Trabajo Fin de Máster

Análisis estadístico de la concentración de distintos metales en muestras de Pseudoscleropodium purum

Pablo Giráldez Suárez

Máster en Técnicas Estadísticas

Curso 2018-2019

Propuesta de Trabajo Fin de Máster

Título en galego: Análise estatístico da concentración de distintos metáis pesados en mostras de Pseudoscleropodium purum.
Título en español: Análisis estadístico de la concentración de distintos metales en muestras de Pseudoscleropodium purum.
English title: Statistical analysis of the concentration of different metals in Pseudoscleropodium purum samples.
Modalidad: Modalidad A
Autor/a: Pablo Giráldez Suárez, Universidad de Santiago de Compostela
Director/a: Rosa María Crujeiras Casais, Universidad de Santiago de Compostela
Breve resumen del trabajo: Análisis estadístico de las muestras de musgo recogidas en 150 puntos de Galicia muestreados en 2000, 2002, 2004, 2006, 2008 y 2014 (en cada año en primavera y otoño). Con el fin de determinar los cambios espacio-temporales que se puedan detectar en las concentraciones de metales pesados del musgo. Para ello se planteará un trabajo de análisis inicialmente descriptivo. Y se determinará la posibilidad de realizar inferencia en base a los datos de la muestra. En lo que respecta a la búsqueda de patrones espacio-temporales, se empleará la alternativa no paramétrica y también Bayes (con modelos jerárquicos).

Doña Rosa María Crujeiras, Profesora Titular de la Universidad de Santiago de Compostela, informa que el Trabajo Fin de Máster titulado

Análisis estadístico de la concentración de distintos metales en muestras de Pseudoscleropodium purum

Fue realizado bajo su dirección por don Pablo Giráldez Suárez para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, dan su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 5 de septiembre de 2019.

La directora: El autor:

Doña Rosa María Crujeiras Casais Don Pablo Giráldez Suárez

Agradecimientos

En primer lugar, al grupo de investigación ECOTOX, de la Universidad de Santiago de Compostela, el haberme proporcionado los datos reales que se emplearon en este trabajo, así como haberme ayudado con todas las dudas que me surgieron a nivel “biológico”.

A mis compañeros y compañeras del Máster (Brais, Andrea, Paloma, Ramiro y Raúl) por todo el apoyo que me brindaron a lo largo de los estudios y por los buenos momentos que compartimos en estos dos últimos años.

A mi novia y mi familia por ayudarme y sobre todo soportarme en los momentos difíciles.

De forma muy especial, a Rosa Crujeiras, mi directora, por su dedicación, entusiasmo y siempre generosa y amable disposición para guiarme y revisar el trabajo a cualquier hora del día. Su apoyo y su actitud animosa fueron tan importantes que, sin ellos, dudo que hubiera podido concluirlo en condiciones.

Índice general

Resumen

Resumen en español

Las concentraciones de metales pesados en musgo se analizan de forma cuantitativa en la mayor parte de los estudios que emplean la técnica “moss bag”. A pesar de esto, la gran variabilidad de estas concentraciones junto con la falta de conocimiento teórico sobre el proceso de acumulación de contaminantes por parte del musgo y la relación real entre la deposición atmosférica y la concentración de contaminantes en los tejidos del musgo, hace que sea muy difícil interpretar las concentraciones de forma cuantitativa. Por ello, en este estudio se propone el tratamiento cualitativo de los datos mediante un nuevo protocolo que nos permita determinar si un punto geográfico está o no contaminado. En este trabajo se desarrolla una nueva regla para la asignación de las observaciones a dos categorías (“no contaminado” y “contaminado”) de la variable binomial “contaminación”, lo que permite realizar un tratamiento cualitativo de los datos. Esta regla se basa en el ajuste de un modelo de mixtura de normales a la distribución de la concentración de los metales pesados en la red de muestreo. Una vez ajustado el modelo, en caso de que una observación se asigne a la primera componente del modelo, será considerada como “no contaminada” y, en otro caso, como “contaminada”. Tras la asignación, se realiza un análisis geoestadístico de los datos y se obtienen mapas de predicciones mediante kriging indicador.

English abstract

The concentrations of heavy metals in moss are quantitatively analysed in most of the studies using the "moss bag" technique. In spite of this, the great variability of these concentrations together with the lack of theoretical knowledge about the process of accumulation of pollutants by moss and the real relationship between atmospheric deposition and the concentration of pollutants in moss tissues, makes it very difficult to interpret the concentrations in a quantitative way. In this work a new rule is developed for the assignment of the observations to two categories ("uncontaminated" and "contaminated"), which allows a qualitative treatment of the data. This rule is based on the adjustment of a normal mixture model to the distribution of the concentration of heavy metals in the sampling network. Once the model has been adjusted, if an observation is assigned to the first component of the model, it will be considered as "uncontaminated" and, in another case, as "contaminated". After assignment, a geostatistical analysis of the data is performed and prediction maps are obtained by indicator kriging.

Capítulo 1

1. Introducción

En este apartado se incluye la motivación del trabajo, un breve análisis descriptivo de los datos que se van a emplear y la descripción de la estructura general del mismo.

1.1. Motivación del trabajo

El uso de musgos terrestres para monitorizar la deposición de contaminantes atmosféricos (principalmente metales pesados) se conoce como la "moss bag technique". Esta técnica fue descrita en los años 60 y, desde entonces, se ha empleado en cientos de estudios científicos. Esta técnica, mediante el uso métodos estandarizados, permite determinar las concentraciones de metales pesados en los tejidos de los musgos. Muchos autores asumen que estas concentraciones reflejan la deposición atmosférica de estos contaminantes y, en la mayor parte de la literatura científica, se han tratado de forma cuantitativa. Sin embargo, en los últimos años, algunos autores han reivindicado y proporcionado argumentos para el uso de la técnica del musgo de forma cualitativa (Aboal *et al.* 2017; Boquete *et al.* 2011, 2017; Fernández *et al.* 2015):

1. La falta de correlación significativa entre la concentración de metales pesados en el musgo y los determinados en la deposición total (Aboal *et al.* 2010; Boquete *et al.* 2015). Se han observado correlaciones significativas en alrededor del 40% de los casos estudiados (con un coeficiente de determinación superior a 0,7 en sólo alrededor del 15% de los casos estudiados). Las características fisicoquímicas de los elementos, como el índice covalencia, podrían explicar las diferencias encontradas entre los metales pesados (Varela *et al.* 2015), y sólo para Cd y Pb se han encontrado correlaciones relevantes (Aboal *et al.* 2010; Harmens *et al.* 2010).
2. La existencia de errores inherentes a la técnica como la variabilidad a corto plazo (Aboal *et al.* 2017; Boquete *et al.* 2011). La mayoría de los estudios llegaron a la conclusión de que la representatividad temporal de la concentración de musgo es generalmente baja (Boquete *et al.* 2011, 2017; Markert y Weckert 1989; Real *et al.* 2008). Recogiendo muestras con un retraso de una semana, la concentración obtenida puede variar de 2 a 3 veces.

Pero, por otro lado, hay muchas evidencias de que en sitios con altos niveles de deposición de metales pesados los musgos alcanzaron concentraciones más altas que en áreas de fondo. Se ha descrito una disminución exponencial de la concentración de contaminantes a medida que aumenta la distancia de la fuente (Fernández *et al.* 2007). Este resultado se ha encontrado para casi todos los metales pesados y es tan robusto que la variabilidad temporal no lo enmascara (Boquete *et al.* 2011). Por esta razón, se ha descrito el uso de musgos para determinar cuándo un foco industrial a pequeña escala está contaminando o no el área circundante (Ares *et al.* 2009; Fernández *et al.* 2007). El uso de concentraciones de musgo de esta forma cualitativa (asignando una probabilidad de contaminación, sin utilizar el valor de las concentraciones) ha permitido obtener resultados de alta calidad (Varela *et al.* 2014), pero este enfoque no se ha aplicado hasta ahora a los estudios de biomonitorización de musgos nacionales o regionales.

1.2. Análisis descriptivo de los datos

Los datos de estudio fueron las concentraciones de metales pesados y de nutrientes en tejidos de musgos terrestres en Galicia, concretamente, en Pseudoscleropodium purum (Hedw.) M.Fleisch (en adelante, P. purum), que se muestra en la Figura 1. Para obtener estas observaciones se realizaron muestreos en primavera y otoño en los años 2000, 2002, 2004, 2006, 2008 y 2014 (en el año 2000 solo se hizo el muestreo de primavera), enmarcados dentro del proyecto BEAG. Los puntos de muestreo se dispusieron formando una red que cubría toda la Comunidad Autónoma de Galicia y las zonas colindantes, compuesta por 150 estaciones de muestreo localizadas de forma equidistante en los vértices de cuadrados de 15 x 15 km. En el año 2000 el musgo P. purum solo se encontró en 132 de las 150 estaciones, las que se muestran en la Figura 1 y que son las que se utilizan en este trabajo para explicar (ejemplificar) el procedimiento. En cada una de las estaciones se recogieron musgo de 30 localizaciones dentro de la estación de muestreo y estas 30 submuestras se mezclaron para formar una única muestra compuesta, que es a partir de la que se obtienen los datos empleados en este estudio.

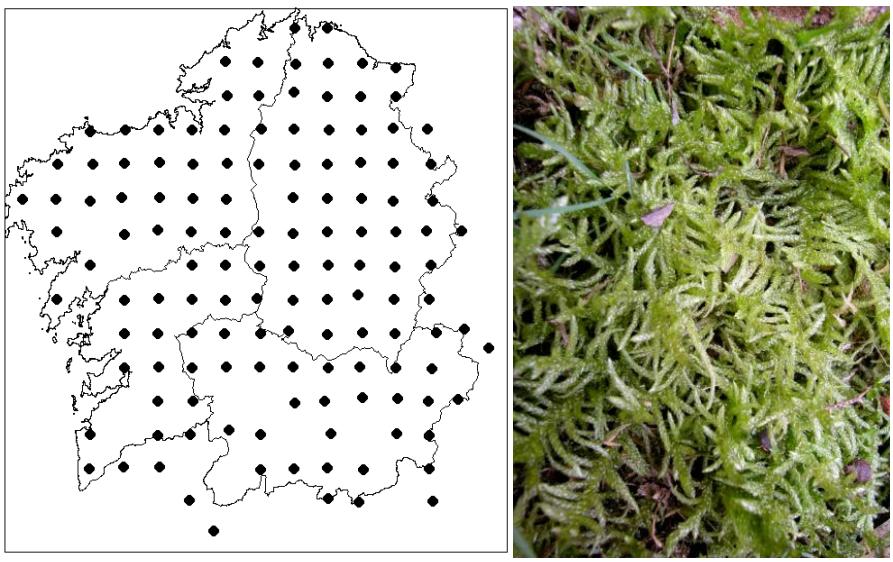

Figura 1. Izquierda: estaciones de muestreo de la Comunidad Autónoma de Galicia en las que se pudo encontrar el musgo P. purum (132 puntos de un total de 150) en el año 2000. Derecha: imagen del musgo P. purum.

De forma general, en las regiones donde hay puntos contaminados, las concentraciones de los metales en el musgo, a diferencia de las concentraciones de los nutrientes, no siguen una distribución normal. Como se puede ver en la Figura 2, esta situación se dio en las muestras de este estudio. La concentración de nutrientes, que está regulada biológicamente y, por lo tanto, no es tan sensible a la contaminación, presenta distribuciones normales que en ocasiones se ven ligeramente distorsionadas en puntos de gran contaminación (Figura 3). Por otra parte, en las gráficas de la estimación no paramétrica de la densidad de la concentración de los metales pesados (Figura 2), se ve lo que parece ser un primer grupo de observaciones que contienen al grueso de la muestra y cuya densidad tiene un comportamiento aproximadamente normal y otros grupos de observaciones cuyas concentraciones son mayores y que identificamos como puntos posiblemente contaminados. Estas estimaciones se han obtenido utilizando un estimador tipo núcleo con núcleo Gaussiano y ventana seleccionada mediante regla del pulgar. Por último, en estas gráficas se puede ver que hay una pequeña cantidad de observaciones cuyas concentraciones en metales pesados son muy elevadas con respecto al resto de las observaciones. Estas observaciones se consideran puntos claramente contaminados.

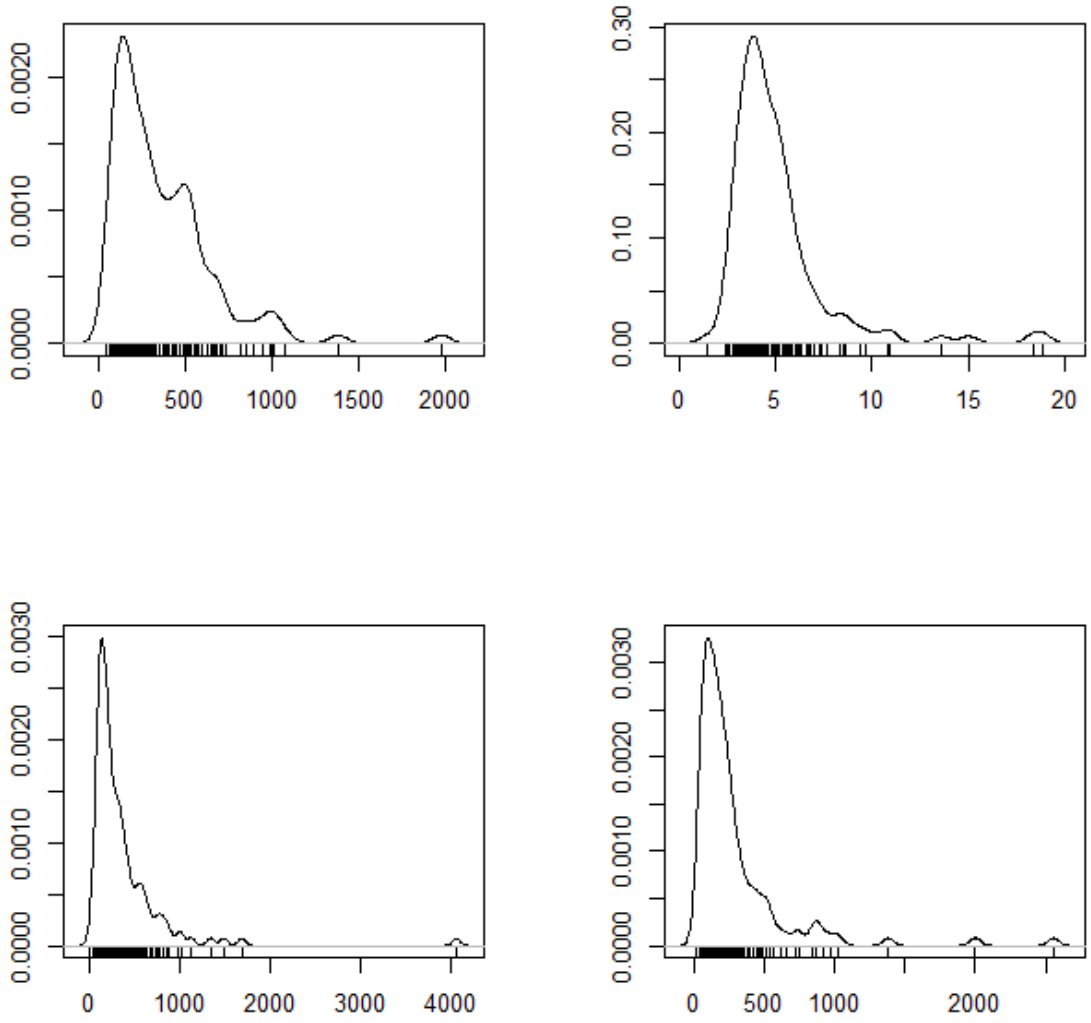

Figura 2. Concentraciones de los metales (de derecha a izquierda y de arriba abajo: mercurio en ng/g, cobre en µg/g, aluminio en µg/g y arsénico en ng/g) en las muestras del musgo P.purum recogidas en 132 estaciones de muestreo de la Comunidad Autónoma de Galicia en el año 2000.

Como ya se comentó en la Sección 1.1, las concentraciones de metales en los musgos presentan una gran variabilidad espacial y temporal, por lo que no se recomienda emplearlas de forma cuantitativa para la diagnosis de la contaminación local. Es decir, si se considera que una localización está contaminada simplemente al tener en cuenta el valor absoluto de la concentración del metal medida en los tejidos del musgo, se pueden cometer errores de juicio a la hora de determinar si un punto está o no contaminado, sobre todo cuando la contaminación en este punto no es muy elevada. Para solventar esta problemática y teniendo en cuenta los grupos de observaciones que se detectan en las concentraciones de metales pesados en el musgo, en este trabajo se planteó un procedimiento para el análisis de cualitativo de la

contaminación a partir de la distribución de las concentraciones y la agrupación de las observaciones.

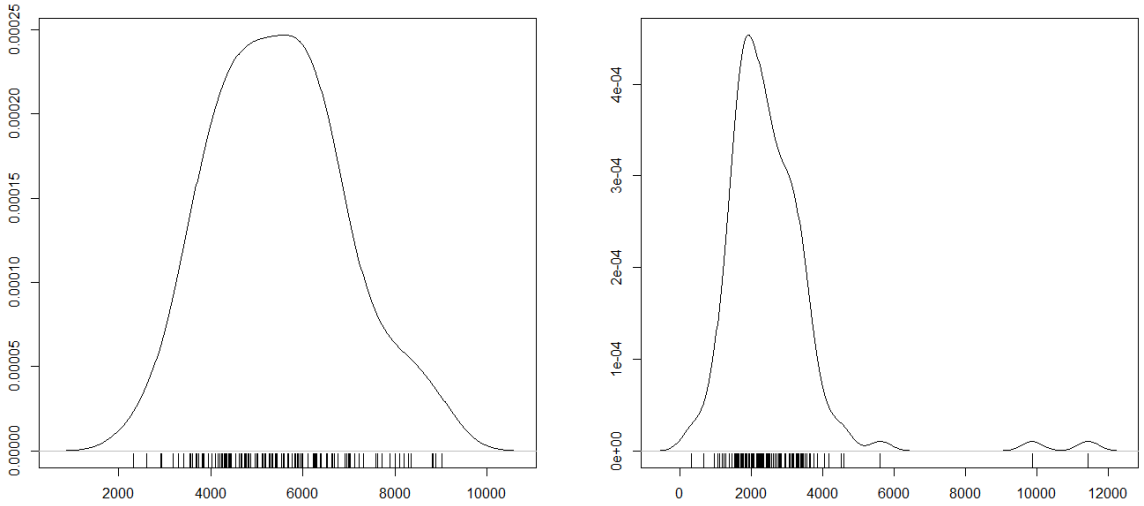

Figura 3. Gráficas de las estimaciones no parámetricas de las densidades de la concentración en $\mu g/g$, de potasio (izquierda) y de calcio (derecha), en las muestras del musgo P. purum recogidas en 132 estaciones de muestreo de la Comunidad Autónoma de Galicia en el año 2000.

1.3. Objetivos y estructura del trabajo

El objetivo de este trabajo es el desarrollo de un nuevo procedimiento que permita determinar si un punto geográfico está o no contaminado, en función una nueva regla de asignación de las observaciones a los niveles de una variable binomial en base a la distribución de concentración de los contaminantes en una red de muestreo regional. Para ello se recurrió inicialmente al ajuste de modelos de mixtura de normales para las concentraciones de metales y, posteriormente, a la elaboración de mapas de probabilidad de contaminación mediante la aplicación de técnicas de estadística espacial (kriging indicador).

Este trabajo se organiza en cinco capítulos. El primer capítulo contiene la introducción y motivación del trabajo, un breve análisis descriptivo de los datos y los objetivos y estructura del estudio. El Capítulo 2 presenta la técnica de modelado de las concentraciones con los modelos de mixturas de normales, atendiendo a la estimación del modelo, la selección del número de mixturas y el contraste de bondad de ajuste de los modelos ajustados. El Capítulo 3 incluye algunos experimentos de simulación realizados para evaluar las técnicas anteriores, presentando resultados relativos al análisis del funcionamiento de los estimadores y al contraste

de bondad de ajuste. El Capítulo 4 se dedica a la aplicación de la técnica a datos reales y su posterior análisis espacial para la creación de mapas de probabilidad de contaminación. Se revisarán algunos conceptos básicos de geoestadística, con atención especial a la predicción kriging. El Capítulo 5 incluye la discusión y las conclusiones del trabajo.

Este es un estudio original, ya que nunca antes se ha aplicado esta técnica en la determinación del nivel de contaminación a partir de concentraciones de metales pesados en musgos. Para ello, se han empleado funciones de paquetes de R ya definidos. En aquellos casos donde se empleen funciones/paquetes disponibles en R, se indicará en el texto.

Capítulo 2

2. Modelado de concentraciones de metales: modelos de mixturas

Este capítulo se dedica a la presentación de los modelos de mixturas, en concreto, de mixturas de distribuciones normales, como una alternativa para la modelización de datos heterogéneos. Además, en las secciones de este capítulo también se abordan cuestiones como la estimación del modelo o del número de sus componentes y se propone un contraste de bondad de ajuste para validar el modelo.

2.1. Introducción

En el Capítulo 1 se presentaron algunas gráficas de las densidades estimadas de las concentraciones de algunos metales pesados. Tal y como se podía observar en la mayoría de los casos, estas densidades estimadas no presentaban una forma paramétrica claramente identificable, como podría ser la de una densidad normal. De hecho, se observaba una importante asimetría positiva e incluso era posible intuir la existencia de distintos grupos de datos, dependiendo de la magnitud de las observaciones. Ante esta circunstancia, no parece adecuado ajustar modelos normales a los datos de concentraciones de metales, haciéndose necesario emplear modelos más flexibles que sean capaces de capturar tanto la asimetría como la posible existencia de grupos.

A pesar de que los modelos no paramétricos nos permiten “intuir” la distribución y cumplen con el requisito de flexibilidad comentado en el párrafo anterior, estos no nos aportan parámetros interpretables y tampoco nos permiten clasificar las observaciones en grupos. Teniendo en cuenta lo expuesto anteriormente, se consideró, como una opción que cumplía los requisitos necesarios, un modelo de mixtura de normales.

La regla de asignación a distintos grupos únicamente tiene sentido en los casos en los que haya duda sobre la contaminación o no de las muestras. Por lo tanto, puntos que tienen concentraciones de metales pesados muy elevados (que superan el límite superior de detección de atípicos, de aquí en adelante límite superior) son directamente considerados como puntos

contaminados y no se tienen en cuenta a la hora de hacer el ajuste del modelo de mixtura de normales.

$$Límite\ superior = Q_3 + 1.5 * RIC$$

donde Q_3 es el tercer cuartil y RIC es el rango intercuartílico.

En la Sección 2.2 se presenta el modelo de mixturas de normales, abordando su estimación en la Sección 2.3. En la Sección 2.4 se trata la selección del número de componentes del modelo y, por último, en la Sección 2.5 se expone el contraste de bondad de ajuste del modelo.

2.2. Modelo de mixtura de normales

Las distribuciones obtenidas mediante mixturas de modelos paramétricos son utilizadas para la modelización de datos heterogéneos en multitud de situaciones experimentales, en donde aquéllos pueden interpretarse como procedentes de dos o más subpoblaciones (componentes). Dado que, en nuestro caso, los datos proceden de puntos contaminados y no contaminados, este tipo de distribución resulta adecuada para identificar los distintos grupos y a la vez tener parámetros interpretables en cada uno de los grupos

Para introducir la construcción de los modelos de mixturas, supongamos que Y es una variable aleatoria unidimensional con densidad f . Esta densidad responde a un modelo de mixtura si se puede escribir como:

$$f(y|\Theta) = \sum_{i=1}^{k} \pi_i f_i(y|\theta_i), \quad y \in \mathbb{R}$$

donde k denota el número de densidades que componen la mixtura, Θ es el vector de parámetros del modelo (que englobaría a los parámetros de cada una de las componentes) y π_i es la proporción o peso de la componente i. De este modo, $f_i(y|\theta_i)$ es la función de densidad de la componente i con vector de parámetros θ_i.

Nótese que las proporciones de las mixturas son valores positivos que suman 1, de modo que se pueden interpretar, para cada valor de y, como la probabilidad de que dicho valor "pertenezca" a cada una de las componentes.

A pesar de que en este caso se presentan los modelos de mixturas para el caso unidimensional la formulación de este tipo de modelos se puede extender a cualquier vector aleatorio multidimensional, si bien debe tenerse en cuenta el incremento del número de parámetros.

Un caso particular de modelos de mixturas es el de las mixturas de distribuciones normales, donde los parámetros de cada componente serían la media y la varianza/desviación típica de cada una de ellas. Marron y Wand (1992) presentaron una colección de modelos de mixturas de normales altamente asimétricos y/o multimodales. Como se puede ver en la Figura 4, obtenida a mediante el paquete nor1mix (Maechler 2019), donde se muestran los modelos diseñados por Marron y Wand (1992), los modelos de mixturas de normales nos permiten modelizar datos con distribuciones muy diversas.

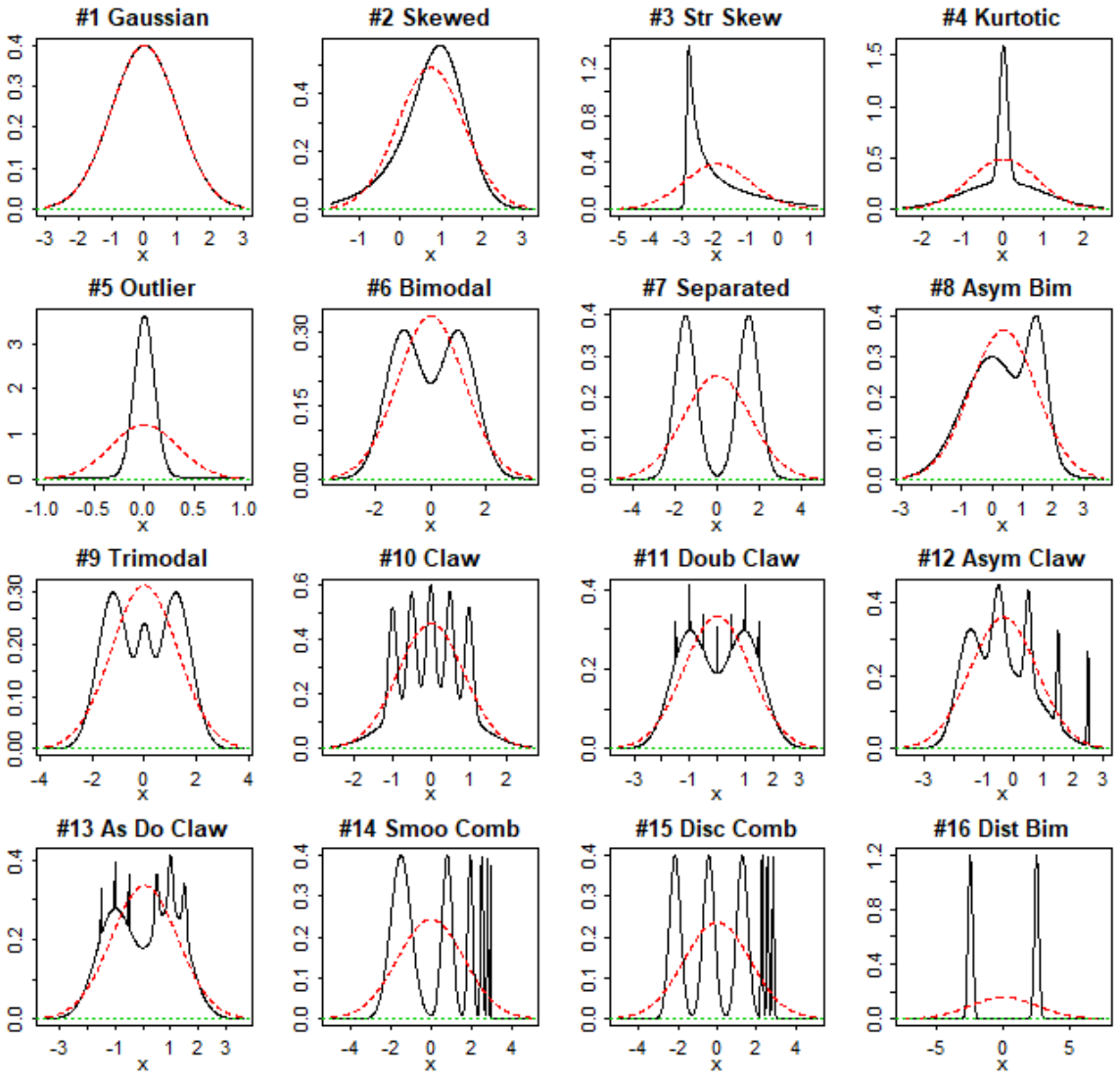

Figura 4. Densidades de los distintos modelos simulados por Marron y Wand (1992). A excepción del modelo #1, que es una distribución normal, el resto de los modelos son mixturas de normales con distintas componentes y parámetros.

Como se acaba de comentar y tal y como se puede deducir de la Figura 4, los modelos de mixturas de normales permiten una gran flexibilidad, a la vez que los parámetros son fácilmente interpretables en los distintos grupos. Por tanto, en este trabajo casi todas las distribuciones mixtas serán distribuciones de normales, con un máximo de 3 componentes gaussianas. Las únicas distribuciones mixtas cuyas componentes no serán todas gaussianas, son las de algunos modelos empleados en los estudios de simulación, en escenarios bajo la hipótesis alternativa. A continuación, se expone un ejemplo de la función de densidad para una mixtura de 3 componentes gaussianas, que sería como sigue:

$$g(y|\Theta) = \pi_1\phi(y|\mu_1,\sigma_1) + \pi_2\phi(y|\mu_2,\sigma_2) + \pi_3\phi(y|\mu_3,\sigma_3)$$

donde $\phi(\cdot)$ es la función de densidad de la gaussiana y μ_i y σ_i la media y desviación típica de la componente i del modelo.

Una última apreciación con respecto a los modelos de mixturas de normales es que el número de componentes no se identifica necesariamente con el número de grupos, entendiendo como tal el número de modas de la densidad. Este es el caso, por ejemplo, del modelo 4 de Marron y Wand (1992) (Figura 4), donde claramente se observa una única moda, si bien la densidad se corresponde con una mixtura de dos normales, ambas con la misma media y distintas varianzas.

2.3. Estimación del modelo

En los modelos de mixturas de normales hay un número elevado de parámetros a estimar ($3k - 1$, incluyendo medias, desviaciones típicas y $k - 1$ pesos). El problema que se da a la hora de obtener el estimador de máxima verosimilitud, ya que no es posible resolver de forma analítica las ecuaciones que se obtienen al derivar respecto al parámetro de interés. El origen de esta problemática está en que se desconoce a qué componente pertenece cada observación.

Por ello, en nuestra muestra $y_1, \dots, y_n$ puede considerarse que falta información, concretamente, la que indica a qué componente pertenece cada observación. Esta información estaría contenida en un vector de variables latentes z de tamaño n $(z_1, \dots, z_n)$ donde z_m toma valores desde 1 hasta k, e indicaría la componente a la que pertenece la observación m. En esta coyuntura es donde se recurre al algoritmo Esperanza-Maximización (*Expectation-Maximization* y, en adelante, EM, introducido por Dempster *et al.* (1977) para tratar de encontrar la estimación de máxima verosimilitud del modelo.

Como ya se comentó, resulta imposible obtener la estimación de máxima verosimilitud en un solo paso, por lo que se recurre al algoritmo EM, que consta de dos pasos, el E (esperanza o *expectation*) y el M (maximización o *maximization*). Estos pasos que se iteran repetidamente hasta que se cumple el criterio de parada y se obtiene una buena aproximación de la estimación de máxima verosimilitud del modelo.

Dado que el vector z es desconocido, y sin él no se pueden estimar los parámetros del modelo, en el paso E se calcula la probabilidad a posteriori de las variables latentes, dada por

$$P(z|y,\theta^0)$$

donde θ^0 es un vector donde se recogen los valores iniciales de los parámetros empleados para inicializar el algoritmo y poder estimar las probabilidades a posteriori de z (es decir, la probabilidad de pertenencia de cada observación a cada uno de los grupos, dados unos valores iniciales de los parámetros del modelo y la muestra observada). Inicialmente, dado un vector de parámetros θ^0 los valores del vector de variables latentes se pueden tomar a partir de asignaciones "soft" (que son valores en (0,1) o bien con asignaciones "hard", que son valores $\{0,1\}$. Estas probabilidades son importantes en la expresión de la esperanza de la log-verosimilitud de los datos completos ($\{y,z\}$) condicionada a la muestra, que viene dada por:

$$Q(\theta,\theta^0) = E_{z|y,\theta^0}\big[\log\big(P(y,z|\theta)\big)\big] = \sum_z P(z|y,\theta^0)\log\big(P(y,z|\theta)\big).$$

En el paso M se determina una nueva estimación $\hat{\theta}$ de los parámetros al maximizar Q:

$$\hat{\theta} = \text{argmax}_\theta Q(\theta,\theta^0)$$

El algoritmo se repite hasta que se cumple el criterio de parada, que normalmente es que el incremento en la verosimilitud del modelo aumente una cantidad despreciable con respecto al modelo estimado en el ciclo anterior del algoritmo.

Para llevar realizar el ajuste del modelo mediante el algoritmo EM, en este trabajo se empleó la función este test se empleó la función **normalmixEM** del paquete de R **mixtools** (Benaglia *et al.* 2009).

2.4. Selección del número de componentes

Para que el algoritmo EM realice una estimación de los parámetros del modelo, se le debe proporcionar previamente el número (k) de componentes de la mixtura (en nuestro caso, de densidades normales). En muchas ocasiones, hay conocimiento o estudios previos sobre los grupos que forman la distribución de los datos, por lo que se puede establecer el número de componentes sin necesidad de hacer estimaciones al respecto. En otros casos, como ocurre en este trabajo, lo que se desconoce e interesa es precisamente identificar estos grupos, por lo que se vuelve imprescindible una estimación precisa del número de componentes que forman la mixtura de normales.

Hay distintos métodos para la estimación de esta k, pero los dos métodos más utilizados, recogidos por McLachlan y Peel (2000), emplean la verosimilitud de la estimación obtenida. El primero de los métodos, considera un criterio que engloba tanto al criterio de información de Akaike (AIC) y como al criterio de información bayesiano (BIC), métodos clásicos en la selección de modelos. Por tanto, se basa en considerar como mejor k aquel donde se maximiza la verosimilitud, pero teniendo en cuenta una "penalización" de la verosimilitud del modelo al aumentar el número de parámetros del mismo. El segundo de los métodos considera un contraste de hipótesis cuyo estadístico es el de razón de verosimilitudes, es decir

$$-2\log\lambda = 2\{\log L(\widehat{\Theta}_1) - \log L(\widehat{\Theta}_0)\}$$

donde L la función de verosimilitud y $\widehat{\Theta}_1$ y $\widehat{\Theta}_0$ son los vectores de parámetros estimados, respectivamente, bajo alternativa ($k+1$ componentes) y bajo la nula (k componentes).

En este trabajo únicamente se utilizó el método basado en el contraste de hipótesis. A pesar de que el AIC y el BIC son más simples, el contraste de hipótesis nos permite calcular un p-valor De manera más específica, la hipótesis nula que se contrasta con este test es:

$$H_0: k = k_0$$

frente a la alternativa

$$H_1: k = k_1$$

donde k_0 es el número de componentes del modelo bajo la hipótesis nula y $k_1 = k_0$+1 el número de componentes del modelo bajo la hipótesis alternativa.

Esta hipótesis se contrasta de forma sucesiva ($k_0 = 1\ vs\ k_1 = 2,\ k_0 = 2\ vs\ k_1 = 3 \ldots$) hasta que no se rechaza la hipótesis nula, es decir, hasta que el p-valor obtenido en el test sea mayor que nivel de significación fijado (en este caso del 5%). Si bien al tener un estadístico de razón de verosimilitudes se pudiera pensar en utilizar una distribución asintótica chi-cuadrado para obtener los p-valores en la práctica, debe notarse que, en el caso de mixturas de distribuciones, se pierden las condiciones de regularidad clásicas, lo que hace que sea necesario recurrir a un algoritmo bootstrap:

1. Dada una muestra $Y_1 \ldots Y_n$ de la variable aleatoria Y, obtener, mediante el algoritmo EM, una estimación de $\widehat{\Theta}_1$ con $k_1 = 2$ y de $\widehat{\Theta}_0$ con $k_0 = 1$ y calcular el estadístico de razón de verosimilitud $-2 \log \lambda^{obs}$.
2. Obtener $B = 1000$ remuestras de tamaño n (igual al de la muestra original, $Y_1^* \ldots Y_n^*$) obtenidas a partir del modelo de mixtura con vector de parámetros $\widehat{\Theta}_0$, ajustado bajo la hipótesis nula en el paso 1.
3. Ajustar en cada muestra dos modelos mediante el algoritmo EM, uno con k_0 y otro con k_1 componentes.
4. Calcular el valor del estadístico $-2 \log \lambda$ en cada una de las remuestras
5. Calcular el p valor como la proporción de veces que el estadístico calculado en cada una de las B remuestras supera al valor del estadístico observado. Si en base al p-valor obtenido en el paso 5 no se puede rechazar la hipótesis nula, entonces el algoritmo se detiene y asumimos que la distribución tiene una única componente. En caso de que la hipótesis nula se rechace, entonces volvemos al paso 1 con $k_0 = 2$ y $k_1 = 3$ y se vuelve a repetir todo el proceso. Los contrastes continúan, y se aumenta el número de componentes de forma sucesiva hasta que en el paso 5 el p-valor es superior al nivel de significación fijado.

Para llevar a cabo este test se empleó la función **boot.comp** del paquete de R **mixtools** (Benaglia *et al.* 2009).

2.5. Contraste de bondad de ajuste

Cuando se considera un modelo paramétrico, aunque sea flexible (como es el caso de los modelos de mixturas), para ajustar la distribución de una variable aleatoria, es importante contrastar la bondad de ajuste del modelo elegido con el fin de evitar una mala especificación

del mismo, que comprometería las conclusiones extraídas a partir del ajuste. Para contrastar la bondad de ajuste de los modelos de mixturas, se empleó un test basado en la distancia L_1, donde el estadístico de contraste mide la integral del valor absoluto de las diferencias entre la estimación tipo núcleo de la función de densidad y la función de densidad de la hipótesis nula a testar.

De modo más concreto, el test de bondad de ajuste se emplea para realizar el contraste,

$$H_0: F(y) = F_0(y)$$

$$H_1: F(y) \neq F_0(y)$$

donde F es la función de distribución desconocida a partir de la que se originó la muestra $y_1 \dots y_n$ y F_0 una función de distribución completamente especificada, la del modelo ajustado mediante el algoritmo EM.

Como primer paso, se estiman los parámetros de la mixtura bajo la hipótesis nula utilizando el algoritmo EM, obteniendo f_n, que sería un estimador paramétrico de la densidad bajo la hipótesis nula. Con la misma muestra de datos, se calcula un estimador tipo núcleo $f_{n,h}$ (donde los subíndices indican que se ha obtenido con una muestra de tamaño n y con un parámetro de ventana h). Finalmente, el estadístico de contraste, que denotaremos por $T_{n,h}$ (Pavia 2015), se obtiene como:

$$T_{n,h} = \int_{-\infty}^{\infty} \left| f_{n,h}(y) - f_n(y) \right| dy.$$

Para calcular el p valor del test basado en el estimador $T_{n,h}$, se emplea integración numérica (para el cálculo del estadístico) y simulación Monte Carlo. Como indica Pavia (2015), tras el cálculo mediante integración numérica entre la función de densidad bajo la hipótesis nula y la estimación tipo núcleo de la densidad podemos obtener un p valor del test mediante simulación.

Dada una muestra de tamaño n de nuestra variable aleatoria, los pasos del algoritmo Bootstrap para la aproximación del p-valor son los siguientes:

1. Dada una muestra $Y_1 \dots Y_n$ de la variable aleatoria Y, obtener una estimación de la función tipo núcleo de la densidad y el valor observado del estadístico de contraste T_{nh}^{obs}.
2. Obtener B remuestras de tamaño n (igual al de la muestra original, $Y_1^* \dots Y_n^*$) obtenidas a partir de f, densidad bajo la hipótesis nula.

3. Obtener la estimación tipo núcleo de la función de densidad para cada una de las muestras, $f_{n,h}^{*}$
4. Calcular el área entre la densidad teórica y cada una de las estimadas en el paso 3.
5. Calcular el p valor como la proporción de veces que el área calculada en cada una de las B remuestras (calculadas en el paso anterior) excede el valor de T_{nh}^{obs} obtenido de la muestra observada.

Para llevar a cabo este test se empleó la función **dgeometric.test** del paquete de R **GoFKernel** (Pavia 2015).

Capítulo 3

3. Estudio de simulación

En este capítulo se realizaron una serie de simulaciones para analizar el comportamiento de los estimadores y del contraste de bondad de ajuste. A lo largo del capítulo se presentan los modelos que se emplean en las simulaciones, el tipo de estudios que se realizan y los resultados de los mismos.

3.1. Introducción

Una vez planteado el procedimiento es preciso analizar tanto el desempeño de los estimadores del modelo ajustado, como el tamaño y potencia del test de bondad de ajuste. Para ello se emplearon técnicas Monte Carlo: en el primer caso, para determinar, con los valores de los estimadores de las distintas remuestras, el sesgo y el error cuadrático medio (ECM) de los estimadores de los distintos parámetros del modelo; y, en el segundo caso, para la obtención de la proporción de rechazos bajo la hipótesis nula y varias alternativas.

Con el fin de determinar el comportamiento de los estimadores se simularon tres escenarios, todos ellos modelos de mixturas de normales cuyas especificaciones se muestran en la Tabla 1 y se encuentran representados en la Figura 5. Puede observarse que los modelos de mixturas de normales considerados son tres de los modelos propuestos por Marron y Wand (1992). Los dos primeros se corresponden con mixturas de dos componentes, con la misma proporción, pero con distinta separación en los grupos (misma desviación en ambas componentes, pero distinta separación en las medias). El tercero de los modelos no es exactamente ninguno de los 16 modelos simulados por Marron y Wand pero sí se puede relacionar estrechamente con uno de ellos, el modelo trimodal (#9), aunque a diferencia de este, el modelo 3 de este trabajo tiene distintos pesos y desviaciones típicas en las tres componentes.

Tabla 1. Especificaciones de los modelos de mixtura de normales empleados en las simulaciones para determinar el sesgo y el error cuadrático medio de los estimadores, así como para determinar el tamaño del contraste de bondad de ajuste.

	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	π_1	π_2	π_3
Modelo 1	200	400	-	50	50	50	0.5	0.5	-
Modelo 2	200	600	-	50	50	50	0.5	0.5	-
Modelo 3	200	400	600	50	50	50	0.5	0.35	0.15

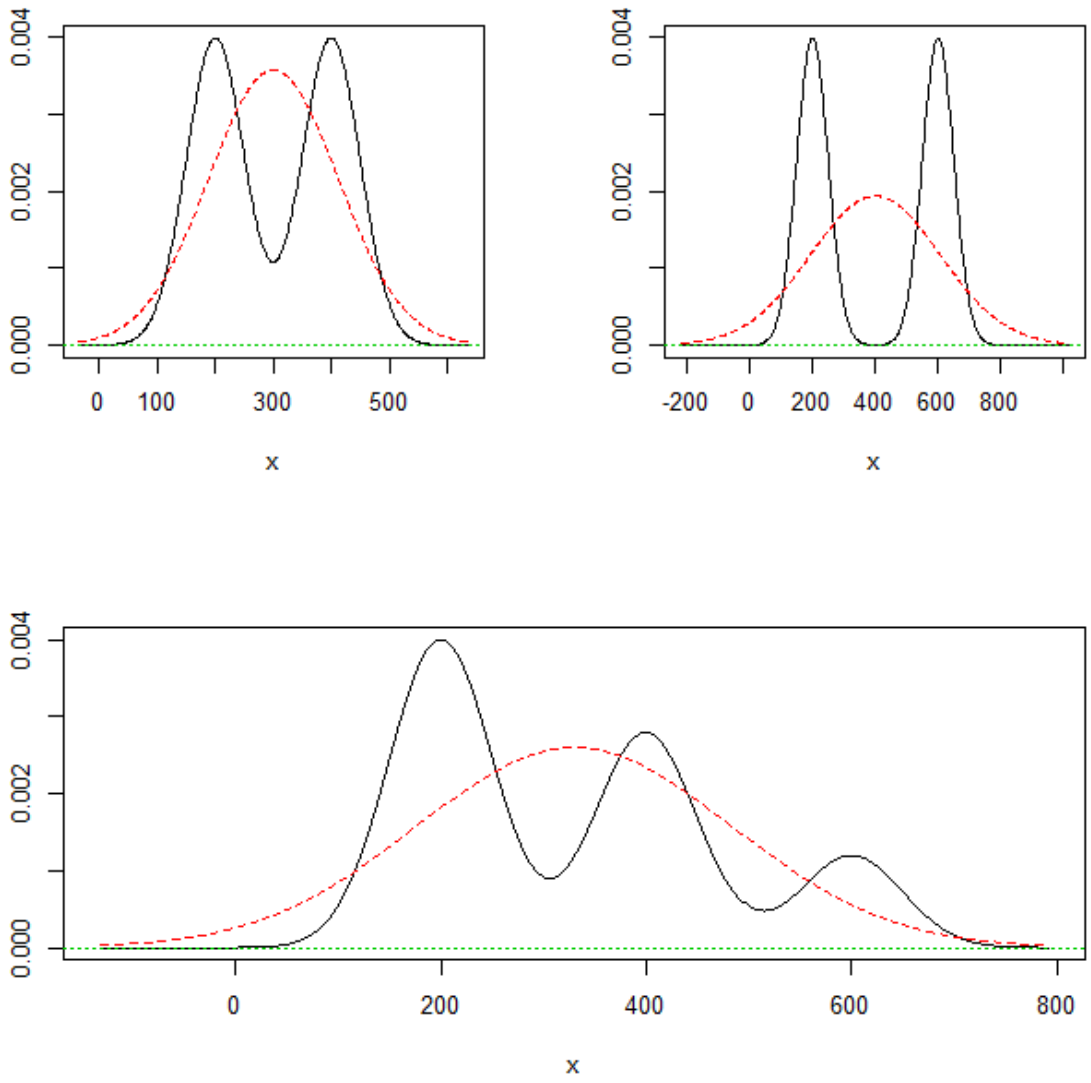

Figura 5. Gráficas con las densidades de los modelos especificados en la Tabla 1 (de derecha a izquierda y de arriba abajo: modelo 1, modelo 2 y modelo 3). En negro se representa la densidad del modelo de mixtura de normales y en rojo la densidad de la distribución normal con la misma media y desviación típica que el modelo de mixtura de normales.

Para analizar el tamaño del test, empleamos los modelos de mixturas de normales con los parámetros que se muestran en la Tabla 1, es decir, los mismos que se emplearon para la determinación del sesgo y el ECM de los estimadores. Para obtener la potencia del contraste de bondad de ajuste se recurrió a la simulación de otros seis modelos que se muestran en la Tabla 2. Los modelos A1, A2 y A3 especificados en esta tabla, son modelos de mixturas de dos componentes, la primera de ellas es una distribución normal y la segunda una distribución gamma. El modelo A4 es una gamma y los modelos A5 y A6 son mixturas de gammas (con 2 y 3 componentes respectivamente).

Los modelos alternativos considerados se han empleado para validar la potencia del test en los distintos escenarios de hipótesis nula (los mostrados en la Tabla 1). Los modelos A1, A2 y A3 de la alternativa se contraponen a los modelos 1 y 2 (es decir, considerando como hipótesis nula los modelos 1 y 2 de la Tabla 1). Por otra parte, los modelos A4, A5 y A6 de la alternativa se confrontan con el modelo 3 de la Tabla 1.

Tabla 2. Especificaciones de los modelos empleados en las simulaciones para determinar la potencia del contraste de bondad de ajuste. Los modelos A1, A2 y A3 son modelos de mixtura de una normal (primera componente) y una gamma (segunda componente), el modelo A4 es una distribución gamma y los modelos A5 y A6 son modelos de mixtura de gammas (dos y tres componentes respectivamente).

	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	π_1	π_2	π_3
Modelo A1	200	385	-	50	51.91	-	0.5	0.5	-
Modelo A2	200	600	-	50	54.77	-	0.5	0.5	-
Modelo A3	300	770	-	50	73.42	-	0.5	0.5	-
Modelo A4	288	-	-	117.58	-	-	-	-	-
Modelo A5	250	550	-	100	75	-	0.75	0.25	-
Modelo A6	200	400	600	50	50	50	0.5	0.35	0.15

En las Figuras 6 y 7 se pueden observar las funciones de densidad de los modelos presentados en la Tabla 2. Las densidades de estos modelos se representan sobre la densidad del modelo del escenario de hipótesis nula bajo el cual se simularon. Es decir, la densidad de los modelos A1, A2 y A3 (Figura 6) se representan, por una parte, sobre la densidad del modelo 1 de la hipótesis nula (Tabla 1), y, por otra, sobre la densidad del modelo 2 de la hipótesis nula; las densidades de los modelos A4, A5 y A6 (Figura 7) se representan sobre la densidad del modelo 3 de la hipótesis nula.

Como se puede observar en la Figura 6 el modelo A1 es muy similar al modelo 1 (a pesar de que las distribuciones que forman las mixturas son distintas en cada caso), por lo que es probable que el test de bondad de ajuste no sea capaz de detectar diferencias entre ambos modelos. Lo mismo le ocurre al modelo A2 comparado con el modelo 2 de la nula y al modelo A6 de la alternativa con el modelo 3 de la nula. En todos estos casos es probable que la proporción de rechazos de la hipótesis nula sea muy baja, acercándose al nivel de significación.

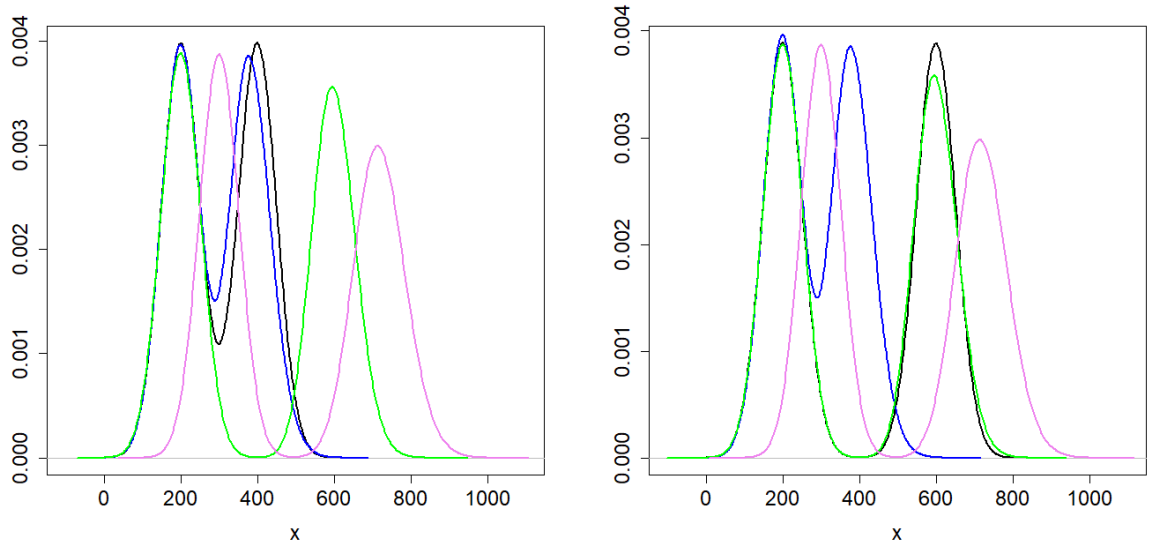

Figura 6. Gráfica de las densidades de los modelos A1 (en azul), A2 (en verde) y A3 (en violeta) bajo la alternativa representados sobre la densidad de los modelos 1 (izquierda) y 2 (derecha) de bajo la nula. La densidad de los modelos bajo la nula está representada, en ambos casos, por una línea negra.

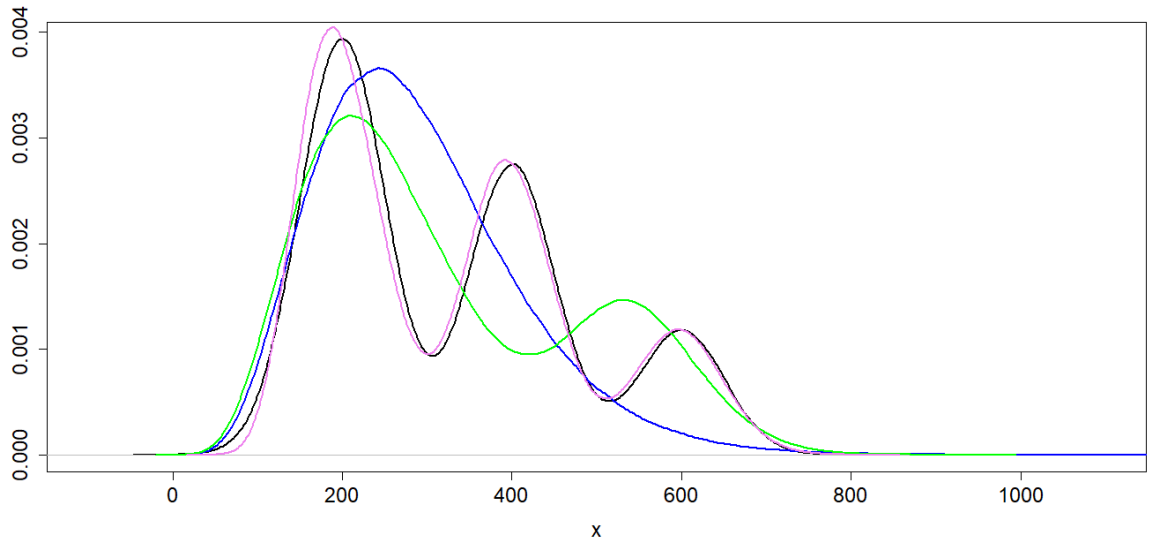

Figura 7. Gráfica de las densidades de los modelos A4 (en azul), A5 (en verde) y A6 (en violeta) bajo la alternativa representados sobre la densidad del modelo 3 bajo la nula (en negro).

3.2. Análisis de sesgo y el error cuadrático medio de los estimadores

El sesgo y el error cuadrático medio de los estimadores obtenidos mediante el algoritmo EM se calcularon mediante Monte Carlo. Los pasos seguidos fueron:

1. Obtener $B = 1000$ remuestras de tamaño n (con n igual a $100, 500$ y 1000) obtenidas a partir de uno de los escenarios de simulación presentados en la Sección 3.1.
2. Ajustar, mediante el algoritmo EM, un modelo de k componentes en cada remuestra, donde k es igual al número de componentes del escenario a partir del cual se obtiene la remuestra.
3. A partir de los modelos ajustados en cada remuestra calcular el sesgo y la cuasivarianza muestral ($\hat{S}^2$) de los estimadores de la manera siguiente

$$Sesgo = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_b - \theta\right)$$

$$\hat{S}^2 = \frac{\sum_{b=1}^{B}(\hat{\theta}_b - \bar{\hat{\theta}})^2}{B-1}$$

donde $\hat{\theta}_b$ es la estimación del parámetro θ obtenida en la remuestra b y $\bar{\hat{\theta}}$ es la media muestral de las estimaciones de lo parámetro θ obtenidas en las B remuestras.

Los resultados obtenidos se muestran en las Tablas 3 y 4; en la primera, se presentan los sesgos obtenidos para cada uno de los estimadores en cada uno de los escenarios y, dentro de cada escenario, cada uno de los tamaños de remuestra; en la segunda se presentan los errores cuadráticos medios (sesgo al cuadrado más varianza).

Como se puede observar en las Tablas 3 y 4, los sesgos y los errores cuadráticos medios de los estimadores disminuyen, como sería de esperar, al aumentar el tamaño de la muestra. A pesar de que también sería esperable que disminuyesen al aumentar la separación entre las medias (manteniendo las varianzas), esto no ocurre, o al menos no en la mayor parte de los casos. De forma contraria, tanto el sesgo como el ECM aumentan mucho al aumentar el número de componentes, sobre todo en las componentes que tienen menor peso.

Tabla 3. Sesgos de los estimadores (medias, desviaciones típicas y proporciones de las componentes) en cada uno de los escenarios de simulación considerados. Nótese que los escenarios 1 y 2 son mixturas de normales de dos componentes mientras que en el escenario 3 la mixtura es de tres componentes.

Escenario	n	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\sigma}_3$	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\pi}_3$
1	100	1.678	-2.027	-	0.115	0.236	-	0.001	-0.001	-
	500	0.724	-0.568	-	0.397	0.230	-	0.002	-0.002	-
	1000	0.488	-0.568	-	0.328	0.361	-	-0.001	0.001	-
2	100	0.849	-0.555	-	-0.499	-0.121	-	-0.002	0.002	-
	500	0.491	-0.619	-	0.316	0.334	-	-0.002	0.002	-
	1000	0.398	-0.320	-	0.131	0.276	-	0.001	-0.001	-
3	100	-7.203	-31.681	-37.928	-5.730	0.000	11.967	-0.067	-0.014	0.081
	500	-4.975	-22.821	-25.519	-2.640	-2.081	11.079	-0.038	-0.021	0.059
	1000	-3.351	-16.657	-19.462	-1.560	-1.497	8.363	-0.026	-0.018	0.044

Dado que las observaciones reales para las que se diseña este protocolo son dependientes, se hizo un pequeño estudio de simulación similar al anterior en el que se simularon dos modelos con datos dependientes y se obtuvieron los sesgos y los ECMs de las estimaciones.

Tabla 4. Errores cuadráticos medios de los estimadores (medias, desviaciones típicas y proporciones de las componentes) en cada uno de los escenarios de simulación considerados. Nótese que los escenarios 1 y 2 son mixturas de normales de dos componentes mientras que en el escenario 3 la mixtura es de tres componentes.

Escenario	n	μ_1	μ_2	μ_3	σ_1	σ_2	σ_3	π_1	π_2	π_3
1	100	264.00	245.79	-	152.34	148.87	-	0.01	0.01	-
	500	83.26	81.34	-	35.58	34.88	-	0.00	0.00	-
	1000	62.62	69.41	-	26.28	26.39	-	0.00	0.00	-
2	100	165.74	172.23	-	101.05	117.02	-	0.00	0.00	-
	500	124.76	129.62	-	79.27	79.43	-	0.00	0.00	-
	1000	81.30	81.21	-	50.80	51.08	-	0.00	0.00	-
3	100	579.37	5534.86	7129.58	221.76	836.77	1432.09	0.02	0.02	0.03
	500	308.30	4137.38	4678.01	67.65	232.10	956.57	0.01	0.01	0.03
	1000	158.22	3031.61	3459.69	34.81	175.90	709.93	0.01	0.01	0.02

Los modelos simulados (véase Tabla 5) se corresponden con mixturas de dos normales (con el mismo peso) de medias 0 y 3 en el primer escenario y 1 y 3 en el segundo escenario, donde las observaciones en cada una de las componentes presentaban una estructura de dependencia espacial. En concreto, en ambos casos dos campos gaussianos con variograma exponencial, de manera que los parámetros del modelo de variograma son la varianza puntual y el rango de

dependencia. Para simular las realizaciones de los campos espaciales, se generaron datos en un grid regular en el cuadrado unidad. Se han considerado dos tamaños de muestra (100 y 400 datos, respectivamente). En la Figura 8 se puede observar una realización de $n = 100$ de ambos escenarios.

Tabla 5. Especificación de los parámetros de los modelos con datos dependientes. Escenarios 1 y 2: mixtura de dos normales con estructura de dependencia exponencial.

	Campo A				Campo B			
Escenario	μ	σ^2	$Rango$	π	μ	σ^2	$Rango$	π
1	0	0.5	0.2	0.5	3	0.1	0.1	0.5
2	1	1	0.2	0.5	3	0.2	0.1	0.5

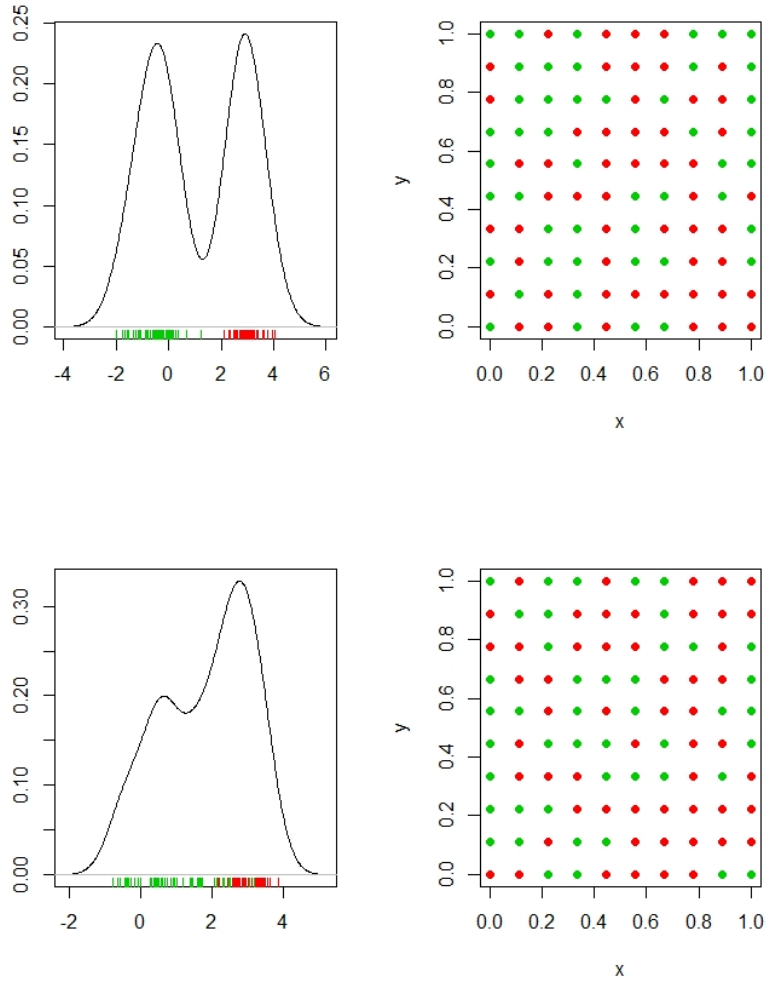

Figura 8. Realización de los modelos simulados bajo dependencia espacial. Izquierda: estimación de la densidad con muestras de 100 datos para el escenario 1 (arriba) y el escenario 2 (abajo) Derecha: localizaciones de los puntos correspondientes a cada una de las componentes de la mixtura. Tamaños de muestra: 100.

Tabla 6. Sesgos y errores cuadráticos medios, con tamaños de muestra de 100 y 400, de los estimadores en los dos escenarios simulados para datos dependientes

Escenario	n		$\hat{\mu}_A$	$\hat{\mu}_B$	$\hat{\sigma}_A^2$	$\hat{\sigma}_B^2$	$\hat{\pi}$
1	100	Sesgo	3.05e04	-3.57e04	2.79e04	5.13e04	-4.18e03
		ECM	0.536	0.505	0.017	0.015	0.004
	400	Sesgo	2.08e04	-1.74e04	2.91e04	4.89e04	1.65e04
		ECM	0.324	0.256	0.010	0.006	0.001
2	100	Sesgo	3.08e04	-4.34e04	-3.14e04	5.69e04	-8.37e03
		ECM	0.512	0.389	0.058	0.036	0.015
	400	Sesgo	2.10e04	-3.12e04	-2.69e04	5.62e04	-1.29e02
		ECM	0.420	0.275	0.037	0.023	0.007

Los sesgos y los ECMs obtenidos para las estimaciones de los datos con dependencia se muestran en la Tabla 6. Se puede observar que los resultados son similares para ambos escenarios, si bien en el escenario 2 se incrementan los ECM con respecto a los obtenidos para el escenario 1, como resulta esperable, ya que en el escenario 2 se aumenta la varianza del campo A y también se aproximan más las medias, dificultando la estimación.

Se puede ver que los ECMs de los datos simulados con dependencia son mucho menores que los de la Tabla 4, lo cual puede resultar sorprendente. Analizando en detalle el proceso de simulación, esto puede deberse a que en el caso de datos independientes se observó que en algunas ocasiones el algoritmo EM no proporcionó resultados satisfactorios, algo que no ocurrió en el caso de datos dependientes. Un ejemplo de estas situaciones se puede ver en la Figura 9, donde se representa la estimación de la densidad no paramétrica de las estimaciones de las medias realizadas en las simulaciones del escenario 1. A pesar de que no se muestra en este trabajo, se obtuvieron simulaciones para datos independientes que arrojaron ECMs de un orden de magnitud menor al que se ve en la Tabla 4 en los casos en que no se detectó la presencia de estas estimaciones anómalas, que por ejemplo en la Figura 9 están en torno a un valor de 300 cuando las medias del modelo son 200 y 400.

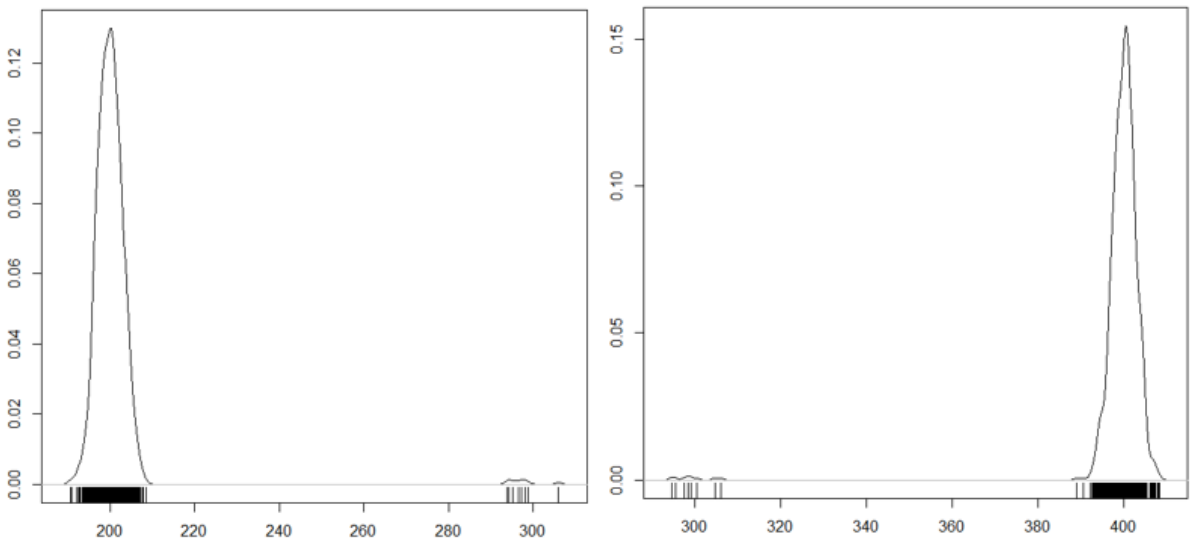

Figura 9. Estimación no paramétricas de la densidad de las estimaciones de las medias de la primera componente (izquierda) y la segunda componente (derecha) ajustadas bajo el escenario de simulación 1 de datos independientes con un tamaño de muestra de 1000.

3.3. Tamaño y potencia del contraste de bondad de ajuste

El comportamiento del contraste de bondad de ajuste se analiza en términos de tamaño (porcentaje de rechazos bajo la hipótesis nula) y potencia (porcentaje de rechazos bajo la hipótesis alternativa). Fijado un nivel de significación teórico (en nuestro caso, del 5%), el porcentaje de rechazos bajo la nula debería aproximar este valor (es decir, debería estar bien calibrado), mientras que bajo la alternativa, el contraste debería ser capaz de identificar las desviaciones de la nula.

Para analizar el tamaño del contraste de bondad de ajuste también recurrimos al bootstrap. Los pasos fueron:

1. Obtener $B = 500$ remuestras de tamaño n (con n igual a $100, 500$ y 1000) a partir de uno de los escenarios de simulación bajo la hipótesis nula presentados en la Sección 3.1.
2. En cada una de las remuestras se aplica el contraste de bondad de ajuste, testando la hipótesis nula de que la remuestra provenga de la densidad bajo la hipótesis nula.
3. Con los p-valores obtenidos en el paso 2 se determina la proporción de rechazos para un nivel de significación de 0.05, es decir, el porcentaje de p-valores menores de 0.05.

Los resultados, que se muestran en la Tabla 7, indican que, para todos los escenarios bajo la hipótesis nula y todos los tamaños de remuestra, la proporción de rechazos de la hipótesis nula se acerca mucho al nivel de significación elegido de 0.05.

Tabla 7. Proporción de rechazos de la hipótesis nula cuando los datos han sido generados bajo la hipótesis nula (calibrado del test).

Escenario	$n = 100$	$n = 500$	$n = 1000$
1	0.064	0.072	0.062
2	0.058	0.066	0.048
3	0.056	0.052	0.066

Para determinar la potencia del contraste de bondad de ajuste el proceso es prácticamente idéntico. Lo único que cambia es que el modelo a partir del que se originan las remuestras es uno de los escenarios de simulación bajo la hipótesis alternativa.

Los resultados para el estudio de la potencia del contraste se muestran en la Tabla 8, donde se puede observar que el test es potente, ya que en la mayor parte de los escenarios la proporción de rechazos de la hipótesis nula es uno o cercano a uno. Incluso en las ocasiones en las que la alternativa es similar al modelo nulo (como, por ejemplo, en el modelo A1 de la alternativa bajo el escenario 1) el porcentaje de rechazos se aleja del nivel de significación y aumenta rápidamente al aumentar el tamaño muestral. En el único caso donde el test no es capaz de rechazar la hipótesis nula es para el modelo A6 bajo el escenario 3, lo que era esperable, ya que como se ve en la Figura 7 las curvas del modelo 3 y el modelo A6 prácticamente se solapan.

Tabla 8. Proporción de rechazos de la hipótesis nula cuando los datos han sido generados bajo la hipótesis alternativa (potencia del test).

H_0	Alternativa	$n = 100$	$n = 500$	$n = 1000$
Escenario 1	1	0.308	0.908	0.996
	2	1	1	1
	3	1	1	1
Escenario 2	1	1	1	1
	2	0.166	0.412	0.630
	3	1	1	1
Escenario 3	4	0.998	1	1
	5	0.990	1	1
	6	0.036	0.042	0.038

Los resultados que se muestran en las Tablas 7 y 8 se han obtenido utilizando la ventada de la regla del pulgar y permitiendo que esta varíe en cada una de las realizaciones bootstrap. Con el fin de determinar si esto altera o no las conclusiones de nuestro estudio, se ha probado (tanto en situaciones de calibrado como de potencia) a fijando la ventana de la regla del pulgar y la plug-in de Seather-Jones, y considerando esta misma ventana en todas las simulaciones y dejando que la ventana varíe en las distintas réplicas. Se han contabilizado las discrepancias en los resultados del test (analizando si la conclusión de rechazo/no rechazo coincide considerando ventana fija en todas las simulaciones -lo que sería lo correcto- y ventanas variando). Tanto cuando se usa la ventana de escala normal como cuando se usa la plug-in, se producen discrepancias (resultados diferentes para fijar/no fijar) en el 1% de los casos para la plug-in y en el 2% de los casos en la regla del pulgar, aproximándose mejor al resultado (en calibrado) al considerar una ventana fija. Parece que se rechaza más con la ventana fija, pero los resultados no parecen invalidar lo reportado en las tablas y para extraer conclusiones sobre el impacto de la ventana fija/variando en remuestras se requeriría de un estudio más completo.

Capítulo 4

4. Aplicación a datos reales

4.1. Introducción

Como ya se comentó en la Sección 1.2, los datos que se emplearon en este trabajo fueron concentraciones de metales pesados en muestras de musgo recogidas en una res de muestreo que cubre toda Galicia. La concentración de estos metales sigue, a nivel regional y en presencia de puntos contaminados, una distribución multimodal con una importante asimetría positiva y, en ella, se puede intuir la presencia de distintos grupos de observaciones.

El objetivo de este trabajo, detallado en la Sección 1.3 es el de desarrollar un nuevo protocolo que permita determinar la probabilidad de que un punto geográfico está o no contaminado. Para ello, se recurrió a la modelización, mediante modelos de mixturas de normales, de la distribución de las concentraciones de metales pesados en la red de muestreo, de tal forma que cada observación se asignó a una componente del modelo. Esta asignación fue la base sobre la cual cada observación fue clasificada como “contaminada” o “no contaminada” y, a partir de esta clasificación se obtuvieron los mapas de probabilidad mediante kriging indicador.

En este Capítulo 4 se emplearon, como ya se indicó al inicio del trabajo, datos del año 2000. El procedimiento se ejemplificó, concretamente, con datos de concentraciones de mercurio.

Dado que el punto crucial del protocolo es la asignación de las observaciones a los niveles de una variable aleatoria binomial con niveles “no contaminado” y “contaminado”, la mayor parte del trabajo se ha centrado en el modelado de los datos. La parte de estadística espacial, no se describirá exhaustivamente, formando parte las herramientas empleadas de los contenidos de la materia de “Estadística Espacial” del Máster en Técnicas Estadísticas.

4.2. Ajuste de modelos de mixturas y asignación de grupos

Como ya se indicó en la Sección 2.1, el primer paso del método es el cálculo del límite superior para, a la hora de hacer el ajuste, no considerar los datos que lo superen, ya que distorsionarían el modelo. Además, estos datos están claramente contaminados, por lo que no tiene sentido considerarlos a la hora de ajustar el modelo. En la Figura 10 se puede ver cómo cambió la

estimación no paramétrica de la densidad de la concentración de Mercurio antes y después de eliminar estos datos “atípicos”.

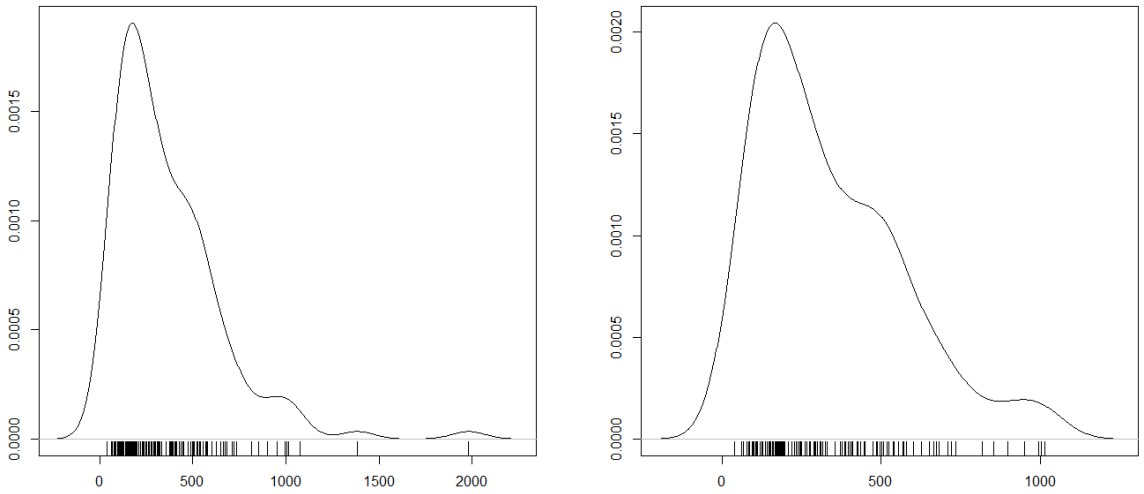

Figura 10. Estimación no paramétrica de la densidad de la concentración de mercurio (ng/g) con datos “atípicos” (izquierda) y sin ellos (derecha). Estas concentraciones se midieron en muestras del musgo P. purum recogidas en 132 estaciones de muestreo de la Comunidad Autónoma de Galicia en el año 2000.

El segundo paso es estimar el número de componentes del modelo de mixtura de normales, por lo que se aplicó el contraste de hipótesis indicado en la Sección 2.4. Para la implementación de este test, como ya se comentó previamente, se empleó la función **boot.comp** del paquete **mixtools** (Benaglia *et al.* 2009).

El primer contraste de hipótesis ($k_0 = 1 \; vs \; k_1 = 2$) arrojó un p-valor menor que los niveles de significación usuales, por lo que se rechazó la hipótesis nula y se procedió al segundo contraste ($k_0 = 2 \; vs \; k_1 = 3$), cuyo p-valor fue de 0.07. Dado que el nivel de significación fijado a priori fue de 0.05, no se rechazó la hipótesis nula y tomó $k = 2$ como estimación del número de componentes del modelo. Además de esto, la función **boot.comp** también proporciona, para cada contraste, un histograma de los valores del estadístico de razón de verosimilitud bootstrap (Figura 11). En esta figura también se puede observar una línea roja que indica el valor del estadístico observado en la muestra original.

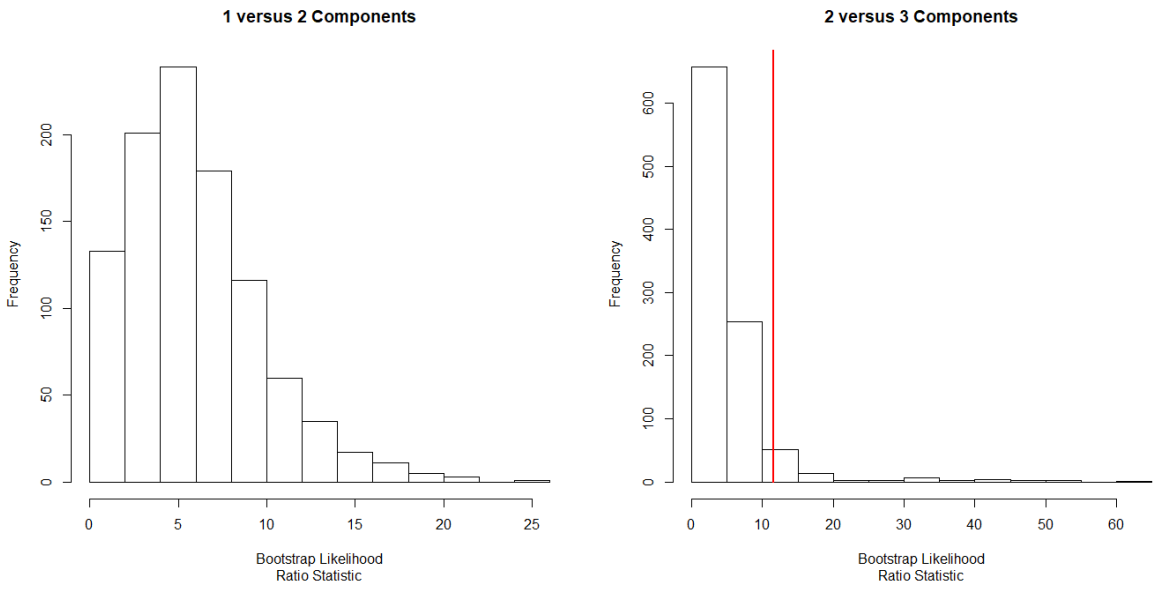

Figura 11. Histogramas de los valores del estadístico de razón de verosimilitud calculados mediante bootstrap durante el proceso de estimación del número de componentes del modelo para la concentración de mercurio (ng/g) en el musgo P. purum. A la izquierda se representan los valores del primer contraste *($k_0 = 1$ vs $k_1 = 2$)* y a la derecha los del segundo contraste *($k_0 = 1$ vs $k_1 = 2$)*. En rojo se indica el valor del estadístico de razón de verosimilitud en la muestra original.

El tercer paso es el ajuste del modelo de mixtura de normales. Es decir, para el caso del mercurio, es la estimación, mediante el algoritmo EM y a partir de la muestra original, de los parámetros de un modelo de mixtura de normales de dos componentes. En este caso las estimaciones del modelo se muestran en la Tabla 9 y su densidad se puede observar en la Figura 12 junto con la estimación no paramétrica de la densidad de los datos originales.

Tabla 9. Estimaciones de los parámetros de un modelo de mixtura de normales ajustado mediante el algoritmo EM para modelizar la distribución de las concentraciones de mercurio (ng/g) en el musgo P. purum.

	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$	$\hat{\pi}_1$	$\hat{\pi}_2$
Estimaciones	155.366	469.424	62.442	214.493	0.411	0.589

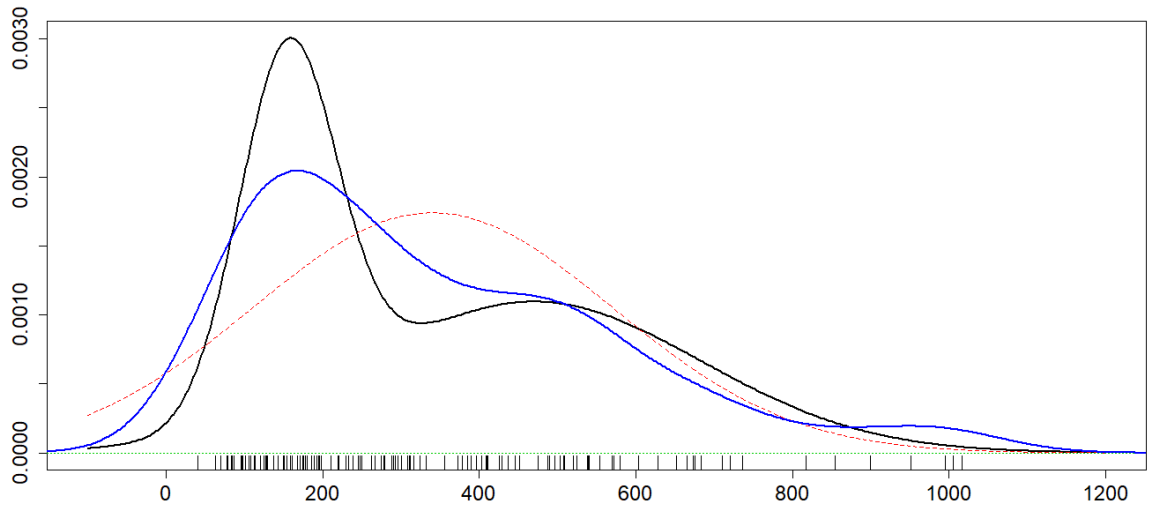

Figura 12. Densidad del modelo de mixtura de normales estimado mediante el algortimo EM (en negro) para modelizar las concentraciones de mercurio (ng/g) en el musgo P. purum. En azul se muestra la estimación no paramétrica de la densidad de los datos.

El cuarto paso es el del contraste de bondad de ajuste, para determinar si el modelo estimado se puede considerar apropiado o no. Como se indicó en la Sección 2.5, para aplicar el test, se recurrió a la función **dgeometric.test**, del paquete **GoFKernel** (Pavia 2015). El valor del estadístico $T_{n,h}$ observado, fue de 0.270, y el p-valor obtenido mediante bootstrap fue de 0.210. En base a estos resultados no se rechazó la hipótesis nula y el modelo estimado se consideró como válido.

Por último, una vez obtenido y validado el modelo, las observaciones que fueron asignadas a la primera componente, se clasificaron como “no contaminadas” mientras que el resto de las observaciones se clasificaron como “contaminadas”. El resultado de esta clasificación puede observarse en la Figura 13, donde se muestra un mapa de Galicia y la localización de las estaciones de muestreo clasificadas como “no contaminadas” (en verde) y “contaminadas” (en rojo).

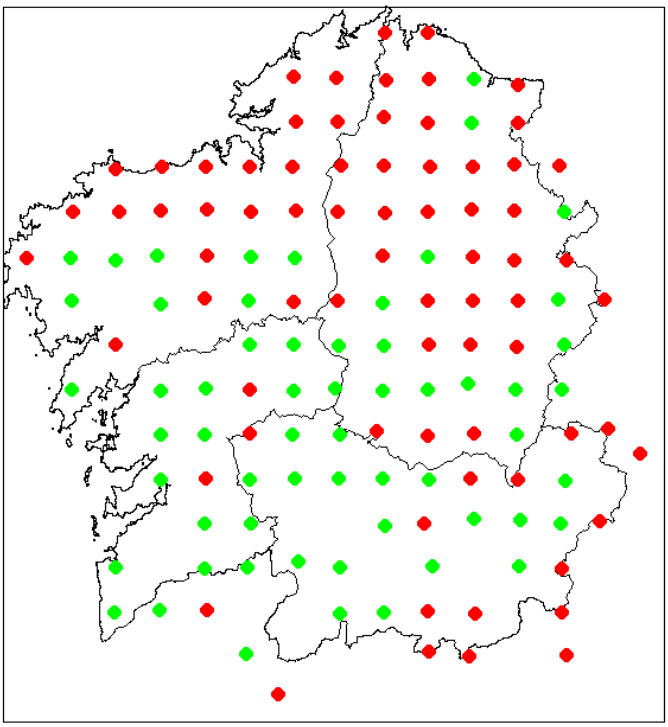

Figura 13. Estaciones de muestreo de la Comunidad Autónoma de Galicia clasificadas como "contaminadas" (en rojo) y "no contaminadas" (en verde) en función de la asignación de las observaciones a las componentes de un modelo de mixtura de normales ajustado a las concentraciones de mercurio (ng/g) en el musgo P. purum.

4.3. Análisis espacial

Como ya se ha indicado, las observaciones de los distintos metales tomadas en la red de localizaciones espaciales pueden considerarse como una realización de un proceso espacial continuo (geoestadístico). Tenemos n=132 localizaciones y denotemos por $y_1 = y(s_1), \dots, y_n = y(s_n)$ las observaciones de un proceso (mediciones de un metal pesado) en dichas localizaciones. Como paso inicial, realizaremos un análisis geoestadístico de esta realización del proceso, a través del análisis del variograma, que nos permitirá determinar si las observaciones presentan realmente dependencia espacial y si el proceso se puede considerar estacionario e isotrópico, dos requerimientos para aplicar técnicas de predicción kriging. Para una revisión clásica de muchas de las herramientas que se utilizarán en este capítulo, puede verse Cressie (1993).

4.3.1. Algunos conceptos básicos en estadística espacial

En primer lugar, para analizar la variabilidad, se construye un variograma empírico robusto basado en la raíz cuadrada del valor absoluto de las diferencias entre los datos. El variograma empírico se estima como

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} \left(y_i - y_j\right)^2$$

donde para cada para de observaciones *(i,j)* a una distancia h se calcula un conjunto de pares de localizaciones con distancia similar: $N(h) = \{(i,j) : h_{ij} \in b(h)\}$ y $b(h)$ es un intervalo que contiene h y $|N(h)|$ denota el cardinal de $N(h)$.

Una versión robusta del variograma empírico que se utiliza en este trabajo es una transformación de la expresión anterior, considerando la raíz cuarta de las diferencias al cuadrado, de manera que se obtiene:

$$\hat{\gamma}^*(h) = \frac{1}{|N(h)|} \sum_{(i,j) \in N(h)} \sqrt{|y_i - y_j|}$$

Como se indica en Bowman y Crujeiras (2013), esta transformación del variograma es, en esperanza,

$$\gamma^*(h) = 0.977741\{\gamma(h)\}^{\frac{1}{4}},$$

lo cual permite trabajar en la escala transformada y recuperar (a través de la inversa de la transformación anterior) el variograma en su escala original. Los autores justifican la consideración de la escala transformada porque es más robusta (se ve menos afectada por diferencias grandes entre las observaciones) y permite disminuir la correlación entre las diferencias.

En la práctica, a partir de una realización de un proceso espacial, se puede estimar este variograma en la escala transformada, obteniendo una nube de puntos, que a su vez se puede suavizar mediante las técnicas habituales empleadas en regresión (por ejemplo, métodos núcleo o splines).

A partir de la versión suavizada, se analiza la dependencia espacial de la muestra mediante un contraste de hipótesis, de tal forma que en caso de que se rechace la hipótesis nula de independencia, se prosigue con el análisis espacial; en caso contrario, no se continúa. Si los

datos son independientes, entonces el variograma debería ser plano, y esta es la idea en la que se basa el contraste propuesto por Diblasi y Bowman (1997) y posteriormente modificado a partir del trabajo de Bowman y Crujeiras (2013). El contraste compara a través de una forma cuadrática un estimador plano (bajo la hipótesis nula de independencia) con un estimador suavizado del variograma (bajo la hipótesis alternativa, donde no se requiere que el variograma sea plano).

Una vez realizados todos los contrastes, se ajusta un semivariograma paramétrico que se acerque al empírico y, a partir de él, se obtienen las predicciones kriging mediante kriging indicador y se representan mediante mapas de probabilidades.

4.3.2. Ajuste de la estructura de dependencia

El primer paso para el ajuste de la estructura de dependencia ha sido construir, a partir de las observaciones de los datos, el variograma empírico descrito en la Sección anterior y se representó en la imagen de la izquierda de la Figura 14. En esta misma figura también puede observarse el variograma junto con sus bandas de confianza, tal y como se detalla en Bowman y Crujeiras (2013). Para obtener ambas imágenes y para el cálculo del variograma se recurrió a la función **sm.variogram** del paquete de R **sm** (Bowman y Azzalini 2018).

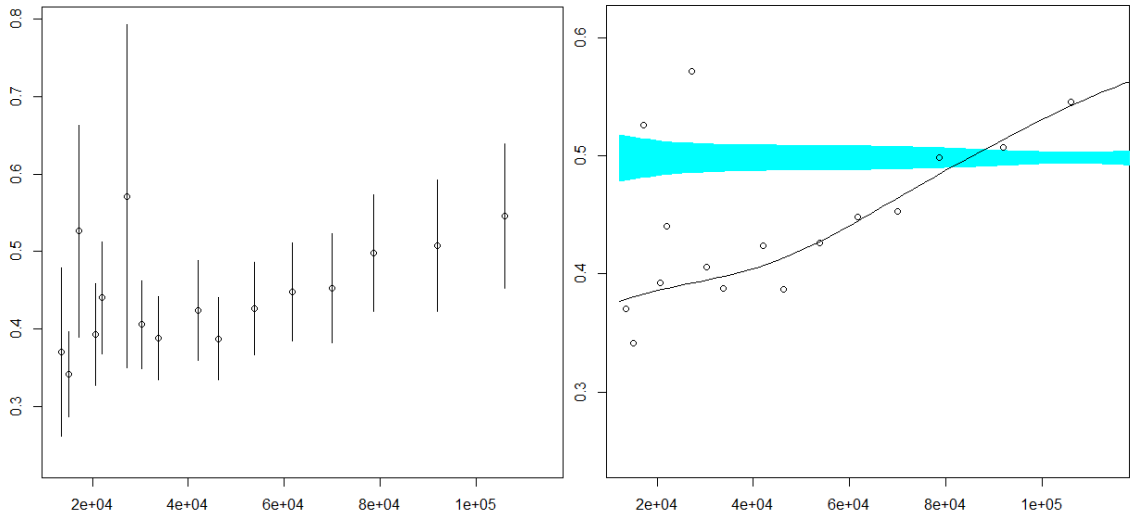

Figura 14. Estimaciones de la transformación del variograma empírico del mercurio junto con las bandas de confianza de la estimación (izquierda) y resultado del contraste de independencia (derecha), donde se ve la estimación del variograma y en azul se muestran la región de confianza en la que se tendría que situar el variograma para que no se rechace la hipótesis nula. En el eje de abscisas se representan las distancias y en el de ordenadas la raíz cuadrada de las diferencias absolutas entre las observaciones.

A continuación, se realizó el contraste de independencia, a través del cual se rechazó la hipótesis nula al obtener en el test un p-valor de 0.02. Al rechazar la hipótesis de independencia espacial, tiene sentido proseguir con el análisis espacial para la obtención de predicciones kriging. Para la realización de este test, también se recurrió a la función **sm.variogram**. El resultado gráfico se muestra en la imagen de la derecha de la Figura 14, donde se observa una versión suavizada del variograma (línea continua) junto con una banda de referencia para la hipótesis nula de independencia (banda azul), que se correspondería con un variograma plano.

Por último, se ajustó un variograma paramétrico circular (por ser el que mejor se ajustaba al variograma empírico), mediante un método de mínimos cuadrados (véase Cressie 1993), en base al cual se obtendrán las predicciones kriging.

4.3.3. Predicciones kriging

Para obtener los mapas de predicciones, es preciso recurrir a métodos kriging. Estos métodos son algoritmos de predicción que parten del principio de que puntos próximos en distancia son más similares entre sí que aquellos que se encuentran alejados. En base a esto emplean combinaciones lineales ponderadas de las observaciones para obtener las predicciones. Las observaciones más cercanas al punto predicho tienen un mayor peso en el valor de la predicción que las observaciones alejadas. Además, debe notarse que los métodos kriging son interpoladores y, por tanto, la predicción que devuelven en los puntos de muestreo coincide con el valor observado en los mismos.

Como ya se comentó, en general la finalidad de las predicciones kriging es obtener predicciones de valores que toman ciertos procesos estocásticos en una superficie y, por lo tanto, los métodos kriging más utilizados son los conocidos como kriging lineal, entre los que destacan el kriging simple, el kriging ordinario y el kriging universal. A pesar de ello, en este trabajo la finalidad no es predecir valores, sino estimar la probabilidad de que el proceso tome un valor menor o igual a un valor determinado en una localización s. Es decir, el objetivo es predecir:

$$F_{s(y)} = P(y(s) \leq y).$$

Por lo tanto, se recurrió al kriging indicador, que es un método kriging no lineal. Pese a ello, y bajo ciertas restricciones, la distribución de $F_{s(y)}$ puede aproximarse mediante kriging ordinario mediante la forma

$$\hat{F}_{s(y)} = \hat{I}(s,y) = \sum_{m=1}^{n} \lambda_m \, I(s_m, y) = \sum_{m=1}^{n} \lambda_m \, I_{\{y(s_m) \leq y\}}$$

donde λ_m es el peso de la observación m $I(s,y) = I_{\{Y(s) \leq y\}}$ es la función indicadora

$$I(s,y) = I_{\{y(s) \leq y\}} = \begin{cases} 1, & si\ Y(s) \leq y, \\ 0, & si\ Y(s) > y. \end{cases}$$

En este trabajo la función indicadora $I(s,y)$ ya viene dada por el vector de unos y ceros que contiene las observaciones clasificadas como “contaminadas” y “no contaminadas”.

Tras la aplicación del kriging indicador (aproximado mediante kriging ordinario) con la función **krige** definida en el paquete de R **gstat** (Pebesma 2004)**,** se obtuvieron las estimaciones de probabilidad de contaminación que se muestran en la Figura 15. En la Figura 16 se observan las varianzas de predicción.

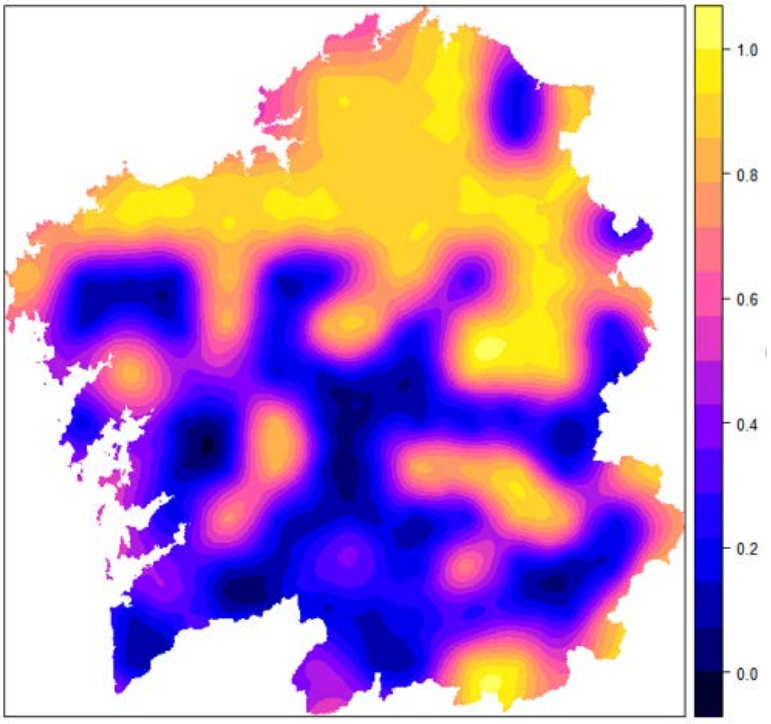

Figura 15. Mapa de Galicia con las estimaciones de la probabilidad de contaminación obtenidas mediante kriging indicador.

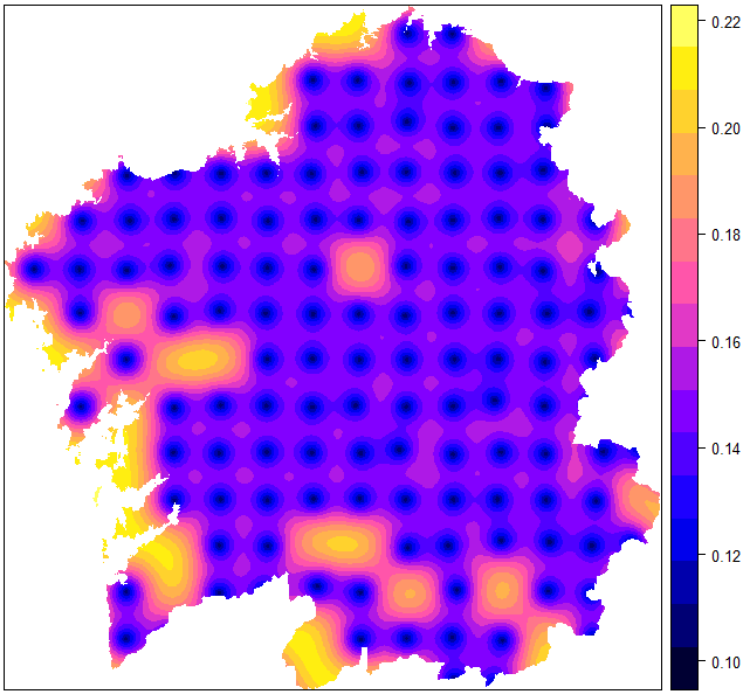

Figura 16. Mapa de Galicia con las varianzas de predicción de las estimaciones de la probabilidad de contaminación obtenidas mediante kriging indicador.

En la Figura 15 se puede observar que en el año 2000, se estima que hay una mayor probabilidad de contaminación por mercurio en la zona norte de Galicia frente a la zona sur, que tiene probabilidades menores. Estos resultados están en consonancia con estudios como el de Real *et al.* (2008) que emplearon aproximaciones cuantitativas para la obtención de conclusiones a partir las mismas observaciones que se emplean en este trabajo.

Capítulo 5

5. Discusión y conclusiones

A pesar de que el protocolo desarrollado parece útil a la hora de obtener predicciones de probabilidad de contaminación, hay distintos puntos del mismo que requieren de un estudio más exhaustivo.

En el análisis de la estimación de los parámetros del modelo de mixtura, si bien se ha estudiado el comportamiento de los estimadores a través del análisis del sesgo y del ECM, así como el calibrado y potencia del contraste de bondad de ajuste final con el que se valida la idoneidad del modelo, también sería interesante realizar un estudio del comportamiento del contraste sobre el número de mixturas. En todo caso, en los datos que nos ocupan, el número de grupos suele estar limitado a 2 y, dado que al realizar el contraste de bondad de ajuste sobre el modelo estimado no se rechaza su validez (siempre y cuando las estimaciones no sean anómalas), las estimaciones del número de componentes se pueden considerar válidas. En este trabajo, los modelos de mixturas se utilizan para asignar las observaciones a los grupos de valores contaminados y no contaminados. Para ello, también se podría recurrir a otro tipo de técnicas basadas en las modas de la densidad, como el clúster modal, aunque habría que tener en cuenta para una correcta aplicación, el carácter dependiente de los datos.

Tal y como se indicó anteriormente, el algoritmo empleado reporta en algunos casos estimaciones anómalas de los parámetros, que hacen que tanto el sesgo como el ECM se incrementen notablemente. Esto no ocurría en el caso de los datos dependientes, de ahí la diferencia en magnitud de los valores obtenidos. Sería conveniente valorar detalladamente a qué son debidas estas estimaciones anómalas y reportar valores de sesgo y ECM considerando el correcto funcionamiento del algoritmo. En todo caso, en este trabajo se ha optado por presentar los resultados completos, aun obteniendo estimaciones anómalas, como nota de precaución para la aplicación en la práctica de los estimadores.

El enfoque adoptado en este trabajo es frecuentista, utilizando tanto técnicas paramétricas como no paramétricas. Otra alternativa a explorar para el análisis de estos datos sería la consideración de modelos jerárquicos en el contexto Bayesiano (véase Schmidt *et al.* 2013).

Es relevante comentar que a pesar de que los resultados mostrados en la Figura 15 parecen concordar con los observados en la bibliografía, las probabilidades fueron obtenidas sin tener

en cuenta las hipótesis de estacionariedad (necesaria para formular el kriging) e isotropía (que simplifica dicha formulación). Estos contrastes se hicieron tal y como proponen Bowman y Crujeiras (2013) para el caso del mercurio en ambos casos se rechazó la hipótesis nula (p-valores de 0.001 y 0 respectivamente). En base a estos resultados, se podría hacer una observación "biológica" de la distribución de las observaciones contaminadas pero la estimación de la probabilidad de contaminación debería tomarse con cautela.

Desde un punto de vista biológico, es interesante destacar que, si bien a lo largo del trabajo se ha hablado en todo momento de observaciones "contaminadas" o "no contaminadas", esto no es necesariamente cierto. Sería más correcto hablar de observaciones "contaminadas con respecto al nivel de contaminación *background*" (nivel de contaminación base) o "no contaminadas con respecto al nivel de contaminación *background*". Esto es así, porque el método que estamos empleando para el estudio se basa en la distribución de las concentraciones de los metales pesados en el musgo, y consideramos como "no contaminadas" las observaciones que se asignan a la primera componente, por ser lo que se considera "nivel base" de contaminación. Pero si este "nivel base" presenta concentraciones muy elevadas, puede ser que las observaciones asignadas a la primera componente estén contaminadas, a pesar de que se distribuyan de forma normal.

En conclusión, el método desarrollado en este trabajo parece tener potencial para el estudio de la contaminación a nivel regional en base a la distribución de las concentraciones de metales pesados en el musgo, pero al mismo tiempo sería necesario la realización de nuevos estudios y pruebas para perfeccionar la técnica.

6. Referencias bibliográficas

Aboal, J. R., Boquete, M. T., Carballeira, A., Casanova, A., Debén, S. y Fernández, J. A. (2017). Quantification of the overall measurement uncertainty associated with the passive moss biomonitoring technique: Sample collection and processing. *Environmental Pollution*, *224*, 235-242. https://doi.org/10.1016/j.envpol.2017.01.084

Aboal, J. R., Fernández, J. A., Boquete, T. y Carballeira, A. (2010). Is it possible to estimate atmospheric deposition of heavy metals by analysis of terrestrial mosses? *Science of The Total Environment*, *408*(24), 6291-6297. https://doi.org/10.1016/j.scitotenv.2010.09.013

Ares, A., Aboal, J. R., Fernández, J. A., Real, C. y Carballeira, A. (2009). Use of the terrestrial moss Pseudoscleropodium purum to detect sources of small scale contamination by PAHs. *Atmospheric Environment*, *43*(34), 5501-5509. https://doi.org/10.1016/J.ATMOSENV.2009.07.005

Benaglia, T., Chauveau, D., Hunter, D. R. y Young, D. S. (2009). Mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6), 1-29. https://doi.org/10.18637/jss.v032.i06

Boquete, M. T., Aboal, J. R., Carballeira, A. y Fernández, J. A. (2017). Do mosses exist outside of Europe? A biomonitoring reflection. *Science of The Total Environment*, *593-594*, 567-570. https://doi.org/10.1016/j.scitotenv.2017.03.196

Boquete, M. T., Fernández, J. A., Aboal, J. R. y Carballeira, A. (2011). Analysis of temporal variability in the concentrations of some elements in the terrestrial moss Pseudoscleropodium purum. *Environmental and Experimental Botany*, *72*(2), 210-216. https://doi.org/10.1016/J.ENVEXPBOT.2011.03.002

Boquete, M. T., Fernández, J. A., Carballeira, A. y Aboal, J. R. (2015). Relationship between trace metal concentrations in the terrestrial moss Pseudoscleropodium purum and in bulk deposition. *Environmental Pollution*, *201*, 1-9. https://doi.org/10.1016/J.ENVPOL.2015.02.028

Bowman, A. W. y Azzalini, A. (2018). R package «sm»: nonparametric smoothing methods (version 2.2-5.6).

Bowman, A. W. y Crujeiras, R. M. (2013). Inference for variograms. *Computational Statistics and Data Analysis*, *66*, 19-31. https://doi.org/10.1016/j.csda.2013.02.027

Cressie, N. A. C. (1993). *Statistics for Spatial Data, Revised Edition* (2.ª ed.). Hoboken, NJ: John Wiley & Sons, Inc. https://doi.org/10.1002/9781119115151

Dempster, A. P., Laird, N. M. y Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm . *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1-22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Diblasi, A. y Bowman, A. W. (1997). Testing for constant variance in a linear model. *Statistics and Probability Letters*, *33*(1), 95-103. https://doi.org/10.1016/S0167-7152(96)00115-0

Fernández, J. A., Aboal, J. R., Real, C. y Carballeira, A. (2007). A new moss biomonitoring method for detecting sources of small scale pollution. *Atmospheric Environment*, *41*(10), 2098-2110. https://doi.org/10.1016/J.ATMOSENV.2006.10.072

Fernández, J. A., Boquete, M. T., Carballeira, A. y Aboal, J. R. (2015). A critical review of protocols for moss biomonitoring of atmospheric deposition: Sampling and sample preparation. *Science of The Total Environment*, *517*, 132-150. https://doi.org/10.1016/J.SCITOTENV.2015.02.050

Harmens, H., Norris, D. A., Steinnes, E., Kubin, E., Piispanen, J., Alber, R., … Zechmeister, H. G. (2010). Mosses as biomonitors of atmospheric heavy metal deposition: Spatial patterns and temporal trends in Europe. *Environmental Pollution*, *158*(10), 3144-3156. https://doi.org/10.1016/j.envpol.2010.06.039

Maechler, M. (2019). nor1mix: Normal aka Gaussian (1-d) Mixture Models (S3 Classes and Methods). R package version 1.3-0.

Markert, B. y Weckert, V. (1989). Fluctuations of element concentrations during the growing season of Polytrichum formosum (Hedw.). *Water, Air, and Soil Pollution*, *43*(1-2), 177-189. https://doi.org/10.1007/BF00175592

Marron, J. S. y Wand, M. P. (1992). Exact Mean Integrated Squared Error. *The Annals of Statistics*, *20*, 712-736. https://doi.org/10.2307/2241980

McLachlan, G. y Peel, D. (2000). *Finite Mixture Models*. Hoboken, NJ: John Wiley & Sons. https://doi.org/10.1002/0471721182

Pavia, J. M. (2015). Testing goodness-of-fit with the kernel density estimator: GoFKernel. *Journal of Statistical Software, Code Snippets*, *66*(1), 1-27. https://doi.org/10.18637/jss.v066.c01

Pebesma, E. J. (2004). Multivariable geostatistics in S: The gstat package. *Computers and Geosciences*, *30*(7), 683-691. https://doi.org/10.1016/j.cageo.2004.03.012

Real, C., Fernández, J. A., Aboal, J. R. y Carballeira, A. (2008). Detection of pulses of atmospheric mercury deposition with extensive surveys and frequently sampled stations: A comparison. *Ecotoxicology and Environmental Safety*, *70*(3), 392-399. https://doi.org/10.1016/j.ecoenv.2008.01.005

Schmidt, A., Hoeting, J., Pereira, J. B. M. y Vieira, P. P. (2013). Mapping malaria in the Amazon rain forest: A spatio-temporal mixture model. En A. O'Hagan & M. West (Eds.), *The Oxford Handbook of Applied Bayesian Analysis* (pp. 90-117). Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198703174.013.5

Varela, Z., Aboal, J. R., Carballeira, A., Real, C. y Fernández, J. A. (2014). Use of a moss biomonitoring method to compile emission inventories for small-scale industries. *Journal of Hazardous Materials*, *275*, 72-78. https://doi.org/10.1016/J.JHAZMAT.2014.04.061

Varela, Z., Fernández, J. A., Real, C., Carballeira, A. y Aboal, J. R. (2015). Influence of the physicochemical characteristics of pollutants on their uptake in moss. *Atmospheric Environment*, *102*, 130-135. https://doi.org/10.1016/J.ATMOSENV.2014.11.061

Índice de figuras

Índice de tablas