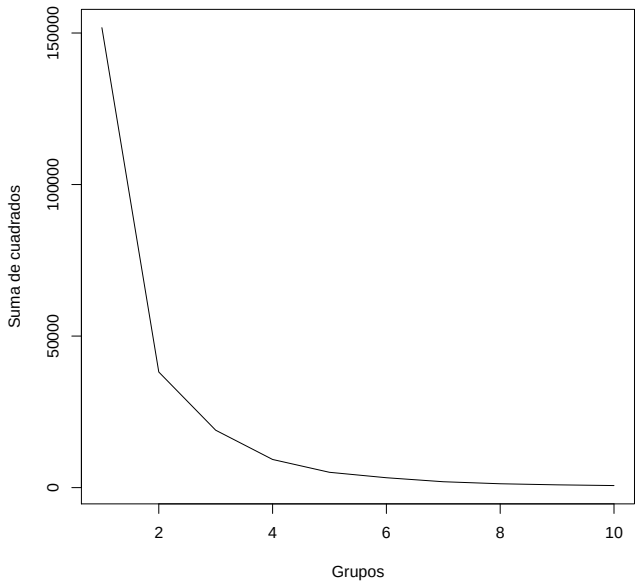


Figura 4.5: Evolución de la suma residual de cuadrados en función del número de grupos.

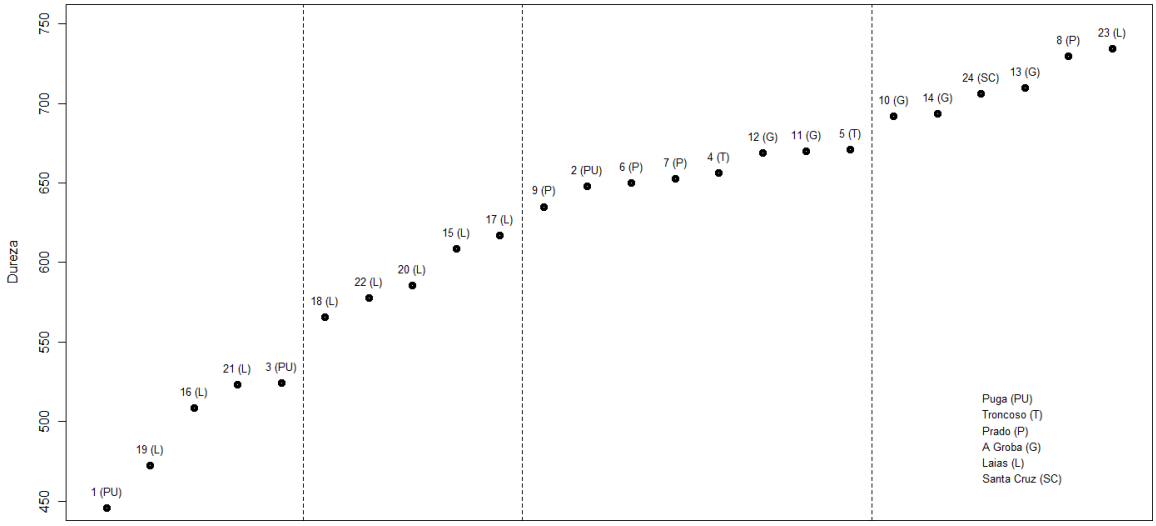


Figura 4.6: Dureza media en cada punto, clasificados por lugares: Puga, Troncoso, Prado, A Groba, Laias y Santa Cruz. Las líneas verticales separan los grupos.

Software

Para la realización de este trabajo se ha utilizado el software `R` (R Development Core Team, 2011).

Se ha hecho uso de la librería `circular`, dedicada a datos circulares. Esta librería nos permitió generar datos de distribuciones von Mises, obtener sus funciones de densidad, hacer representaciones circulares (como los diagramas de rosa), etc. Para las representaciones como la de la Figura 3.10 (izquierda) se ha utilizado la librería `plotrix`. En la librería `sm` se encuentra la función que permite realizar el test de no-efecto de una variable sobre otra que se ha aplicado en el análisis de datos.

Tanto los métodos de estimación no paramétrica de la función de densidad como de la función de regresión, así como los respectivos métodos de selección del parámetro de suavizado han formado parte del trabajo desarrollado en esta memoria. Así, se han implementado concretamente funciones para calcular:

- El estimador no paramétrico de la función de densidad para datos lineales introducido en (2.9).
- El estimador no paramétrico de la función de densidad para datos circulares introducido en (2.10).
- La ventanas plug-in para datos lineales introducida en la sección 2.3.1.
- La ventana de Seather y Jones a la que se hace referencia en la sección 2.3.2.
- Los parámetros de suavizado con los métodos de validación cruzada ($LSCV$ y LCV), plug-in y RIC para datos circulares de la sección 2.3.1.
- Los estimadores local lineal y Nadaraya-Watson de la función de regresión para variable respuesta y variable explicativa lineal introducidos en la sección 2.4.1.

- El estimador local lineal de la función de regresión para variable respuesta lineal y variable explicativa circular introducido en (2.23).
- El estimador de Nadaraya-Watson de la función de regresión para variable respuesta lineal y variable explicativa circular introducido en (2.24).
- La función de validación cruzada para la regresión introducida en (2.25).

Bibliografía

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

Bowman, A. W. y Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*. Oxford Science Publications.

Di Marzio, M., Panzera A. y Taylor, C. C. (2009). Local polynomial regression for circular predictors. *Statistics & Probability Letters*, **79**, 2066-2075.

Habbema, J.D.F., Hermans, J. y van der Broek, K. (1974). A stepwise discrimination analysis program using density estimation. *In Compstat 1974: Proceedings in computational statiscs (G. Bruckman, ed.)* 101-110. Physica Verlag, Vienna.

Jammalamadaka, S. R. y SenGupta, A. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.

Kompatscher, M. (2004). Equotip - Rebound Hardnes Testing After D. Leeb. En *Hardmeko 2004. Proceeding of Hardness Measurements Theory and Application in Laboratories and Industries*.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, **9**, 65-78.

Sheather, S. J. y Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society, Series B*, **53**, 683-690.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Taylor, C. C. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics and Data Analysis*, **52**, 3493-3500.

Pérez, A., Gomes, A., García, H. y Oliveira, M., enviado (2011). Distinguish fluvial terraces by clast weathering with an Equotip Hardness Tester: An example from Miño River in the Northwestern Iberian Peninsula.