Trabajo Fin de Máster

Modelado y visualización con R-INLA de contaminantes atmosféricos y mortalidad por cáncer de pulmón en Galicia

Fikru Meilán Garrido

Máster en Técnicas Estadísticas

Curso 2021-2022

Propuesta de Trabajo Fin de Máster

Título en galego: Modelado e visualización con R - INLA de gradientes de concentración de metais a nivel atmosférico e mortalidade por cancro de pulmón en Galicia.
Título en español: Modelado y visualización con R - INLA de gradientes de concentración de metales a nivel atmosférico y mortalidad por cáncer de pulmón en Galicia.
English title: Modelling and visualization with R - INLA of atmospheric metal concentration gradients and lung cancer mortality in Galicia.
Modalidad: Modalidad A
Autor: Fikru Meilán Garrido, Universidad de Santiago de Compostela
Directores: Rosa María Crujeiras Casais, Universidad de Santiago de Compostela; Jesús Ramón Aboal Viñas, Universidad de Santiago de Compostela
Breve resumen del trabajo: Análisis estadístico y modelado predictivo con R-INLA de datos para la monitorización y control de la calidad del aire (concentración de contaminantes atmosféricos en tejidos de P. purum durante el intervalo 2000 - 2010) y datos de Salud Pública en Galicia (muertes por municipio durante el 2010).

Doña Rosa María Crujeiras Casais, Profesora Titular de Universidad en el Departamento de Estatística, Análisis Matemático y Optimización de la Universidad de Santiago de Compostela, y don Jesús Ramón Aboal Viñas, Catedrático del Área de Ecología de Universidad de Santiago de Compostela, informan que el Trabajo Fin de Máster titulado

Modelado y visualización con R-INLA de contaminantes atmosféricos y mortalidad por cáncer de pulmón en Galicia

fue realizado bajo su dirección por don Fikru Meilán Garrido para el Máster en Técnicas Estadísticas. Estimando que el trabajo está terminado, da su conformidad para su presentación y defensa ante un tribunal.

En Santiago de Compostela, a 13 de junio de 2022.

CRUJEIRAS CASAIS ROSA MARIA - 52931780V

Firmado digitalmente por CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-52931780V, givenName=ROSA MARIA, sn=CRUJEIRAS CASAIS, cn=CRUJEIRAS CASAIS ROSA MARIA - 52931780V
Fecha: 2022.06.12 22:22:18 +02'00'

La directora:
Doña Rosa María Crujeiras Casais

ABOAL VIÑAS JESUS RAMON - 33288734Y

Firmado digitalmente por ABOAL VIÑAS JESUS RAMON - 33288734Y
Nombre de reconocimiento (DN): c=ES, serialNumber=IDCES-33288734Y, givenName=JESUS RAMON, sn=ABOAL VIÑAS, cn=ABOAL VIÑAS JESUS RAMON - 33288734Y
Fecha: 2022.06.13 10:19:25 +02'00'

El director:
Don Jesús Ramón Aboal Viñas

El autor:
Don Fikru Meilán Garrido

Declaración responsable. Para dar cumplimiento a la Ley 3/2022, de 24 de febrero, de convivencia universitaria, referente al plagio en el Trabajo Fin de Máster (Artículo 11, Disposición 2978 del BOE núm. 48 de 2022), **Fikru Meilán Garrido declara** que el Trabajo Fin de Máster presentado es un documento original en el que se han tenido en cuenta las siguientes consideraciones relativas al uso de material de apoyo desarrollado por otros/as autores/as:

- Todas las fuentes usadas para la elaboración de este trabajo han sido citadas convenientemente (libros, artículos, apuntes de profesorado, páginas web, programas,...)
- Cualquier contenido copiado o traducido textualmente se ha puesto entre comillas, citando su procedencia.
- Se ha hecho constar explícitamente cuando un capítulo, sección, demostración,...sea una adaptación casi literal de alguna fuente existente.

Y, acepta que, si se demostrara lo contrario, se le apliquen las medidas disciplinarias que correspondan.

Agradecimientos

Son muchas las personas que han sido bálsamo para el duro proceso que culmina con este documento, y pocas me parecen las líneas que puedo dedicarles agradeciéndoselo.

A los que habéis estado ahí siempre, amigos, familia, gracias por esos tés mañaneros, gracias por esas comidas, paseos y cenas quitándole hierro a mis preocupaciones y haciendo del proceso una experiencia enriquecedora. En especial, gracias a mi "abu" por haberme enseñado la importancia de afrontar los retos con la cabeza alta y a "ma", por la confianza que depositas en mí cada día, y por tu visión de las cosas.

Gracias a todo el grupo de investigación ECOTOX de la Universidad de Santiago de Compostela por compartir conmigo los datos de diez años de trabajo. También agradecerle a Alberto Ruano Raviña, catedrático de Medicina Preventiva y Salud Pública en la Universidad de Santiago de Compostela, las gestiones realizadas con el SERGAS para poder hacer este estudio.

Finalmente, a mis directores Rosa y Jesús. Rosa, gracias por tu pasión docente, sin tu capacidad para trasladar a lenguaje común la "formulitis" matemática no habría sido capaz de ver la belleza de este trabajo; también agradecerte todas y cada una de las correcciones, observaciones y comentarios a este documento. Jesús, muchas gracias por una propuesta de trabajo tan interesante y por haberme abierto los ojos al potencial del Máster en Técnicas Estadísticas.

Índice general

Resumen

Resumen en español

Los gradientes de contaminación atmosférica pueden ser analizados utilizando biomonitores bajo una metodología estandarizada. En este trabajo se utiliza el musgo *P. purum* como biomonitor de la contaminación atmosférica por un metaloide (arsénico) y ocho elementos metálicos (aluminio, cadmio, cobre, cromo, hierro, mercurio, níquel y plomo), en 517 puntos de muestreo distribuidos por toda la superficie gallega y su tejido industrial durante el intervalo 2000-2010. Utilizado estos datos geoestadísticos y el paquete R-INLA, que implementa la aproximación de Laplace basada en integrales anidadas (INLA) y el enfoque de las ecuaciones diferenciales parciales estocásticas (SPDE), se diseñaron mapas de predicción de los gradientes de concentración de los elementos en tejidos de *P. purum*, que indican a su vez el comportamiento de los gradientes de contaminación a nivel atmosférico. Por otro lado, utilizando datos reticulares de mortalidad por cáncer de pulmón en Galicia durante 2010, se diseñaron tanto mapas con las tasas de mortalidad estandarizadas por municipio, como mapas de riesgos relativos de mortalidad. Una vez diseñados todos los mapas, se trató de hacer un análisis cualitativo de asociación entre los mapas de contaminación por metales y los mapas de riesgo relativo de mortalidad.

English abstract

Atmospheric pollution gradients can be analysed using biomonitors under a standardized methodology. In this work, the moss *P. purum* is used as a biomonitor of atmospheric pollution by one metalloid (arsenic) and eight metallic elements (aluminium, cadmium, copper, chromium, iron, mercury, nickel and lead), in 517 sampling points distributed over the whole Galician surface and its industrial fabric during the period 2000-2010. Using these geostatistical data and the R-INLA package, which implements the integrated nested Laplace approximation (INLA) and the stochastic partial differential equations approach (SPDE), prediction maps of the concentration gradients of the elements in *P. purum* tissues were designed. Maps which in turn indicate the behaviour of the pollution gradients at the atmospheric level. On the other hand, using lattice data on lung cancer mortality in Galicia during 2010, both maps with standardized mortality rates by municipality and maps of relative mortality risks were designed. Once all the maps were designed, a qualitative analysis of the association between the metal contamination maps and the relative mortality risk maps was performed.

Prefacio

La célebre frase del filósofo inglés Thomas Hobbes en su obra *El Leviatán*, *"homo homini lupus"* o *el hombre es un lobo para el hombre*, podría ser reinterpretada en la actualidad como una metáfora acertada del ciclo de interacciones entre algunas actividades antrópicas y el medio ambiente.

En el presente Trabajo Fin de Máster (TFM), se comienza a explorar si podemos estar siendo testigos de un ejemplo más de dicho ciclo de interacciones. Para ello, se desarrollará una metodología destinada a modelar y mapear distintos tipos de datos espaciales y se aplicará la metodología a datos de contaminantes atmosféricos y de Salud Pública en Galicia.

Para llevar a cabo el estudio, nos centraremos en dos de los cuatro subsistemas de la Tierra. Por una parte, la atmósfera y por otra la biosfera. Partiendo de la perspectiva de que la atmósfera es un bien común indispensable para la vida, respecto del cual, todas las personas tienen el derecho de su uso y la obligación de su conservación[1], y teniendo en cuenta que, según la Agencia Internacional para la Investigación del Cáncer (IARC), la contaminación atmosférica es un agente cancerígeno (del grupo I) para los humanos[2], resulta acuciante estudiar el origen, así como modelar el comportamiento y distribución de dicha contaminación.

Por una parte, cabe mencionar que la gama de contaminantes atmosféricos es muy diversa tanto en su composición como en su origen. La composición y origen de los contaminantes son factores esenciales a la hora de diseñar planes específicos para la prevención o reducción de sus riesgos. En este TFM nos centraremos en la contaminación atmosférica por metales y metaloides emitidos mayoritariamente por el tejido industrial de la comunidad de Galicia.

Por otro lado, destacar que para el modelado de un determinado riesgo, se pueden incluir variables con las que presente asociación. Como en este estudio se están dando los primeros pasos para ver si hay indicios de asociación entre los contaminantes considerados y la mortalidad por cáncer de pulmón en Galicia, no conviene incluir ninguno de los elementos analizados para la estimación de los riesgos de mortalidad. En su lugar, se utilizará como variable de interés la concentración media por municipio de radón residencial. El radón es un gas radiactivo incoloro, inodoro e insípido que se produce por desintegración radiactiva natural del uranio presente mayormente en suelos y rocas y que se asocia con la incidencia de cáncer de pulmón en Galicia [3, 22].

Finalmente, indicar que la herramienta central para el desarrollo de este estudio será el software R [26] y las librerías:

- `cowplot`: complementa las funciones gráficas de `ggplot()`. En este documento se utiliza la función `plot_grid()` para combinar en una única figura varios gráficos [32].

[1] Contemplado en la Ley 34/2007 (2007) sobre la calidad del aire y protección de la atmósfera, por la cual todos tenemos unos derechos y obligaciones para con la atmósfera. Para más información consultar https://www.boe.es/buscar/pdf/2007/BOE-A-2007-19744-consolidado.pdf.

[2] IARC: International Agency for Research on Cancer - WHO. En el 2016 se incluyó a la contaminación atmosférica como un agente cancerígeno para los humanos. Consultado en https://monographs.iarc.fr/list-of-classifications.

- `dplyr`: permite la manipulación de los conjuntos de datos usando un lenguaje alternativo a la sintaxis habitual de R [29, 26].
- `ggplot2`: útil para la visualización y representación gráfica de los datos [30].
- `gridExtra`: de forma similar al paquete `cowplot`, permite la manipulación y fusión de gráficos con funciones como `grid.arrange()` [2].
- `INLA`: entre otros aspectos, implementa múltiples funciones para realizar inferencia bayesiana aproximada para campos gaussianos latentes [17].
- `lwgeom`: librería dependiente de `sf` y `tidyverse` que se puede utilizar para la lectura, filtrado y visualización de datos vectoriales geoespaciales en R [24, 25, 26, 28].
- `mapSpain`: de sus múltiples aplicaciones, la principal utilidad de esta librería consiste en facilitar la creación de mapas de los diferentes niveles administrativos de España [10].
- `moments`: librería que implementa funciones para resolver contrastes de normalidad basados en los momentos (como el contraste de hipótesis de normalidad basado en el test de D' Agostino) [13].
- `normtest`: implementa funciones para resolver contrastes de hipótesis basadas en el métodos de los momentos (ejemplo del contraste de normalidad basado en el estadístico de Jarque-Bera) [8].
- `nortest`: implementa diversas funciones para realizar contrastes de normalidad como los que se comentarán en el análisis descriptivo de los datos del primer capítulo de este documento (contrastes basados en el estadístico de Anderson-Darling, en el estadístico de Cramer-von Mises o el basado en el estadístico de Kolmogorov-Smirnov-Lilliefors) [9].
- `raster`: resulta de utilidad para el manejo de datos raster (matriz de celdas en las que cada celda representa información de fenómenos del mundo real, como pueden ser los niveles de un determinado contaminante) [11].
- `sf`: librería para trabajar con datos espaciales acorde al código estándar abierto que utilizan los programas de los sistemas de información geográfica (puntos, líneas, polígonos, *etc.*) Es una librería que a diferencia de `sp` carece de soporte para datos raster [23, 25].
- `sp`: librería para la manipulación de datos espaciales [23].
- `spdep`: consta de un conjunto de funciones para la construcción de matrices de pesos para polígonos contiguos (véase capítulo 4) [4].
- `stringr`: paquete con funciones útiles para el procesamiento y manipulación de cadenas de forma simple [31].
- `tidyverse`: librería que implementa funciones para el tratamiento de datos tanto espaciales como no espaciales [28].
- `viridis`: paquete con funciones que permiten implementar gradientes de coloración en representacines gráficas. Es muy útil a la hora de representar fenómenos naturales como son las concentraciones de un determinado contaminante en una región de estudio [7].

Capítulo 1

Introducción

En este capítulo se comenzará por presentar la motivación y los objetivos del estudio. A continuación, se expondrá el origen, naturaleza y estructura de los distintos conjuntos de datos empleados. Finalmente, se procederá con el análisis descriptivo de los dos principales conjuntos de datos (los de contaminación atmosférica y los de mortalidad por cáncer de pulmón), prestando especial atención a las particularidades de cada conjunto.

1.1. Motivación del trabajo

La biomonitorización es generalmente descrita como el uso sistemático de organismos vivos o sus respuestas, para determinar el estado o los cambios del medio ambiente [16]. Son cientos los estudios en los que se utilizan como biomonitores los musgos terrestres, ya sea para comprobar su fiabilidad, para analizar las tendencias espaciales y temporales de los contaminantes atmosféricos o para estudiar la asociación de dicha contaminación con afecciones sanitarias [1, 5, 6, 15, 27, 33]. Los motivos para la utilización de los musgos como biomonitores son claros:

- La determinación de la concentración de un contaminante en un organismo proporciona una medida de la exposición a la que ha estado sometido.
- Los organismos retienen la contaminación absorbida, revelando así eventos de contaminación episódicos en un ambiente dinámico como es la atmósfera.
- Existen especies de musgos con distribución cosmopolita, lo que permite diseñar estudios que abarcan áreas geográficas extensas, así como la comparación de zonas muy distantes.
- Son organismos sésiles, por lo tanto fáciles de recolectar y reflejan la contaminación de su punto de muestreo.
- Carecen de raíces, por lo que es mínima la bioacumulación de contaminantes procedentes del suelo. En su lugar, absorben el agua y los nutrientes que llegan a su superficie a través de la deposición.

Partiendo de estas premisas generales se podría pensar que la aplicación de técnicas de biomonitorización de la deposición de contaminantes atmosféricos empleando musgos, popularmente conocida como “moss bag technique”, es un procedimiento sencillo y altamente eficiente a la hora de obtener datos. Nada más lejos de la realidad, es un procedimiento que puede arrojar resultados con una variabilidad muy alta entre estudios, fruto de variaciones en la metodología de muestreo, especies consideradas para el muestreo, vegetación circundante, variabilidad temporal, etc. De todos modos, la mayor parte de la variabilidad debida a estas causas se puede minimizar estandarizando las metodologías de

muestreo[1].

Existe una comprobación sistemática en estudios epidemiológicos de la asociación entre concentraciones de materia particulada inhalable (PM_{10}) y la morbilidad o la mortalidad. Al margen del tamaño de las partículas, sus efectos sobre la salud dependen de su composición química. Dada la amplia gama de contaminantes atmosféricos, todavía se están investigando los efectos provocados por las diferentes especies químicas; la mayoría de los estudios apuntan que el mayor impacto en la salud viene causado por las partículas de carbono elemental (CE), compuestos orgánicos (CO), especialmente hidrocarburos aromáticos policíclicos (HAPs), sulfatos, nitratos y determinados metaloides y metales como arsénico (As), cadmio (Cd), hierro (Fe), zinc (Zn), cromo (Cr), cobre (Cu), aluminio (Al) , níquel (Ni) y plomo (Pb) [18].

A pesar de la amplia bibliografía existente sobre la biomonitorización de la contaminación atmosférica con musgos terrestres, así como sobre estudios epidemiológicos de los efectos de la contaminación atmosférica por PM_{10} y metales asociados, son muy pocos los estudios en los que se han fusionado ambos campos; de hecho, tras una búsqueda exhaustiva, sólo se han identificado tres publicaciones distribuidas entre Estados Unidos, Francia y Países Bajos [5, 15, 33].

1.1.1. Objetivos y organización del documento

Para entender los objetivos de este estudio conviene estar familiarizado con los tipos de datos espaciales y su representación (véase sección 2.1).

Destacar que la motivación central de este trabajo es, llevar a cabo un estudio de carácter interdisciplinar en el que se fusionen la monitorización y control de la calidad del aire con Salud Pública. Por un lado, se busca modelar datos geoespaciales para obtener mapas de concentración de elementos metálicos tóxicos en Galicia durante el intervalo 2000-2010. Por otra parte, se pretenden usar datos de área de mortalidad por cáncer de pulmón, para modelar y mapear la mortalidad por cáncer de pulmón en Galicia durante el 2010. Finalmente, se pretenden utilizar los mapas predictivos para hacer un análisis cualitativo de asociación entre los contaminantes y el riesgo de mortalidad.

Conviene resaltar que, mientras los datos para generar el mapa de concentración de contaminantes atmosféricos engloba datos del 2000-2010, los mapas de mortalidad se generarán en base a los datos exclusivamente del 2010. Esto se debe a que el cáncer de "origen ambiental" es una condición asociada generalmente a una exposición prolongada al detonante.

Llegado a este punto, uno estará familiarizado con la motivación y los objetivos del estudio. Por tanto, ya se puede exponer la estructura que va a presentar el documento para alcanzar los objetivos establecidos. En este primer capítulo, nos familiarizaremos con la naturaleza de los datos y su obtención; a continuación, se hará un análisis descriptivo de los mismos utilizando nociones de estadística espacial desarrolladas en la sección 2.1. En el capítulo 2, tras hacer una breve introducción de conceptos generales necesarios para el estudio, se exponen las bases estadísticas de las dos etapas necesarias para la obtención de los mapas predictivos; primero el análisis estructural y a continuación, la predicción por inferencia bayesiana utilizando la aproximación de Laplace basada en integrales anidadas (INLA) y el enfoque de las ecuaciones diferenciales parciales estocásticas (SPDE). A continuación, se procederá con los capítulos 3 y 4 de forma más práctica (casi a modo de tutorial). Mientras que la primera parte del capítulo 3 trata de familiarizar al lector con las librerías y funciones de R [26] utilizadas para el modelado y representación de los datos geoestadísticos, la segunda parte consiste en la aplicación de la metodología a datos reales de contaminación atmosférica en Galicia. El capítulo 4 consta de la misma estructura que el 3, pero en este caso, la aplicación es sobre datos de mortalidad por cáncer de pulmón en Galicia. En el capítulo 5 se comentarán los resultados obtenidos junto al análisis cualitativo de asociación entre los contaminantes y los riesgos de mortalidad. También se expondrán las conclusiones,

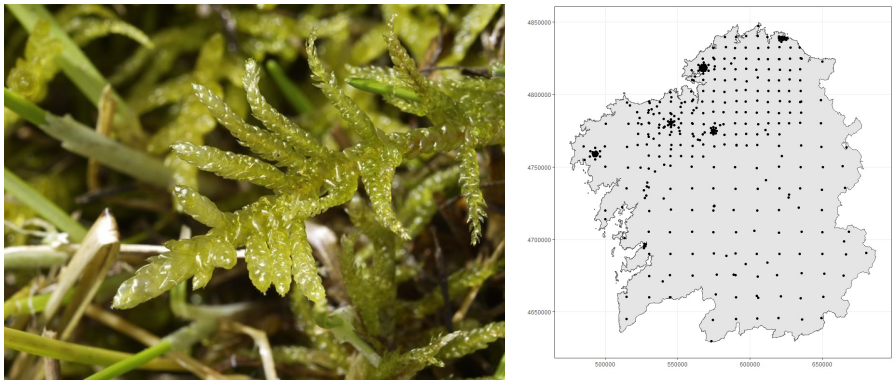

Figura 1.1: A la izquierda un ejemplar de *P. purum* [19]. A la derecha los 517 puntos de biomonitorización de la contaminación atmosférica por metales con *P. purum* y empleando una proyección UTM ED50 zona 29 N.

problemáticas y posibles mejorías del estudio. Finalmente, habrá un apéndice con figuras adicionales que complementan el cuerpo del texto y otro apéndice con la bibliografía empleada.

1.2. Obtención de los datos

En el presente estudio se usan datos procedentes del "Instituto Galego de Estatística" (IGE), del "Servizo Galego de Súde" (SERGAS) y del grupo de Ecotoxicología y Ecofisiología Vgetal de la USC.

- Del IGE se extrajeron datos reticulares: tamaño poblacional por municipio gallego durante el 2010 con un código único para cada municipio [12].
- El SERGAS suministró datos reticulares: muertes por cáncer de pulmón en municipios de Galicia con más de 10000 habitantes durante el 2010.
- Se generó una base de datos geoestadísticos a partir del pull de muestreos que se realizaron durante la primera década del siglo XXI por el grupo de Ecotoxicología y Ecofisiología Vegetal de la USC. Se comprobó que todos los datos utilizados procediesen de una metodología estandarizada[1].

Dicha metodología estandarizada implica que el muestreo fue realizado empleando exclusivamente *Pseudoscleropodium purum* (Hedw.) M. Fleisch (denominado de aquí en adelante como *P. purum*), musgo que se observa en la Figura 1.1. Utilizando *P. purum* como biomonitor se diseñó una red regular con puntos de muestreo eqiespaciados cada 15Km. Esta red equiespaciada fue muestreada de forma sistemática (cada 6 meses), durante el intervalo 2004-2010, también hay medidas simples durante los años 2000 y 2002.

Las muestras se recogieron a más de 300m de carreteras principales, siempre que fuera posible, y a más de 100m de otro tipo de carreteras y casas aisladas. También se trató de recolectar en zonas abiertas, y de no ser posible, en claros de zonas arboladas. En el caso de que dichas estaciones no cumpliesen alguno de estos requisitos o de que no existiese musgo en alguna de ellas, se permitió un pequeño alejamiento del punto teórico, siempre que la nueva estación estuviese lo suficientemente alejada de las estaciones contiguas. En cada estación de muestreo, el área de recogida de musgo no fue mayor de $50m^2$ ni menor de $35m^2$. En cada una de ellas se recolectaban, como mínimo 30 submuestras lo mas separadas posibles por todo el área, evitándose la recogida concentrada en una zona. Las muestras se

transportaron diariamente al laboratorio y se almacenaron en una cámara a temperatura y humedad constantes (T^a= 20^{o}C y H.R= 70 %). Tras un procesado estandarizado de los musgos, se obtuvieron los datos geoestadísticos.

A continuación, se agregaron más datos geoestadísticos procedentes de las inmediaciones de las empresas que constaban en el Registro Estatal de Emisiones y Fuentes de Contaminantes Terrestres (PRTR) en Galicia durante el intervalo 2000-2010. Además, también se muestrearon un conjunto de cogeneraciones con emisiones a nivel atmosférico. En este caso el muestreo no es regular. Finalmente, comentar que para el muestreo de las centrales térmicas de As Pontes, Meirama y Sabón, se diseñó otra red regular que abarcase su posible área de efecto, con puntos de muestreo equiespaciados 7Km; esto se debe a que en el caso de las térmicas, los contaminantes son inyectados a la atmósfera a gran altura, por lo que a diferencia del resto de industrias, los contaminantes tienen una dispersión potencialmente superior. En la Figura 1.1 se muestran los puntos de muestreo descritos.

1.3. Análisis descriptivo de los datos

Los datos pueden ser divididos en dos bloques atendiendo a su naturaleza. Por una parte, los datos geoestadísticos de concentración de elementos metálicos en tejidos de *P. purum* y por otro lado, los datos reticulares de mortalidad por cáncer de pulmón.

1.3.1. Datos de contaminantes

Se parte de un conjunto de datos georreferenciados y etiquetados según formen parte del muestreo de una industria o de la red regular de muestreo. Las coordenadas de los datos constan de dos componentes $s = (s_x, s_y)$, acordes a las coordenadas proyectadas UTM ED50 zona 29N. Destacar que cada muestreo, además de estar georreferenciado, también consta de un identificador de su fecha de recolección. Las variables georreferenciadas de interés son las concentraciones medidas en $\mu g/g$ o en ng/g de metaloides (ng/g As) y metales ($\mu g/g$ Al, ng/g Cd, $\mu g/g$ Cr, $\mu g/g$ Cu, $\mu g/g$ Fe, ng/g Hg, $\mu g/g$ Ni y ng/g Pb). Una nomenclatura habitual para los $\mu g/g$ y los ng/g suelen ser las respectivas partes por millón (PPM) y partes por billón (PPB).

Como se indicará posteriormente, no todos los elementos han sido muestreados en la totalidad de estaciones de muestreo, por lo tanto, debe evitarse el enfoque multivariante para los contrastes que se quieran plantear. Un contraste habitual en estos escenarios es el de bondad de ajuste a un modelo paramétrico, más concretamente, el contraste de bondad de ajuste a una normal:

$$\begin{cases} H_0 : X \text{ se distribuye de acuerdo a una } N(\mu, \sigma), \text{con } \mu \text{ y } \sigma \text{ desconocidas} \\ H_1 : X \text{ no se distribuye de acuerdo a una } N(\mu, \sigma) \end{cases} \tag{1.1}$$

donde X es una variable aleatoria (v.a.). En este caso, X representa la concentración del elemento metálico para el que se quiere comprobar la hipótesis nula de normalidad. Para resolver el contraste (1.1) para los distintos elementos, se emplearon tanto pruebas analíticas como gráficas.

Las pruebas analíticas se basaron en los estadísticos de contraste de Kolmogorov-Smirnov-Lilliefors, Cramér-von Mises, Anderson-Darling, Shapiro-Wilk, Jarque-Bera, D' Agostino-Pearson y pruebas adicionales de asimetría y kurtosis. Con independencia del método empleado y como se puede comprobar gráficamente en la Figura 1.2, para todos los elementos considerados en el estudio, se rechaza la hipótesis nula de normalidad para un nivel de significación $\alpha = 0,01$. Es decir, no se puede asumir que la distribución de las concentraciones de los elementos metálicos de las muestras siga una distribución

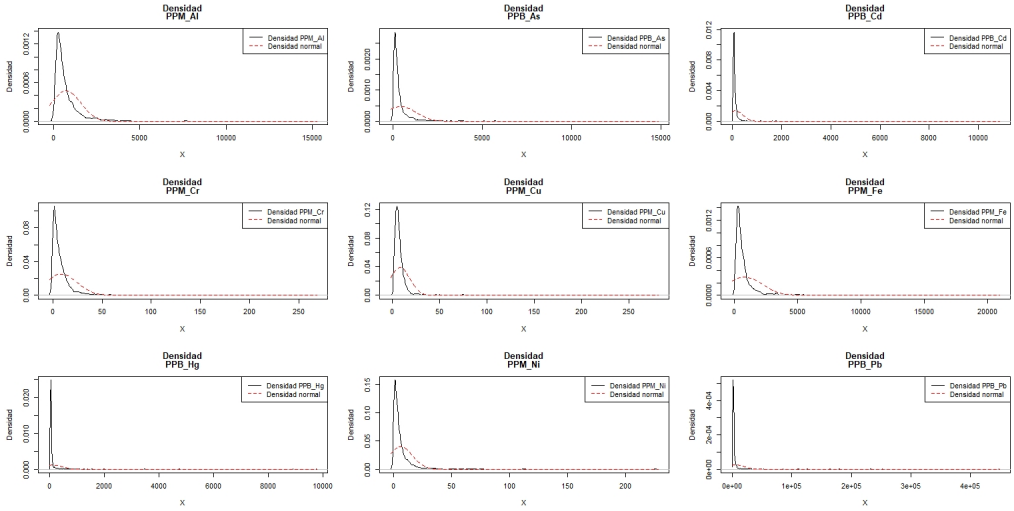

Figura 1.2: Representación de la estimación tipo núcleo de la densidad muestral con selector de ventana por la "regla del pulgar" (en negro), frente a la densidad teórica de una normal (en rojo punteado). La densidad teórica fue obtenida con la media y la desviación típica de las muestras de cada elemento. De izquierda superior a derecha inferior, se tienen las densidades para el Al, As, Cd, Cr, Cu, Fe, Hg, Ni y Pb. Se representan en distinta escala porque los elementos han sido medidos en distinta escala dados sus niveles de emisión y toxicidad.

normal.

Tras aplicar una transformación logarítmica para corregir la asimetría de los datos, se comprobó que no se producían cambios relevantes en los resultados del contraste (1.1). A excepción del resultado obtenido para el Al y el As con el estadístico de contraste basado en asimetría D' Agostino-Pearson, se sigue rechazando, con un 99 % de confianza, la hipótesis nula de normalidad, con independencia del estadístico empleado o elemento analizado. Aunque empleando métodos analíticos se rechazaría la hipótesis nula de normalidad, viendo la representación de la densidad empírica frente a la teórica de los datos log-transformados (Figura 1.3), no resulta tan evidente dicho rechazo, especialmente para el Al, el As y el Cu. Dicho rechazo es más evidente en el gráfico cuantil-cuantil (Q-Q) (Figura 1.4).

Por otra parte, tras aplicar la transformación logarítmica, se puede ver (Figura 1.5), que la distribución de la concentración de la mayoría de los elementos es simétrica en torno a la mediana, a excepción de la dispersión para el Cr y el Ni. También se puede apreciar que, los datos atípicos están presentes para todos los elementos, siendo más frecuentes los datos atípicos por concentraciones elevadas. Estos datos atípicos probablemente se correspondan a las inmediaciones de alguna de las industrias.

Como se puede ver en las Figura 1.6, no hay ningún elemento medido en la totalidad de puntos de muestreo. Del total de 2089 muestreos, sólo en 1150 se determinaron las concentraciones de los nueve elementos metálicos considerados en este estudio. La totalidad de muestreos han sido tomados en 517 puntos de muestreo únicos repartidos por toda la superficie gallega, y en únicamente 277 de esos puntos se han tomado medidas de todos los elementos metálicos.

En la Figura 1.6, se pueden apreciar las concentraciones en escala logarítmica de cada uno de los elementos en los puntos de muestreo de la Figura 1.1. Además de lo comentado previamente, destacar que hay un patrón generalizado de mayor concentración de los elementos metálicos analizados en los

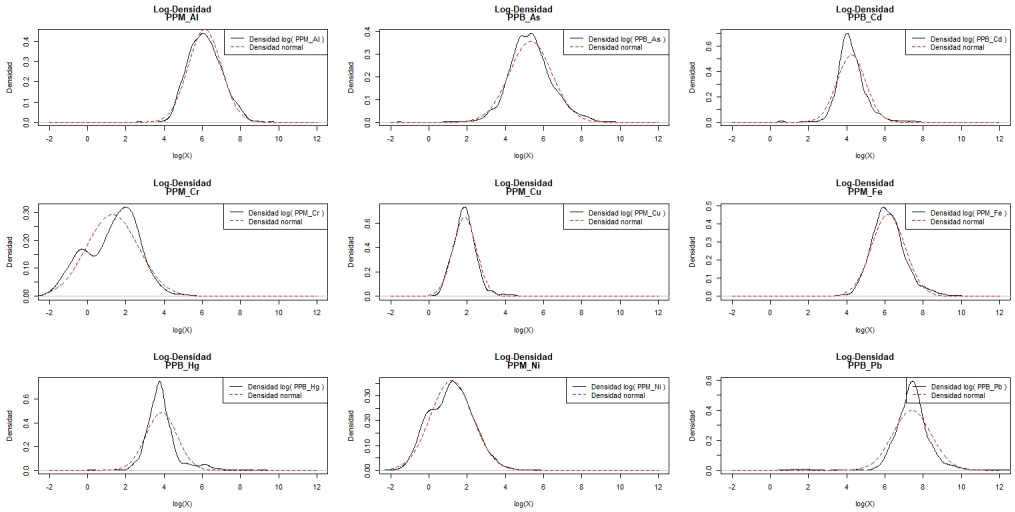

Figura 1.3: Representación de la estimación tipo núcleo de la densidad muestral con selector de ventana por la "regla del pulgar" (en negro), frente a la densidad teórica de una normal (en rojo punteado) de los datos log-transformados. La densidad teórica fue obtenida con la media y la desviación típica de las muestras log-transformadas de cada elemento. De izquierda superior a derecha inferior, se tienen las densidades para el Al, As, Cd, Cr, Cu, Fe, Hg, Ni y Pb.

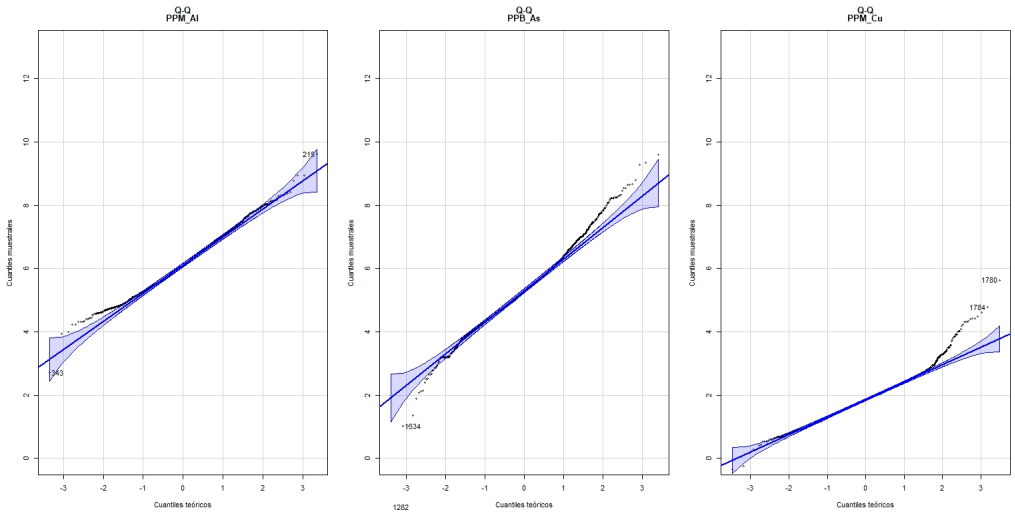

Figura 1.4: Gráficos Q-Q de normalidad de las log-concentraciones de Al, As y Cu.

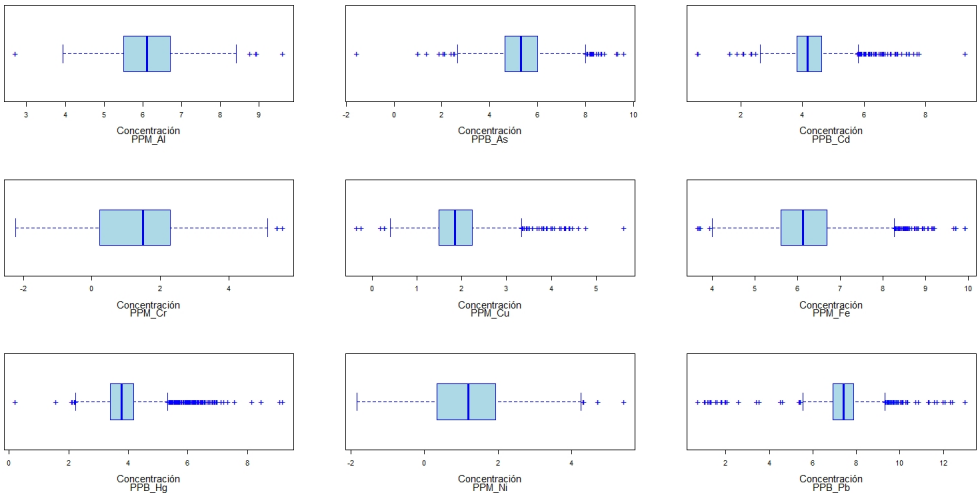

Figura 1.5: Gráficos de cajas de log-concentraciones de Al, As, Cd, Cr, Cu, Fe, Hg, Ni y Pb.

puntos de muestreo próximos o en las inmediaciones de industrias, especialmente claro en el caso del Co, Cr, Fe, Ni y Pb. Dicho patrón no es tan evidente a simple vista para los restantes elementos.

1.3.2. Datos de mortalidad

Para el análisis de los datos de mortalidad se emplean dos conjuntos de datos, uno procedente del SERGAS y otro del IGE.

Tanto el conjunto de datos procedente del SERGAS como el del IGE contienen datos reticulares agregados por municipios. Ambos comparten un código de cinco dígitos para la nomenclatura de las particiones del área de estudio. El área de estudio es Galicia y las particiones son los municipios. Los dos primeros dígitos del código son identificadores de la provincia y los restantes tres son los identificadores del municipio dentro de la provincia.

El SERGAS suministró una base de datos reticular de los municipios con más de 10000 habitantes, donde se registraban todos los eventos de defunción, durante el 2010, por un conjunto de afecciones que se enmarcan dentro del cáncer de pulmón o de las vías respiratorias (tumor maligno del bronquio principal, tumor maligno del lóbulo superior del bronquio o del pulmón, tumor maligno del lóbulo medio del bronquio o del pulmón, tumor maligno del lóbulo inferior del bronquio o del pulmón, tumor maligno de sitios contiguos de los bronquios y del pulmón, o tumor maligno de los bronquios o del pulmón sin una parte especificada). En dicha base de datos reticular, también se registra el sexo del individuo difunto.

Del IGE se extrajeron el número total de habitantes por municipio y el número de habitantes por municipio desglosado por su sexo durante el 2010 [12].

Según el registro del SERGAS, durante el 2010, se contabilizaron un total de 1267 muertes, de las cuales 1041 eran hombres y 226 eran mujeres. Resulta llamativa la diferencia en proporción del número de muertos segregado por su sexo. Del total de 317 municipios en Galicia, solamente 59 tienen al menos 10000 habitantes. Por lo tanto, el máximo número de municipios que podrían constar de datos de mortalidad por cáncer de pulmón durante el 2010 serían de 59. La realidad es que en 9 de

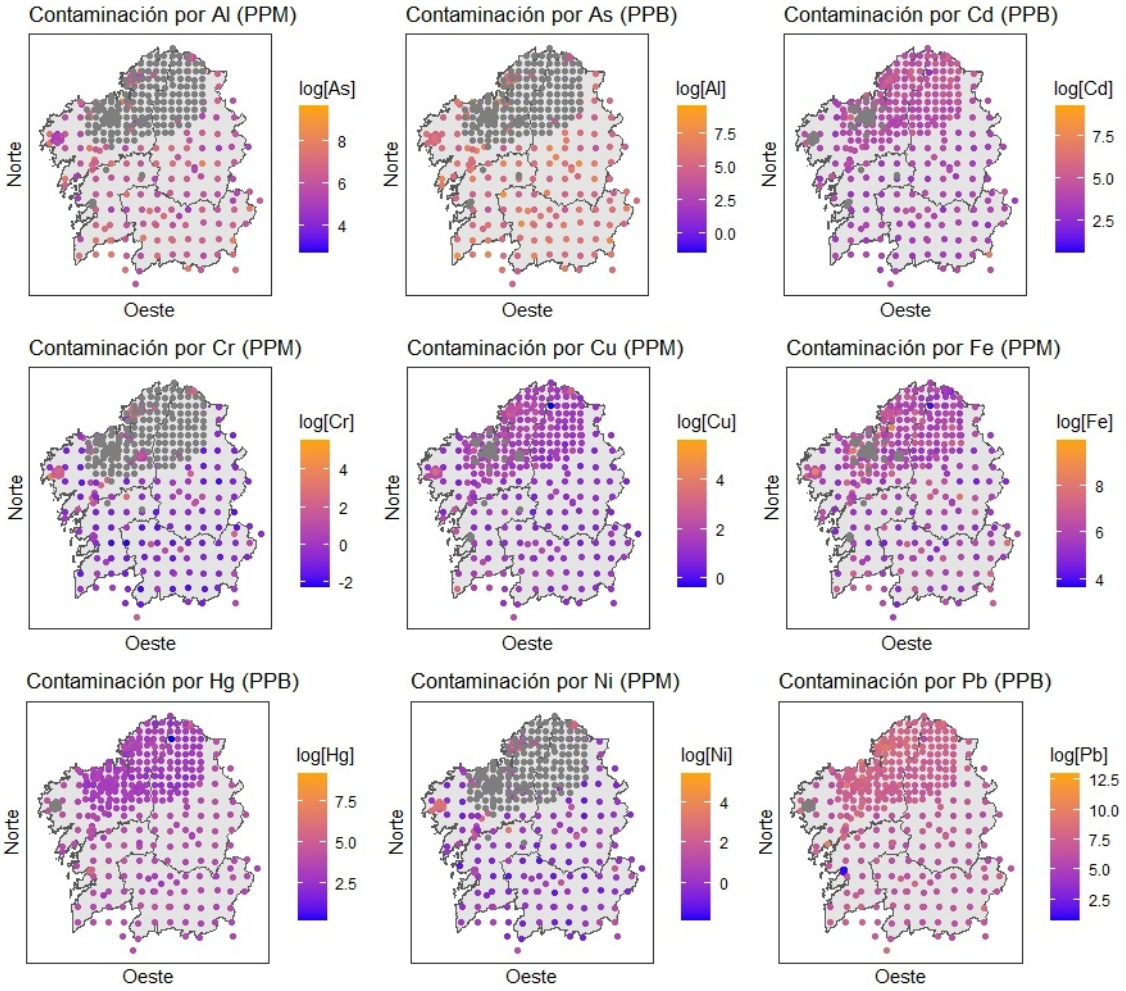

Figura 1.6: Concentraciones de los contaminantes en escala logarítmica en cada estación de muestreo. Gradiente de concentración con código de colores de azul a naranja a medida que se incrementa la concentración del elemento; en gris datos faltantes en los puntos de muestreo. Se emplea una escala distinta para cada elemento, porque estos han sido medidos en distinta escala debido a sus niveles de emisión y toxicidad.

esos municipios, no se contabilizó ninguna muerte por cáncer de pulmón durante el 2010 (únicamente hay datos de 50 municipios). Si estos datos de mortalidad se estratifican por sexos se vería que este número es aún menor, siendo de 32 y 49 respectivamente con mujeres y hombres.

También conviene destacar que la media de muertes por municipio con el total de la población (para los municipios de los que se consta de datos), está en torno a 25 muertes por municipio, con una desviación típica de aproximadamente 58 muertes por municipio. Esta alta dispersión en los muertos por municipio era esperable, dado que el número de muertos varia notablemente en función del tamaño poblacional del municipio.

Para poder abordar plenamente el resto del análisis descriptivo de los datos de área conviene estar familiarizado con algunos conceptos y notación de estadística espacial (véase sección 2.1.3).

En la Figura 1.7 se pueden apreciar las tasas estandarizadas de mortalidad (SMRs) para los tres escenarios (población total y estratificada por sexos). Se trata de las SMR en (2.2) estimadas durante el 2010 en base a los datos del SERGAS y del IGE. Como se puede apreciar, son realmente pocos los municipios de los que se tiene información sobre la SMR estimada (en torno al 16 % de los municipios en el escenario con la población sin estratificar).

Analizando la Figura 1.7 uno se percata rápidamente que las SMR obtenidas para la población total y para la población exclusivamente masculina son prácticamente idénticas, lo cual es lógico teniendo en cuenta que la gran mayoría de muertes por cáncer de pulmón se produjeron en hombres. También se puede ver que, tanto en el escenario con toda la población, como con el que tiene en cuenta sólo a la población masculina, únicamente Ferrol, A Coruña, Santiago de Compostela, Lugo, Ribeira, Pontevedra, Ourense, Monforte de Lemos, Vigo y Verín tienen un SMR > 1, lo que estaría indicando que para estos municipios la mortalidad por cáncer de pulmón es superior a la esperada. En el escenario que tiene en cuenta sólo a la población femenina, los municipios con un SMR > 1 serían A Guardia y los mismos que antes, menos Lugo y Monforte de Lemos. Aunque no hay ningún municipio con SMR $= 1$, lo que estaría indicando que las muertes observadas en el municipio son iguales a las esperadas para ese municipio, hay algunos que están cerca.

Destacar que las SMRs más altas se han producido para algunas de las regiones en las que intuitivamente uno esperaría encontrar tasas de mortalidad elevadas por afecciones asociadas a contaminación atmosférica, ejemplo de los municipios de Ferrol y Coruña, regiones con una alta carga de emisiones a nivel atmosférico dentro de Galicia. Resulta llamativo que una de las industrias más contaminantes analizadas por el grupo de Ecotoxicología y Ecofisiología Vegetal de la USC (Megasa Siderurgias S.L.), se localiza próximo al municipio con mayor SMR (Ferrol).

La empresa Megasa es una siderurgia ubicada en el municipio de Narón, a ella están asociadas altas emisiones producto de la fundición de chatarra (Cu, Fe) y pinturas de la chatarra (As, Cd, Cr, Co, Hg y Pb). En los análisis realizados, los valores de concentración más elevados se daban sobre la fábrica, donde la concentración de Cd, Cu, Hg y Pb mostraba una clara estructura espacial con anisotropía dependiente de la orografía y vientos. Teniendo en cuenta esto y lo comentado previamente se ha decidido prescindir de los métodos clásicos de kriging para el modelado predictivo tanto de la contaminación como de los riesgos de mortalidad por cáncer de pulmón. Es decir, partiendo de la observación hecha (existencia de procesos subyacentes anisotrópicos), uno podría plantearse no utilizar predictores kriging, ya que estos asumen generalmente normalidad e isotropía. De todos modos no sería motivo suficiente, dado que existen correcciones de los predictores para escenarios con anisotropía. Si además, se le suma lo destacado en la obtención de los datos (que se va a intentar integrar información de distintas fuentes y con datos faltantes), resulta problemático utiliza los predictores kriging.

A pesar de todo, el punto clave para rechazar el uso de los predictores kriging para el modelado

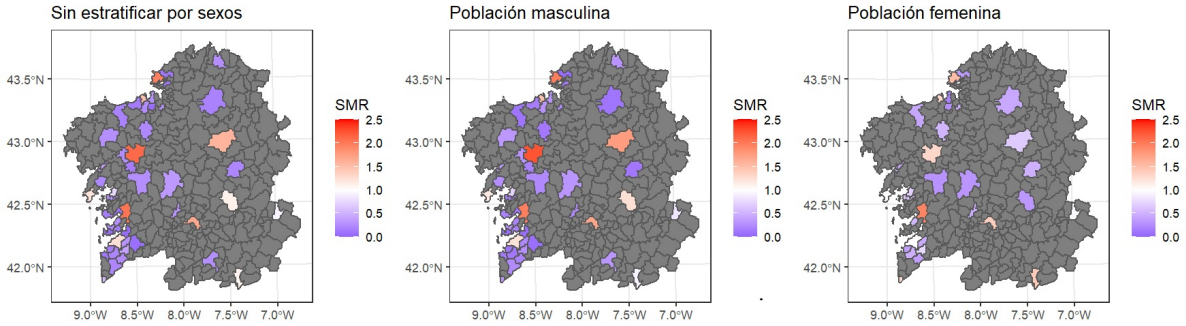

Figura 1.7: Representación de las tasas estandarizadas de mortalidad durante el 2010 en base a los datos del Sergas y el INE. Se utiliza una escala de coloración (rojo para las tasas altas, azul para las tasas bajas, blanco para las tasas neutras). En gris oscuro aparecen los municipios para los que no hay datos.A la izquierda los resultados utilizando la población total, en el medio sólo con la población masculina y a la derecha sólo la población femenina.

de los datos geoestadísticos y reticulares previamente mencionados reside en que no sólo estamos interesados en realizar una interpolación sin más, si no que estamos interesados en meter un modelo para inferir parámetros en los procesos estudiados.

Capítulo 2

Introducción a la geoestadística

El término "Geoestadística" fue acuñado en los años 50 por el matemático George Matheron, para hacer referencia a las técnicas estadísticas aplicadas al análisis geográfico. Fue concebida por su potencial aplicación para predecir las reservas de mineral útil a partir de observaciones espacialmente distribuidas en una región. Aunque en la actualidad se emplea la geoestadística para resolver una gran variedad de problemas, todos ellos se engloban bajo la Teoría de las Variables Regionalizadas, por la cual, los datos se pueden ver como una realización, habitualmente parcial, de un proceso estocástico sobre una región espacial continua [20].

Como se explicó en el primer capítulo, uno de los objetivos centrales de este estudio es generar mapas con gradientes continuos de concentración de la contaminación o tasas de mortalidad. Para generar estos mapas predictivos, tendremos que hacerlo en dos etapas. La primera será el análisis estructural, en el cual se describe la correlación entre puntos en el espacio. En la segunda etapa, se harán las predicciones en sitios de la región no muestreados por medio de la aproximación de Laplace basada en integrales anidadas (INLA) y mediante el enfoque de las ecuaciones diferenciales parciales estocásticas (SPDE) [21].

2.1. Algunos conceptos previos

En esta sección se van a exponer algunos conceptos y términos transversales a todo el documento.

2.1.1. Tipos de datos espaciales

Cuando se trabaja en $d = 2$ dimensiones, un proceso espacial se indica como:

$$\left\{X(s) : s \in D \subset R^2\right\}, \tag{2.1}$$

siendo X la variable observada, por ejemplo el número de muertes por cáncer de pulmón o la concentración de un contaminante, y s hace referencia a la localización de la variable observada. Generalmente, los datos espaciales se suelen agrupar en tres clases según las características del dominio D [21]:

- Datos de área: también conocidos como datos reticulares o de agregado. Son observaciones de una variable X, que se asocian a un conjunto numerable de polígonos o áreas con unas fronteras bien definidas dentro de la región de observación D (que puede tener una forma regular o irregular).
- Datos geoestadísticos: situación en la que el dominio D es un conjunto fijo continuo. Que sea continuo quiere decir que s varía de forma continua sobre D y que por tanto, $X(s)$ puede observarse en cualquier punto dentro de D. Con fijo se quiere indicar que los puntos en D no son aleatorios.

- Procesos puntuales: a diferencia de los datos geoestadísticos o de área, el dominio D en los procesos puntales es aleatorio. Los conjuntos de datos pertenecen a eventos distribuidos aleatoriamente. Un ejemplo perfecto serían las coordenadas en las que residen una serie de individuos diagnosticados con una determinada enfermedad.

Las variables de interés X pueden ser continuas (como la concentración de los contaminantes atmosféricos), o discretas (como el número de muertes por cáncer de pulmón). La única restricción que se considera es que exista algún tipo de dependencia entre dos realizaciones con distinta referencia o posición. Dos características de los procesos espaciales (2.1) son, que las observaciones son más similares cuanto más cercanas son sus posiciones (independientemente del valor observado) y que, a medida que la distancia de las localizaciones aumenta, la correlación entre las variables tenderá a anularse [20].

2.1.2. Sistemas de referencia de coordenadas

Partiendo de la definición de proceso espacial dada en la ecuación (2.1), destacar que los puntos s de los datos geoestadísticos o de los procesos puntuales constan de dos componentes $s = (s_x, s_y)$ y para su representación espacial, se debe tener en cuenta el sistema de referencia de coordenadas (CRS). En la Tierra se pueden emplear dos sistemas de localización, por una parte se puede usar un sistema de información geográfica (también llamado no proyectado), o se pueden usar sistemas de información proyectados [21].

- Sistemas de información geográficos: usan un sistema de localización de latitud y longitud sobre una superficie elipsoidal tridimensional de la Tierra. Medidas: -90 a 90 (S-N) y -180 a 180 (O-E). Medidas en grados y minutos (DD) o grados, minutos y segundos (DMS).
- Sistema de información proyectado: utiliza un sistema de coordenadas Cartesiano (E-N), para hacer referencia a una localización en dos dimensiones de la representación de la Tierra. La Tierra está dividida en 60 zonas y 6 grados de longitud en ancho, cada zona usa una proyección transversa de Mercator distinta y no pueden considerarse todas las proyecciones de forma simultánea.

2.1.3. Tasa de mortalidad estandarizada

La aplicación de la tasa de incidencia estandarizada (SIR) a datos de mortalidad es lo que se conoce como tasa de mortalidad estandarizada (SMR). Resulta de gran utilidad en estudios epidemiológicos, donde interesa ofrecer estimaciones de riesgos de mortalidad por enfermedades en cada una de las particiones de un área de estudio [21].

Partiendo del área de estudio D, que sería la comunidad autónoma de Galicia, se tienen:

- Particiones de dicho área de estudio en municipios i, indexados en $i = 1, \ldots, n$.
- En cada municipio hay un número de muertes observado que se denomina Y_i.
- Para cada municipio también hay un número de muertes esperadas denominado E_i.
- Puede haber j estratos de una población estándar, indexados en $j = 1, \ldots, m$.
- La ratio r_j: es el número de muertos dividido por la población en el estrato j de una población estándar.
- n_j^i: es la población en el estrato j del municipio i.

Para cada área $i, i = 1, \ldots, n$, el SMR se define como el ratio de muertes observadas frente a las esperadas.

$$SMR_i = Y_i/E_i. \tag{2.2}$$

E_i se puede calcular usando estandarizaciones indirectas como:

$$E_i = \sum_{j=1}^{m} r_j n_j^i. \tag{2.3}$$

En la práctica, cuando la información del estrato no está disponible, se puede calcular el número de muertes esperado como:

$$E_i = rn^i, \tag{2.4}$$

donde r es la tasa en la población estándar y n^i es la población del municipio i.

2.2. Hipótesis sobre los procesos

Generalmente, no se dispone de varias realizaciones del proceso, si no que se observa una única muestra (una realización) por lo que resulta complicado hacer inferencia sin mayores suposiciones. Es habitual asumir que el proceso de estudio sea estacionario. En adelante, se asumirá que los procesos espaciales estudiados tiene una media, $\mu(s) = E(X(s))$ y que la varianza de $X(s)$ existe para todo $s \in D$.

El proceso $X(s)$ es Gaussiano, si para cualquier tamaño muestral $n \geq 1$ y para cualquier conjunto de sitios $\{s_1, \ldots, s_n\}$, $X = (X(s_1), \ldots, X(s_n))$ tiene una distribución normal multivariante. El proceso se dice que es *estrictamente estacionario*, si para cualquier $n \geq 1$, para cualquier conjunto de sitios $\{s_1, \ldots, s_n\}$ y para cualquier distancia $h \in R^d$, la distribución de $Z = (X(s_1), \ldots, X(s_n))^T = (X(s_1 + h), \ldots, X(s_n + h))$ [20]. Trabajando con datos reales, pocas veces se verifica la estacionariedad estricta, ya que establece que las distribuciones de probabilidad conjunta permanezcan invariantes a cambios de localización, es decir:

$$F_{s_1,\ldots,s_n}(x_1, \ldots, x_n) = F_{s_1+h,\ldots,s_n+h}(x_1, \ldots, x_n) \quad \forall \quad n \in \mathbb{N}. \tag{2.5}$$

Una condición menos exigente es la *estacionariedad de segundo orden*, o *estacionariedad débil*, que implica que la esperanza sea constante y que la función de covarianzas sea invariante a traslaciones:

$$E[X(s)] = \mu, \quad \forall \quad s \in D. \tag{2.6}$$

$$Cov[X(s), X(s')] = C(s - s'), \quad \forall \quad s, s' \in D, \tag{2.7}$$

siendo $C(\cdot)$ el *covariograma* o *función de covarianza*. Cuando la esperanza de la variable no es constante no habrá estacionariedad. La varianza de un proceso estacionario débil es finita e independiente de la localización espacial s porque se verifica que $\sigma^2(s) = Var[X(s)] = C(0) = \sigma^2$.

Si la correlación entre los datos no depende de la dirección en la que esta se calcula, se dice que el fenómeno es *isotrópico*, en caso contrario se habla de *anisotropía*. Es decir, si:

$$Cov(X(s), X(s')) = C(||s - s'||), \tag{2.8}$$

donde $||\cdot||$ representa la norma euclídea, se dirá que el proceso es *isotrópico*, de lo contrario el proceso es *anisotrópico*. Una forma más intuitiva de indicar esto consiste en aclarar que s y s' son las localizaciones donde se observa la variable en estudio y que $s - s'$ es el vector de diferencias entre las dos localizaciones. Dicho vector tendrá una longitud (distancia entre las localizaciones) y un sentido; si la covarianza es únicamente función de distancia (longitud del vector diferencia entre las localizaciones)

entonces habrá isotropía, pero si también depende del sentido se tiene anisotropía.

Dado que muchos fenómenos físicos poseen una capacidad infinita de dispersión, de modo que no existe la varianza correspondiente a la variable regionalizada, resulta útil emplear lo que se conoce como *estacionariedad intrínseca*, en la cual:

$$E[X(s) - X(s')] = 0 \quad \forall \quad s \in D. \tag{2.9}$$

$$Var[X(s) - X(s')] = 2\gamma(s - s'), \quad \forall \quad s, s' \in D, \tag{2.10}$$

donde la función $2\gamma(s - s') = 2\gamma(h)$ recibe el nombre de *variograma* y $\gamma(h)$ es el *semivariograma*. En este caso, si:

$$Var(X(s) - X(s')) = 2\gamma(||s - s'||), \tag{2.11}$$

donde $||\cdot||$ representa la norma euclídea, se dirá que el proceso es isotrópico. En caso contrario se diría que el proceso intrínseco es anisotrópico.

Tras ver estos dos últimos tipos de estacionariedad, queda claro que la clase de procesos intrínsecamente estacionarios es más general que la clase de procesos estacionarios débiles, ya que no requiere que $Var[X(s)]$ (2.11) sea finita ni independiente de la localización espacial s. De todos modos,cuando un proceso es gaussiano, entonces la estacionariedad intrínseca, la estacionariedad de segundo orden y la estacionariedad fuerte son equivalentes. De hecho, hay una relación entre el covariograma y el semivariograma que en el caso isotrópico para una distancia h:

$$C(h) = \sigma^2 - \gamma(h) \tag{2.12}$$

2.3. Modelado de la correlación espacial

Como se mentó previamente en este capítulo, la primera etapa en el desarrollo del análisis geoestadístico es la determinación de la dependencia espacial, también conocida como análisis estructural de los datos. Para llevarla a cabo, con base en la información muestral, se pueden usar tres funciones (el semivariograma, el covariograma y el correlograma). A continuación, se hace un breve resumen de los conceptos asociados a los conceptos de semivariograma y covariograma que son los relevantes para este TFM [20].

2.3.1. Variograma y semivariograma

Se conoce como semivariograma, $\gamma(h)$, a la mitad del variograma $2\gamma(h)$. El semivariograma consta de los siguientes elementos:

- Umbral o meseta: denominación que recibe el limite $\lim_{h\to\infty} \gamma(h)$ si γ está acotado.

  De esta definición de meseta se infiere que no siempre existe. Los semivariogramas que tienen meseta finita cumplen con la hipótesis de estacionariedad estricta, mientras que cuando ocurre lo contrario, el semivariograma define un fenómeno natural que cumple sólo con la hipótesis de estacionariedad intrínseca [20]. Si el proceso es estacionario de segundo orden, el umbral o meseta coincide con la varianza σ^2.

- Pepita o efecto pepita: es un término extraído de la aplicación a la minería. Alude a la situación en que el semivariograma no tiende a 0 al acercarse al origen. Es decir, siempre se verifica que $\gamma(0) = 0$, pero puede ocurrir que $\lim_{h\to 0} \gamma(h) \neq 0$. En este caso, al valor $c_0 = \lim_{h\to 0} \gamma(h)$ se le

denomina pepita y puede deberse a diversas razones (discontinuidad en el origen, inexistencia de datos para distancias pequeñas, etc.)

- Meseta parcial o umbral parcial: cuando el semivariograma tiene efecto pepita, a la diferencia $\sigma^2 - c_0$ se le denomina meseta parcial o umbral parcial.
- Rango o alcance: Si σ^2 es el umbral, el rango o alcance, si existe, es un valor real r, tal que si $||h|| \geq r$, entonces $\gamma(h) = \sigma^2$. Es el valor a partir del cual las variables $X(s)$ y $X(s+h)$ son incorreladas. No siempre existe el rango, ya que si la función carece de meseta, consecuentemente carecerá de rango. Cuando el proceso es estacionario de segundo orden, siempre puede definirse el rango asintótico o efectivo como el valor real r', tal que si $||h|| \geq r'$, entonces se ha alcanzado la meseta parcial $\gamma(t) \geq c_0 + 0,95$.

Dado un conjunto de $X(s_1), \ldots, X(s_n)$ observaciones de un proceso que suponemos isotrópico, considerando una distancia h, el variograma empírico para dicha distancia h puede ser estimado mediante

$$2\hat{\gamma}(h) = \frac{\sum_{s_i, s_j \in B(h)} (X(s_i) - X(s_j))^2}{|B(h)|}, \tag{2.13}$$

donde $|B(h)|$ es el número de pares $B(h) = \{(s_i, s_j), ||s_i - s_j|| \simeq h\}$. El variograma empírico $2\hat{\gamma}(h)$ es un estimador de $2\gamma(h) = Var(X(s_i) - X(s_j))$.

Partiendo de la ecuación (2.13) se puede llegar a la expresión del semivariograma empírico

$$\hat{\gamma}(h) = \frac{\sum_{s_i, s_j \in B(h)} (X(s_i) - X(s_j))^2}{2|B(h)|}, \tag{2.14}$$

donde $B(h)$ es el conjunto de pares de localizaciones a distancia aproximada h.

2.3.2. Covariograma

Dado un conjunto de $X(s_1), \ldots, X(s_n)$ observaciones de un proceso que suponemos isotrópico, separadas por una distancia h, el covariograma empírico entre pares de observaciones puede calcularse mediante

$$\hat{C}(h) = \frac{1}{|B(h)|} \sum_{s_i, s_j \in B(h)} (X(s_i) - \hat{\mu})(X(s_j) - \hat{\mu}) = \frac{1}{|B(h)|} \sum_{s_i, s_j \in B(h)} X(s_i)X(s_j) - \hat{\mu}^2, \tag{2.15}$$

donde $\hat{\mu}$ es la media muestral en todo punto de la región de estudio y $|B(h)|$ es el número de pares de puntos que se encuentran a una distancia aproximada h [20]. $\hat{C}(h)$ es un estimador de $C(h) = Cov(X(s_i), X(s_j))$.

Asumiendo estacionariedad, tanto el semivariograma (2.14) como el covariograma (2.15) son válidos para determinar la relación espacial entre los datos. Sin embargo, el semivariograma es el único que no requiere hacer estimación de parámetros; por ello, en la práctica se suele utilizar el semivariograma en lugar del covariograma.

2.4. Descomposición de la variabilidad

Lo más habitual es que un proceso aleatorio presente una parte determinista y otra aleatoria:

$$X(s) = \mu(s) + \varepsilon(s),$$

siendo $\mu(s) = E[X(s)], \quad \forall \quad s \in D$, es la media o tendencia determinista (no aleatoria) del proceso y representa los cambios o evolución a gran escala. $\varepsilon(s)$ es la componente aleatoria o proceso de error y representa el comportamiento local o evolución a pequeña escala. Generalmente, se suele imponer que $\varepsilon(s)$ sea un proceso estacionario. Para la estimación del variograma o la obtención de predicciones, normalmente se plantea alguna de las siguientes condiciones para la tendencia determinista $\mu(s)$:

- $\mu(s)$ es conocida.
- $\mu(s) = \mu$, es decir, la media es constante.
- $\mu(s)$ se puede escribir como una combinación lineal de funciones conocidas.
- $\mu(s)$ se puede estimar a partir de los datos de una forma adecuada.

Una vez descritas las relaciones entre puntos, el siguiente paso sería hacer las predicciones en sitios de la región no muestreados.

2.5. Modelado predictivo

Se entiende por modelado predictivo al proceso por el cual se construye una herramienta matemática o un modelo capaz de generar una predicción. Desde la estadística espacial se han realizado numerosos avances en el ajuste de regresiones con datos espaciales tanto desde una perspectiva frecuentista, como desde un enfoque Bayesiano [14].

2.5.1. Enfoque frecuentista vs. Bayesiano

La principal diferencia entre el enfoque frecuentista y el Bayesiano radica en su formulación de lo que es la probabilidad de un evento. Tradicionalmente se define la probabilidad de un evento como, la representación de la medida de ocurrencia de un evento. Ésta puede variar entre 0 y 1, lo que representa, en sus extremos, una certeza completa de que el evento está ocurriendo (probabilidad = 1) o no está ocurriendo (probabilidad = 0); cualquier cosa entre estos dos valores proporciona el grado de incertidumbre sobre si el evento es verdadero.

Mientras que en el enfoque frecuentista, los parámetros del modelo a ajustar se consideran constantes fijas desconocidas y deben ser estimadas a partir de los datos, sin considerar información extra-muestral, en el enfoque Bayesiano se aprovecha tanto la información procedente de los datos muestrales así como la información extra-muestral disponible, siendo esto último, toda la información relevante que sirve para disminuir la incertidumbre o ignorancia en torno a un fenómeno aleatorio de interés. En el enfoque Bayesiano, se considera que los parámetros del modelo de regresión son variables aleatorias y consecuentemente se pueden asumir distribuciones de probabilidad asociadas a los parámetros [14].

Partiendo de estas observaciones, si no hay datos, resulta evidente que la estadística frecuentista no puede operar; del mismo modo, si hay muy pocos datos, también surgen problemas, ya que muchos de los métodos se apoyan en resultados asintóticos, tales como la ley de los grandes números o el teorema central del límite. En definitiva, para un funcionamiento fiable de los métodos frecuentistas, hacen falta tamaños muestrales grandes. Como se puede apreciar en el análisis descriptivo realizado en el primer capítulo, sobre los datos de mortalidad por cáncer de pulmón en municipios gallegos, la muestra de municipios con datos es reducida, por lo que el presente estudio se abordará con un enfoque Bayesiano, de modo que [14]:

- La probabilidad describe grados de creencia, no frecuencias límite. Por lo tanto, se podrán realizar afirmaciones probabilísticas sobre los parámetros.

- Las distribuciones de probabilidad de los parámetros permitirán realizar inferencia sobre el valor esperado del parámetro y la credibilidad asociada a esa estimación.
- Las distribuciones de los valores predichos permitirán derivar no solo predicciones puntuales sino también medidas de incertidumbre para cada predicción obtenida por el modelo.

Los modelos Bayesianos jerárquicos son utilizados con frecuencia para modelar datos espaciales y espacio-temporales. Estos modelos permiten total flexibilidad en el modelado espacial y espacio-temporal. Además, los métodos Bayesianos permiten trabajar con modelos complejos que son difíciles de ajustar mediante métodos clásicos, como sería el escenario con medidas repetidas, datos faltantes o multivariantes [14, 21].

Se procederá a utilizar una notación común para los eventos de interés de ambos tipos de procesos espaciales. Tanto la concentración de los elementos metálicos, X_s, en los datos geoespaciales, como las muertes observadas en los datos reticulares, Y_i, serán denominados como la variable de interés y para la explicación de la aproximación Bayesiana.

En una aproximación Bayesiana, la distribución de probabilidad $\pi(y|\theta)$, es especificada para unos datos observados $y = (y_1, \ldots, y_n)$, dado un vector de parámetros desconocido θ. Para ello, una distribución previa $\pi(y|\eta)$ es asignada a un θ, siendo η un vector de hiperparámetros. La distribución previa de θ representa lo que se sabe sobre θ antes de obtener los datos de y. Si η es conocida, la inferencia respecto a θ se basa en la distribución a posteriori de θ definida por el Teorema de Bayes como [21]:

$$\pi(\theta|y) = \frac{\pi(y,\theta)}{\pi(y)} = \frac{\pi(y|\theta)\pi(\theta)}{\int \pi(y|\theta)\pi(\theta)d\theta} \tag{2.16}$$

El denominador $\pi(y) = \int \pi(y|\theta)\pi(\theta)d\theta$ es la probabilidad marginal de y. Esto es independiente de θ y se puede establecer como una constante de escala que no afecta a la forma de la distribución a posteriori.

2.5.2. MCMC e INLA

Una dificultad a la hora de aplicar métodos bayesianos es el cálculo de las probabilidades a posteriori $\pi(\theta|y)$, que habitualmente involucra integrales múltiples. Esto tiene como consecuencia que aunque la probabilidad y su distribución a priori tengan expresiones cerradas, la distribución a posteriori no tiene por qué tener expresiones cerradas. Este tipo de problemas suelen ser resueltos empleando Cadenas de Markov Monte Carlo (MCMC) para aproximar sus valores [21].

Los métodos MCMC trabajan generando una muestra de valores $\{\theta^g, g = 1, \ldots, G\}$, de una cadena de Markov convergente, cuya distribución es la posterior $\pi(\theta|y)$. A partir de estas muestras, resúmenes empíricos de los valores de θ^g pueden ser usados para resumir la distribución a posteriori de los parámetros de interés. Para ver si las cadenas MCMC han alcanzado la distribución estacionaria se pueden examinar los *traceplots*, que son gráficos en los que se muestra el valor del parámetro en cada iteración y se puede ver a partir de qué punto comienzan a converger las distintas simulaciones [21].

La aproximación de Laplace basada en integrales anidadas (INLA), es computacionalmente menos costosa que las MCMC, por lo que resulta una alternativa para llevar a cabo inferencia Bayesiana aproximada en modelos Gaussianos latentes. Estos modelos van desde modelos lineales generalizados mixtos, a modelos espaciales y espacio-temporales. INLA utiliza una combinación de aproximaciones analíticas y algoritmos numéricos para matrices dispersas, para aproximar las distribuciones a posteriori con expresiones cerradas. Esto permite hacer inferencia de forma más rápida, evitando problemas de convergencia de muestras o mezcla, lo que permite ajustar un gran número de conjuntos de datos

y explorar modelos alternativos [21].

2.5.3. Soporte teórico de INLA

Como se comentó previamente, se parte de dos conjuntos de datos principales. Por una parte, los datos geoestadísticos de concentración de contaminantes, donde la variable observada de interés (concentración del contaminante) se denomina $X_{s_i} = x_i$; y por otro lado, los datos reticulares de mortalidad, donde la variable observada de interés (muertes observadas por municipio) se denomina Y_i. Como la aplicación de la aproximación de Laplace basada en integrales anidadas (INLA) es común tanto para los datos reticulares como para los geoestadísticos, se denominará y a los datos observados, del mismo modo que se hizo para la explicación de la inferencia bayesiana. $y = (y_1, \ldots, y_n)$ donde las observaciones pueden denominarse $y(s_i)$ en caso de ser datos geoestadísticos o y_i para datos reticulares.

Se empleará como referencia el libro Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny [21] para abordar los fundamentos de INLA. INLA se aplica sobre modelos del tipo

$$\begin{aligned} y_i|x, \theta &\sim \pi(y_i|x_i, \theta), i = 1, \ldots, n, \\ x|\theta &\sim N(\mu(\theta), Q(\theta)^{-1}), \\ \theta &\sim \pi(\theta), \end{aligned}$$

donde y son los datos observados, x representa el campo Gaussiano y θ son los hiperparámetros. $\mu(\theta)$ es la media y $Q(\theta)$ es la matriz de precisión del campo x. Aquí, x e y pueden ser multidimensionales; pero para hacer inferencia de forma rápida es conveniente que las dimensiones del vector de hiperparámetros θ sea pequeño, dado que las aproximaciones son calculadas usando integrales numéricas sobre el espacio de hiperparámetros.

En muchos casos, se asume que las observaciones y_i forman parte de una familia exponencial con media $\mu_i = g^{-1}(\eta_i)$. El predictor lineal η_i registra el efecto de las covariables de forma aditiva

$$\eta_i = \alpha + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}). \tag{2.17}$$

En (2.17), α es el intercepto, β_k cuantifica el efecto lineal de las covariables z_{ki} en la respuesta, y $f^{(j)}(\cdot)$ es el conjunto de efectos aleatorios dados en términos de covariables u_{ji}. Dada la gran variedad de formas que puede tomar la función $f^{(j)}$, esta formulación resulta flexible para ajustar una gran variedad de modelos.

Como se mencionó previamente, INLA usa una combinación de aproximaciones analíticas y algoritmos numéricos para aproximar las distribuciones a posteriori de los parámetros. Sea $x = (\alpha, \beta_k, f^{(j)})|\theta \sim N(\mu(\theta), Q(\theta)^{-1})$ un vector de variables del campo Gaussiano latente, INLA calcula de forma rápida y precisa las aproximaciones de las distribuciones marginales a posteriori de las variables de las componentes del campo Gaussiano latente

$$\pi(x_i|y), i = 1, \ldots, n, \tag{2.18}$$

así como la distribución marginal para los hiperparámetros del modelo Gaussiano latente

$$\pi(\theta_j|y), j = 1, \ldots, dim(\theta). \tag{2.19}$$

Las marginales a posteriori de cada elemento de x_i del campo latente x, viene dado por

$$\pi(x_i|y) = \int \pi(x_i|\theta, y)\pi(\theta|y)d\theta, \tag{2.20}$$

y las marginales a posteriori de los hiperparámetros pueden indicarse como

$$\pi(\theta_j|y) = \int \pi(\theta|y)d\theta_{-j}. \tag{2.21}$$

Esta formulación anidada es utilizada para aproximar $\pi(x_i|y)$ mediante combinaciones de aproximaciones analíticas (a las condicionales $\pi(x_i|\theta, y)$ y $\pi(\theta|y)$) y rutinas de integrales numéricas para integrar θ. Análogamente, $\pi(\theta_j|y)$ es aproximada mediante aproximaciones de $\pi(\theta|y)$ e integrando θ_{-j}. Más específicamente, la densidad a posteriori de los hiperparámetros es aproximada usando una aproximación Gaussiana a posteriori del campo latente, $\tilde{\pi}_G(x|\theta, y)$, evaluada del siguiente modo, $X^*(\theta) = arg \;\; max_x \pi_G(x|\theta, y)$,

$$\tilde{\pi}(\theta|y) \propto \left.\frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x|\theta, y)}\right|_{x=x^*(\theta)}. \tag{2.22}$$

Después INLA construye la siguiente aproximación anidada:

$$\tilde{\pi}(x_i|y) = \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y)d\theta, \tilde{\pi}(\theta_j|y) = \int \tilde{\pi}(\theta|y)d\theta_{-j}. \tag{2.23}$$

Finalmente, estas aproximaciones pueden ser integradas numéricamente respecto a θ

$$\tilde{\pi}(x_{i|y}) = \sum_k \tilde{\pi}(x_i|\theta_k, y)\tilde{\pi}(\theta_k|y) \times \triangle_k,$$
$$\tilde{\pi}(\theta_j|y) = \sum_l \tilde{\pi}(\theta_l^*|y) \times \triangle_l^*,$$

donde $\triangle_k$ ($\triangle_l^*$) indica el peso del área correspondiente a θ_k (θ_l^*).

La aproximación para las marginales a posteriori para las x_i condicionadas a los valores seleccionados de θ_k, $\tilde{\pi}(x_i|\theta_k, y)$, puede ser obtenida utilizando una aproximación Gaussiana, una de Laplace o una aproximación simplificada de Laplace. La forma más simple y rápida es utilizar una aproximación Gaussiana derivada de $\tilde{\pi_G}(x|\theta, y)$. De todos modos, en algunas situaciones, esta aproximación produce errores en la localización y falla para detectar un comportamiento asimétrico. La aproximación de Laplace es preferible a la Gaussiana, pero es más costosa. La aproximación de Laplace simplificada (que es la que viene implementada por defecto en el paquete "R-INLA") tiene un coste más pequeño y remedia los problemas que surgían con la aproximación Gaussiana referente a la asimetría y la localización.

2.5.4. Soporte teórico de la aproximación SPDE

Cuando se consideran datos geoestadísticos, generalmente se asume que hay una variable espacialmente continua subyacente a las observaciones. Dicha variable puede ser modelada usando campos gaussiano $\{X(s), \;\; s \in D\}$. Por lo tanto, se pueden utilizar las ecuaciones diferenciales parciales estocásticas implementadas en el paquete "R-INLA" para ajustar un modelo espacial y predecir las variables de interés en localizaciones no muestreadas. Un campo gaussiano con una matriz de covarianzas Matérn puede ser expresado como solución al siguiente dominio continuo SPDE:

$$(\kappa^2 - \triangle)^{\alpha/2}(\tau X(s)) = \mathcal{W}(s). \tag{2.24}$$

Donde $X(s)$ es un campo gaussiano y $\mathcal{W}(s)$ es un es un proceso de ruido blanco espacial gaussiano. α controla el grado de suavidad del campo gaussiano, τ controla la varianza y $\kappa > 0$ es un parámetro de

escala. $\triangle$ es el Laplaciano definido como $\sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$, donde d es la dimensión del dominio espacial D.

El parámetro de suavizado del parámetro ν de la matriz de covarianzas Matérn está relacionada con las SPDE mediante:

$$\nu = \alpha - \frac{d}{2}, \tag{2.25}$$

y la varianza marginal σ^2 está relacionada con las SPDE mediante:

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}. \tag{2.26}$$

Para $d = 2$ y $\nu = 1/2$ (que se corresponde a la función de covarianzas exponencial), el parámetro $\alpha = \nu + d/2 = 1/2 + 1 = 3/2$. En el paquete "R-INLA" el valor por defecto es $\alpha = 2$, aunque hay otras versiones disponibles ($0 \leq \alpha < 2$).

Una solución aproximada de las SPDE puede ser obtenida utilizando el método de los elementos finitos. Este método divide el dominio espacial D en un conjunto de triángulos que no se intersecan, de modo que se tiene una malla de triangulación con N nodos y N funciones base. Las funciones base $\psi(\cdot)$ se definen como funciones lineales por partes en cada triángulo que es igual a 1 en el vértice k, y es igual a 0 el resto de vértices. Por tanto, el campo Gaussiano continuamente indexado x se representa como un campo aleatorio de Markov Gaussiano discretamente indexado mediante las funciones de base finita definidas en la malla de triangulación:

$$X(s) = \sum_{k=1}^{N} \psi_k(s) X_k, \tag{2.27}$$

donde N es el número de vértices de triangulación, $\psi_k(\cdot)$ denota las funciones lineales base por partes y X_k.

A la distribución conjunta del vector de pesos se le asigna una distribución Gaussiana $x = (x_1, \ldots, x_n) \sim N(0, Q^{-1}(\tau, \kappa))$ que aproxima la solución $x(s)$ de las SPDE en los nodos de la malla, y las funciones base transforman la aproximación $x(s)$ de los nodos de la malla a otras localizaciones de interés.

Capítulo 3

Modelado de datos geoestadísticos

En este capítulo se expone la forma de implementar en R tanto la aproximación de Laplace basada en integrales anidadas, como el enfoque de las ecuaciones diferenciales parciales estocásticas para el análisis de datos geoestadísticos. Este capítulo está casi a modo de tutorial para que el modelado sea reproducible. Se comenzará haciendo una introducción al paquete "R-INLA" y su función principal `inla()` [17]. A continuación, se expondrá la metodología para modelar los datos geoestadísticos y finalmente, se aplicará a datos reales de concentración de contaminantes atmosféricos [21].

3.1. Paquete R-INLA

Lo primero que hay que destacar es que el paquete "R-INLA" [17] no está en el repositorio de CRAN (hay que descargarlo de enlaces como http://stackoverflow.com/ o del repositorio de R-INLA). Para instalar la versión estable basta con usar la siguiente instrucción: `install.packages("INLA", repos = "https://inla.r-inla-download.org/R/stable", dep = TRUE)`.

Una vez instalado el paquete, se carga con la función `library(INLA)` y se puede proceder con el ajuste del modelo. Para ajustar un modelo usando INLA se tiene que hacer en dos pasos. En primer lugar, se escribe el predictor lineal del modelo con un objeto fórmula en R [26]. A continuación, se corre el modelo llamando a la función `inla()`, donde se especifica la fórmula, la familia, los datos y otras opciones.

La ejecución de `inla()` devuelve un objeto que contiene la información del modelo ajustado, incluyendo varios resúmenes y las distribuciones marginales a posteriori de los parámetros, los predictores lineales y los valores ajustados. Dichas distribuciones a posteriori pueden ser pos-procesadas utilizando un conjunto de funciones disponibles en "R-INLA" [17]. El paquete "R-INLA" también proporciona estimaciones acorde a distintos criterios para evaluar y comparar modelos bayesianos. Estos incluyen el criterio de información de la desviación del modelo (DIC), el criterio de información de Watanabe-Akaike (WAIC), la probabilidad marginal y las ordenadas predictivas condicionales (CPO).

3.1.1. Predictor Lineal

Se entiende por predictor lineal a una combinación lineal de un conjunto de coeficientes y variables explicativas (variables independientes), cuyos valores son utilizados para predecir los valores de una variable respuesta o dependiente.

La sintaxis de un predictor lineal en R-INLA es similar a la sintaxis usada para ajustar los modelos lineales con la función `lm()`. Es decir, primero se indica la variable respuesta, a continuación se añade la virgulilla, `~`, y finalmente se indican los efectos fijos y aleatorios separados por operadores `+`. Los

efectos aleatorios se especifican usando la función `f()`. El primer argumento de `f()` es un vector de índices que indican los efectos aleatorios que actúan sobre cada observación y el segundo argumento, es el nombre del modelo. Parámetros adicionales de la función `f()` pueden ser consultadas escribiendo `?f`.

3.1.2. Ajuste del modelo

La función `inla()` es utilizada para ajustar el modelo. Sus principales argumentos son:

- `formula`: objeto fórmula que establece el predictor lineal utilizado.
- `data`: datos suministrados en formato *data.frame*. Si se desea predecir la variable respuesta para algunas observaciones, es necesario especificar la variable respuesta de esas observaciones como `NA`.
- `family`: argumento que requiere una cadena o vector de cadenas que establezcan la verosimilitud de la familia como Gaussiana (`"gaussian"`), de Poisson (`"poisson"`) o binomial (`"binomial"`). Por defecto la familia es Gaussiana. Un listado de familias alternativas pueden ser vistas ejecutando `names(inla.models()$likelihood)`; además, se pueden consultar los detalles para cada familia con `inla.doc("nombre de la familia")`.
- `control.compute`: listado con las especificaciones de numerosas variables computacionales como `dic`, que es una variable booleana que indica si se debería calcular el DIC.
- `control.predictor`: listado con las especificaciones de numerosas variables computacionales como `link`, que es la función de enlace del modelo, y `compute`, que es la variable booleana que indica si las densidades marginales para el predictor lineal deberían ser calculadas.

3.1.3. Especificaciones a priori

Los nombres de las especificaciones a priori disponibles en "R-INLA" [17] pueden ser consultados con `names(inla.models()$prior)`, y un listado de las opciones sobre cada especificación puede ser consultado mediante `inla.models()$prior`. La documentación respecto a una determinada especificación a priori se puede visualizar a través de la consulta `inla.doc("nombre especificación a priori")`.

Por defecto, al intercepto del modelo se le asigna una especificación a priori Gaussiana con media y precisión igual a 0. Al resto de los efectos establecidos también se les asignan distribuciones Gaussianas con media igual a 0 y precisión igual a 0,001. Estos valores pueden observarse con `inla.set.control.fixed.default()[c("mean.intercept", "prec.intercept", "mean", "prec")]`. Estos valores de las especificaciones a priori se pueden cambiar con el argumento `control.fixed` de `inla()`, utilizando una lista con la media y la precisión de las distribuciones Gaussianas. Más concretamente, la lista contiene la media y la precisión a priori para el intercepto (`mean.intercept` y `prec.intercept`) así como la media y la precisión para todos los efectos fijos menos el intercepto (`mean` y `prec`). Quedando la estructura general como:

```
prior.fixed <- list(mean.intercept = <>, prec.intercept = <>, mean = <>, prec = <>)
```

Las especificaciones a priori de los hiperparámetros θ se asignan con el argumento `hyper` de la función `f()`:

```
formula <- y ~ 1 + f(<>, model = <>, hyper = prior.f)
```

Las especificaciones a priori de los parámetros de la verosimilitud son asignados con el argumento `control.family` de la función `inla()`:

```
res <- inla(formula, data = d, control.fixed = prior.fixed,
control.family = list(..., hyper = prior.l))
```

`hyper` trabaja con un listado que contiene los nombres de cada hiperparámetro y los valores de las especificaciones a priori. De forma más específica, el listado contiene los siguientes campos:

- `initial`: valor inicial del hiperparámetro (unos valores iniciales adecuados de los hiperparámetros pueden hacer que el proceso de inferencia sea mucho más rápido).
- `prior`: nombre de la distribución a priori.
- `param`: vector con los valores de los parámetros de la distribución a priori.
- `fixed`: variable booleana que indica si el hiperparámetro es un valor fijo.

```
prior.prec <- list(initial = <>, prior = <>, param = <>, fixed = <>)
prior <- list(prec = prior.prec)
```

"R-INLA" [17] también proporciona un entorno de trabajo útil para construir especificaciones a priori con penalización por complejidad (PC). Las especificaciones a priori con PC penalizan las desviaciones del modelo base. Las especificaciones con PC controlan la flexibilidad, reducen el sobreajuste y mejoran el rendimiento predictivo. Destacar además, que tienen un único parámetro que controla el grado de flexibilidad del modelo:

$$P(T(\xi) > U) = \alpha, \tag{3.1}$$

donde $T(\xi)$ es una transformación interpretable del la flexibilidad del parámetro ξ, U es el limite superior que especifica el evento de cola y α es la probabilidad del evento.

3.2. Metodología de modelado de datos geoestadísticos

El primer paso para abordar de forma adecuada el modelado de datos tanto reticulares como geoestadísticos es conocer su naturaleza y propiedades. Este aspecto se adelantó en el primer capítulo con el análisis descriptivo de los datos. Tras familiarizarse con las características de los datos, el paso natural en el enfoque bayesiano consiste en establecer el modelo que mejor se pueda ajustar al comportamiento del evento de interés.

3.2.1. Establecer un modelo

Para abordar esta sección se va a considerar que el proceso espacial (2.1) ocurre de forma continua en el espacio y por tanto, se pueden usar modelos geoespaciales para predecir los valores en otras localizaciones a partir de un conjunto de observaciones del evento de interés X_s, en una serie de localizaciones s_i. Por ejemplo, se puede asumir que el evento de interés en la localización s_i, X_i, sigue una distribución Gaussiana con media μ_i y varianza σ^2. La media μ_i es expresada como la suma de un intercepto, β_0 y un efecto aleatorio espacialmente estructurado que sigue un proceso Gaussiano con media cero y función de covarianza Matérn:

$$X_s \sim N(\mu_i, \sigma^2), u = 1, 2, \ldots, n, \tag{3.2}$$

donde $\mu_i = \beta_0 + Z(s_i)$.

A continuación, se describen los pasos necesarios para ajustar este modelo usando la aproximación por SPDE con el paquete "R-INLA" [17].

3.2.2. Construcción de la malla de triangulación

Con la aproximación por SPDE se aproxima el campo Gaussiano continuo $Z(\cdot)$ como un campo aleatorio de Markov Gaussiano mediante una función de base finita sobre una malla de triangulación de la región de estudio. Dicha malla de triangulación puede ser diseñada con la función `inla.mesh.2d()` del paquete "R-INLA". Los principales argumentos de esta función son:

- `loc`: coordenadas utilizadas como vértices iniciales de la malla de triangulación.
- `boundary`: objeto que define las fronteras del área de estudio.
- `offset`: distancia que especifica el tamaño de las extensiones internas y externas alrededor de las localizaciones de los datos.
- `cutoff`: distancia mínima permitida entre puntos. Se utiliza para evitar construir demasiados triángulos pequeños en localizaciones agrupadas.
- `max.edge`: valores que indican el máximo perímetro de los triángulos en la región de estudio y alrededores.

Suponiendo que se ha denominado `mesh` a la malla de triangulación del área de estudio, el número de vértices de triangulación de la malla puede ser consultado con `mesh$n`.

Otra forma de construir la malla de triangulación sería utilizando la frontera de la región de estudio. Por ejemplo, se podría construir un límite no convexo para las coordenadas usando la función `inla.nonconvex.hull()`, para posteriormente pasar dicho límite a la función `inla.mesh.2d()` y así construir la malla.

3.2.3. Construcción del modelo SPDE

Una vez construida la malla de triangulación, se puede utilizar la función `inla.spde2.matern()` para construir el modelo SPDE. Como se mencionó en los fundamentos teóricos del SPDE ($\alpha = \nu + d/2$), `alpha` está relacionado con el grado de suavidad del proceso. Con el argumento `constr = TRUE` se puede establecer una restricción de integración a cero.

3.2.4. Generación del conjunto de índices

El siguiente paso necesario es la generación de un conjunto de índices para el modelo SPDE; para ello se utiliza la función `inla.spde.make.index()`, donde se especifican los nombres de los efectos y el número de vértices en el modelo SPDE.

3.2.5. Diseño de una matriz de proyección

Se necesita construir una matriz de proyección A para proyectar el GRF desde las observaciones hacia los vértices de triangulación. La matriz A tiene el número de filas igual al número de observaciones y el número de columnas igual al numero de vértices de triangulación. La fila i de la matriz A (correspondiente a una observación en la localización s_i), posiblemente tenga tres valores distintos de cero en las columnas que corresponden a los vértices de triangulación que contienen la localización. Si s_i está dentro del triángulo, estos valores son iguales a las coordenadas baricéntricas; esto quiere decir que son proporcionales a las áreas de cada uno de los subtriángulos definidos por la localización s_i y los vértices de cada triángulo (su suma es igual a 1). Si s_i es igual al vértice del triángulo, la fila i tiene solo un valor distinto de 0, que será 1 en la columna que corresponda al vértice. Intuitivamente, el valor $Z(s)$ en la localización que cae dentro del triángulo es la proyección del plano formado por los pesos de los vértices del triángulo en la localización s.

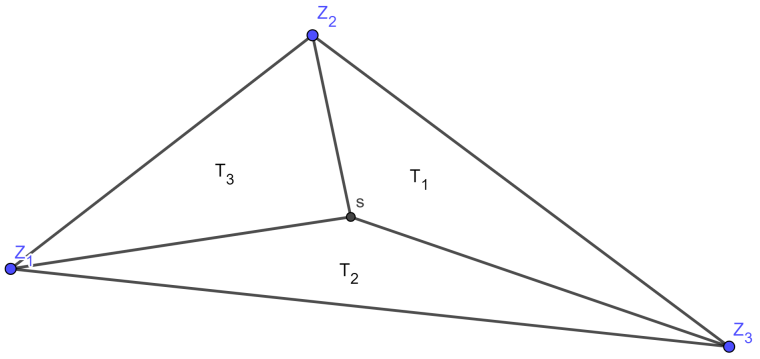

Figura 3.1: Triángulo de una matriz de triangulación. Figura diseñada con Geogebra.

Un ejemplo de una matriz de proyección se da a continuación (3.3) . Esta matriz de proyección proyecta n observaciones en G vértices de triangulación. La primera fila de la matriz corresponde a una observación con una localización que coincide con el vértice número tres. La segunda y última fila corresponden a observaciones con localizaciones que caen dentro del triángulo.

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} & \cdots & A_{1G} \\ A_{21} & A_{22} & A_{23} & \cdots & A_{2G} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & A_{nG} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & \cdots & 0 \\ A_{21} & A_{22} & 0 & \cdots & A_{2G} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & A_{n3} & \cdots & 0 \end{bmatrix} \tag{3.3}$$

Para visualizar lo comentado previamente, imaginemos que estamos en el escenario descrito en la Figura 3.1, donde se muestra una localización s que cae dentro de uno de los triángulos de la red de triangulación. El valor del proceso $Z(\cdot)$ en el punto s se expresa como un peso promedio de los valores del proceso en los vértices del triángulo (Z_1, Z_2 y Z_3) y con pesos igual a T_1/T, T_2/T y T_3/T, donde T denota el área de un gran triángulo que contiene s, y T_1/T, T_2/T y T_3/T son las áreas de los subtriángulos.

$$Z(s) \approx \frac{T_1}{T} Z_1 + \frac{T_2}{T} Z_2 + \frac{T_3}{T} Z_3 \tag{3.4}$$

Dicha matriz de proyección se puede diseñar de forma sencilla introduciendo la malla de triangulación y las coordenadas en la función `inla.spde.make.A()` del paquete "R-INLA" [17]. Asumiendo que se ha llamado `A` a la matriz A, se puede comprobar si el número de filas es igual al número de observaciones y el número de columnas igual al número de vértices de la malla mediante la función `dim(A)`. También se puede comprobar que los elementos de cada fila suman 1 con la función `rowSum(A)`.

3.2.6. Obtención de los datos para la predicción

El siguiente paso consiste en diseñar una matriz con las coordenadas de las localizaciones en las que se desea realizar la predicción. Para ello, en primer lugar se diseña una rejilla con localizaciones equiespaciadas utilizando la función `expand.grid()`. A continuación, se seleccionan únicamente los puntos de

la matriz que caen dentro de las fronteras del área de estudio con la función `point.in.polygon()` del paquete "sp" [23]. También se crea una matriz de proyección con las localizaciones de las predicciones (se vuelve a usar la función `inla.spde.make.A()` con la correspondiente malla y coordenadas).

3.2.7. Agrupación de los datos

Llegados a este punto, es necesario organizar los datos, sus efectos y las matrices de proyección asociadas. Para ello se utiliza la función `inla.stack()` con los siguientes argumentos:

- `tag`: cadena para identificar los datos.
- `data`: lista de los vectores de datos.
- `A`: lista con las matrices de proyección.
- `effects`: lista con los efectos fijos y aleatorios.

3.2.8. Formulación del modelo

Para la implementación del modelo (3.2) usando R-INLA basta con definir un objeto fórmula estableciendo la variable respuesta a la izquierda y añadiendo los efectos fijos y aleatorios a la derecha. Dadas las características del modelo (3.2) se necesita prescindir del intercepto (para lo que se añade un `0`), añadir una covariable (`b0`) y especificar el modelo SPDE con sus correspondientes índices (`s`). La estructura de la fórmula sería:

```
formula <- y ~ 0 + b0 + f(s, model = spde)
```

3.2.9. Llamada de inla()

Para terminar de ajustar el modelo se llama a la función `inla()` utilizando las especificaciones a priori que vienen implementados por defecto en "R-INLA" [17]. Indicar que en el argumento `control.predi ctor` se establece `compute=TRUE` para calcular las predicciones a posteriori. Para que permanezca la continuidad entre subsecciones, se puede suponer que la llamada de esta función se denomina `res` (ya que el objeto que genera contiene los resultados).

3.2.10. Obtención de resultados

Para consultar resultados como la media y los cuantiles $2,5$ y $97,5$ de los valores ajustados, se puede recurrir a la instrucción `res$summary.fitted.values`. Además, los indices de las filas correspondiente a las predicciones pueden ser obtenidas con la función `inla.stack.index()` introduciendo los datos agrupados e indicando la etiqueta `"pred"`. Utilizando los índices y el resumen se pueden crear variables con las medias a posteriori, así como con el límite superior e inferior a posteriori con un 95 % de confianza.

Con los resultados obtenidos se pueden diseñar mapas de predicción utilizando `ggplot()`. Se puede utilizar `geom_tile()` para representar las distintas predicciones en celdas y `facet_wrap()` para mostrar los tres mapas de predicción en el mismo gráfico. También se debe usar la función `coord_fixed(ratio = 1)` para asegurarse que una unidad en el eje de las ordenadas sea igual a una unidad en el eje de abscisas del mapa. Finalmente, se recurre a la función `scale_fill_gradient()` para establecer una gradiente de coloración en función del valor de la variable de interés (generalmente se usa el azul para indicar valores bajos y el naranja para valores altos).

3.2.11. Proyección del campo espacial

De los resultados se pueden extraer vectores (`res$summary.random$s$mean` y `res$summary.random $s$sd`) con la media y desviación típica a posteriori en los nodos del campo espacial. Se pueden proyectar estos valores en distintas localizaciones calculando una matriz de proyección para nuevas localizaciones y luego multiplicando dicha matriz por los valores del campo espacial.

También se pueden proyectar los valores del campo espacial en distintas localizaciones utilizando las funciones `inla.mesh.projector()` e `inla.mesh.project()`. Primero se debe utilizar la función `inla.mesh.projector()` para calcular la matriz de proyección para las nuevas localizaciones (ya sea introduciendo las coordenadas o especificando los límites en el eje de ordenadas y abscisas). A continuación, se utiliza la función `inla.mesh.project()` para proyectar la media y la desviación típica a posteriori del campo espacial. Nuevamente, por aspectos estéticos, se puede seleccionar que sólo se proyecten los campos espaciales que caen dentro de las fronteras de la comunidad de Galicia.

Los valores proyectados se pueden graficar utilizando el paquete "ggplot2"[30]. Para ello, primero se crea un `data.frame()` con las coordenadas de la malla de localizaciones y los valores del campo espacial. A continuación, se utilizan las funciones `ggplot()` y `geom_raster()` para crear los mapas con los valores del campo espacial. Finalmente, se representan los mapas uno al lado del otro en una cuadrícula utilizando la función `plot_grid()` del paquete "cowplot" [32].

3.3. Aplicación a datos reales

Como se mentó en el primer capítulo, se tienen medidas de las concentraciones de un metaloide (arsénico) y ocho elementos metálicos (Al, Cd, Cr, Cu, Fe, Hg, Ni y Pb) en tejidos de *P. purum*. Cada medida está georreferenciada acorde a un sistema de coordenadas proyectadas UTM ED50 zona 29N. Partiendo de este escenario, uno se puede dar cuenta que se van a tener que modelar por separado los 9 elementos considerados. El proceso de modelado, consta de una serie de pasos y resultados comunes para todos los elementos y otros, como el agrupamiento, que varían en función del elemento considerado.

Se comenzará por obtener el mapa del área de estudio (Galicia) con la proyección UTM ED50 zona 29N. Para ello, lo primero es obtener el mapa de España usando la función `getData()` del paquete "raster". Se accede a la base de datos de las fronteras administrativas globales mediante la instrucción `name="GADM"`, se selecciona España con el argumento `country="Spain"` y se especifica que estamos interesados en la subdivisión administrativa de las comunidades autónomas, para lo que se usa el argumento `level=1`. De este modo, se genera un objeto `SpatialPolygonsDataFrame` que alberga la información necesaria para la representación de España con las subdivisiones en comunidades autónomas.

Como sólo nos interesa el área de Galicia, se puede prescindir de los polígonos correspondientes a las islas y el resto de comunidades autónomas. Para ello, se utilizan funciones asociadas a las librerías "sf" [25] y "dplyr" [29]. En primer lugar se convierte el objeto `SpatialPolygonsDataFrame` en un objeto `sf`. A continuación, se selecciona la Comunidad de Galicia. Finalmente, se procede a transformar el sistema de coordenadas del mapa para que haya concordancia con las coordenadas empleadas para el diseño del muestreo. El objeto `sf` que contiene el mapa viene por defecto con un sistema de coordenadas geográfico. Para transformarlo a un objeto con proyección UTM se utiliza la función `st_transform()` especificando el código ESPG de España (23029) con la proyección que le correspondería a Galicia (UTM zona 29 Norte). Llegados a este punto se tendría el mapa sobre el que se proyectan los puntos de la Figura 1.1.

Una vez cargados los datos, hecha su transformación logarítmica y comprobado su sistema de referencia, conviene establecer un modelo que rija su comportamiento. Se puede asumir que la conta-

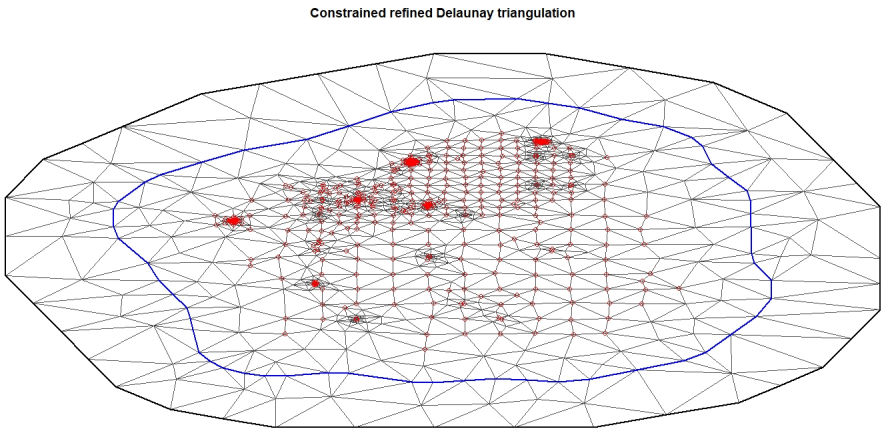

Figura 3.2: Malla de triangulación para construir el modelo SPDE. La zona interior (rodeada en azul) representa una red de triangulación sobre la superficie Gallega que contiene a todos los puntos de muestreo (en rojo). La zona exterior es la malla de triangulación con menor densidad de triángulos (resultaría de utilidad en escenarios en los que se quisiesen realizar predicciones en regiones circundantes a Galicia).

minación atmosférica por cada elemento sigue una distribución Gaussiana con sus respectivas medias μ_i y varianzas σ^2. Además, se decidió que la media μ_i fuese expresada como la suma de un intercepto, β_0, y un efecto aleatorio espacialmente estructurado que seguía un proceso Gaussiano con media cero y función de covarianzas Matérn. Es decir, se propone que los contaminantes obedezcan el comportamiento del modelo (3.2).

A continuación, se construye la malla de triangulación que se muestra en la Figura 3.2. Para ello, se diseña un límite no convexo para las coordenadas muestreadas con la función `inla.nonconvex.hull()`. Dicho límite se introduce en la función `inla.mesh.2d()`, obteniéndose así la malla de triangulación sobre la que se superponen en rojo los puntos de muestreo.

Tras el diseño de la red de triangulación, se procede con la construcción del modelo SPDE sobre la malla, para lo que se usa la función `inla.spde2.matern()`. En el modelado de los datos de contaminantes se ha decidido que el parámetro de suavizado ν sea igual a 1. Dado que $\alpha = \nu + d/2$ y teniendo en cuenta que se trabaja en dos dimensiones ($d = 2$), $\alpha = 2$.

Con el modelo establecido, se generan el set de índices y la matriz de proyección para dicho modelo SPDE siguiendo las instrucciones de la metodología. Llegados a este punto, se procede con el último paso común al modelado de todos los elementos, que es la construcción de una matriz de datos con las localizaciones para la predicción de la concentración del elemento metálico considerado. Primero, se crea una malla de localizaciones de 65×65 utilizando la función `expand.grid()` y combinando los vectores `x` e `y` que contienen las coordenadas en el rango de las localizaciones de las observaciones. A continuación sólo se almacenan las localizaciones que caen dentro del mapa de Galicia (Figura 3.3).

Una vez obtenidos los puntos para las predicciones, se aplica una metodología análoga para cada elemento por separado. El primer paso de esta metodología es el agrupamiento de los datos para la estimación de la predicción, donde el argumento `data` de la función `inla.stack()` será el listado de

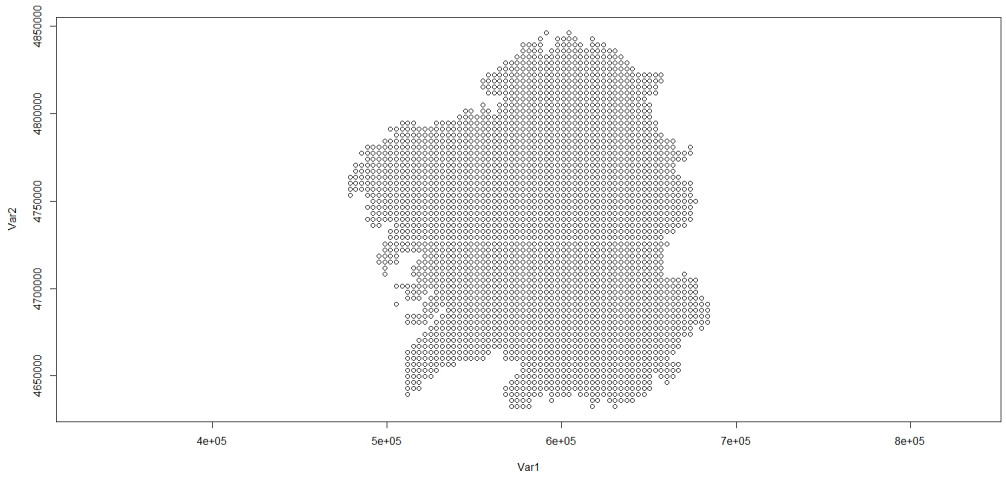

Figura 3.3: Localizaciones para la predicción en Galicia.

concentraciones log-transformadas de cada elemento por separado. Para obtener las predicciones se llama a la función `inla()` con el objeto fórmula pertinente. Con todo ello, se obtienen las predicciones de la media de la log-concentración de los distintos contaminantes con sus respectivos límites. A continuación se muestran los mapas de predicción para el único metaloide del estudio (As), el resto de figuras se adjuntan en el apéndice "A". Todos los mapas de predicción serán explicados en el capítulo 5.

Como se explicó en la metodología, otras representaciones de interés son las medias y desviaciones típicas a posteriori del campo espacial para cada elemento. En este caso también se muestra a continuación el resultado para el As y las figuras para los restantes elementos se pueden consultar en el apéndice "B". Nuevamente, los resultados obtenidos se comentarán de forma detallada en el quinto capítulo.

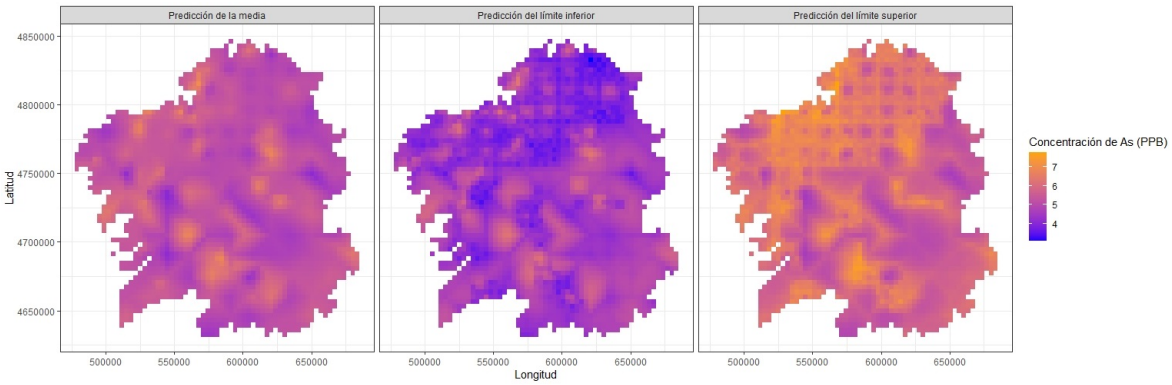

Figura 3.4: Mapas de predicción de la concentración de As. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

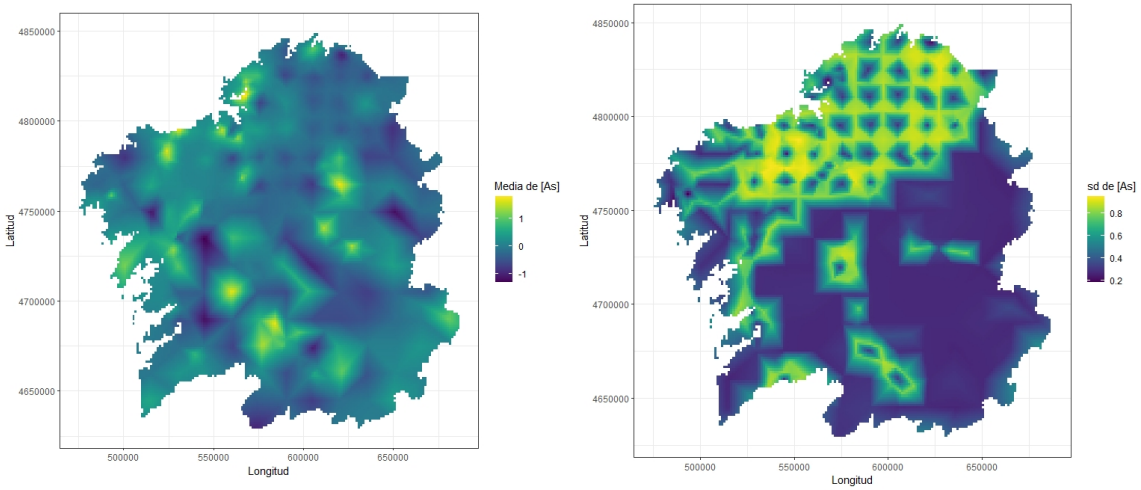

Figura 3.5: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del As.

Capítulo 4

Modelado de datos reticulares

En este capítulo se aborda la implementación de INLA y SPDE en R [26] para el modelado de datos reticulares. Se comenzarán introduciendo las limitaciones del SIR, o su variante SMR, para obtener estimaciones del riesgo de enfermedad o muerte para áreas pequeñas. A continuación, se introducirá el Modelo Besag-York-Mollié (BYM) como una alternativa para mitigar dichas limitaciones. Finalmente, se expondrá una metodología destinada a modelar y mapear los datos reticulares y se procederá a su aplicación a datos de mortalidad por cáncer de pulmón. [21]

4.1. Estimaciones del riesgo de mortalidad en áreas pequeñas

A menudo, los modelos de riesgo de enfermedades tienen como objetivo la obtención de estimaciones del riesgo de enfermedad o mortalidad dentro de las áreas de donde se extraen los datos. Medidas simples del riesgo de enfermedad o mortalidad en las áreas de estudio son las respectivas SIRs y SMRs (2.2), que se definen respectivamente como la proporción de los recuentos de enfermos o muertos observados frente a los esperados. Sin embargo, en muchas situaciones, las áreas pequeñas pueden presentar SIRs o SMRs extremas debido al reducido tamaño poblacional o muestral. En estas situaciones, estas tasas pueden ser engañosas y poco informativas, por lo que sería preferible estimar el riesgo de mortalidad mediante el uso de modelos jerárquicos bayesianos que permitiesen incluir tanto información de áreas vecinas, como información de covariables, de forma que se suavizasen o redujesen los valores extremos.

Los datos reticulares de variables discretas, como es el número de muertes por municipio, obedecen a la siguiente formulación general del modelo espacial:

$$Y_i \sim \text{Po}(E_i\theta_i), \quad i = 1, \ldots, n, \tag{4.1}$$

$$\log(\theta_i) = \alpha + u_i + v_i. \tag{4.2}$$

Esto indica por una parte (4.1), que los conteos observados Y_i en el área i, son modelados utilizando una distribución de Poisson con media $E_i\theta_i$, donde E_i es el conteo esperado y θ_i es el riesgo relativo en el área i. Por otro lado (4.2), también establece que el logaritmo del riesgo relativo, es expresado como la suma de un intercepto, α y efectos aleatorios a tener en cuenta para la variabilidad adicional a la variabilidad de la distribución de Poisson (u_i y v_i). α modela el nivel general del riesgo relativo en la región de estudio, u_i es un efecto aleatorio específico del área i para modelar la dependencia espacial del riesgo relativo y v_i es una componente desestructurada intercambiable que modela el ruido incorrelado, $v_i \sim N(0, \sigma_v^2)$.

El riesgo relativo, θ_i, cuantifica si el área i tiene un riesgo mayor ($\theta_i > 1$) o menor ($\theta_i < 1$) que el riesgo general de la población estándar. Por ejemplo, si $\theta_i = 3$, esto significa que el riesgo del área i es tres veces el riesgo promedio de la población estándar.

Para la formulación del modelo, también es común incluir tanto covariables para cuantificar el factor de riesgo, como otros efectos aleatorios para tratar otras fuentes de variabilidad. Por ejemplo $\log(\theta_i)$ se puede expresar como

$$\log(\theta_i) = d_i\beta + u_i + v_i, \tag{4.3}$$

donde $d_i = (1, d_{i1}, \ldots, d_{ip})$ es el vector del intercepto y p covariables correspondientes al área i, y $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$ es el coeficiente del vector. En este escenario, para un incremento de una unidad en la covariable $d_j, j = 1, \ldots, p$,manteniendo el resto de covariables constantes, el riesgo relativo se incrementa por un factor de $\exp(\beta_j)$.

Un modelo espacial popular que mitiga parte de la problemática de las estimaciones de las SMRs en áreas reducidas es el modelo BYM. En este modelo, se le asigna al efecto aleatorio espacial u_i una distribución condicional autorregresiva (CAR) que suaviza los datos acorde a una determinada estructura de vecinos; por la cual, dos áreas son vecinas si comparten frontera. Más concretamente

$$\left(u_i|u_{-i} \sim N(\bar{u}_{\delta_i}, \frac{\sigma_u^2}{n_{\delta_i}})\right), \tag{4.4}$$

donde $\bar{u}_{\delta_i} = n_{\delta_i}^{-1} \sum_{j \in \delta_i} u_j$. δ_i y n_{δ_i} representan respectivamente el conjunto de vecinos y el número de vecinos de cada área i. La componente desestructurada v_i es modelada como una variable normal independiente e idénticamente distribuida con media cero y varianza σ_v^2.

4.2. Metodología de modelado de datos reticulares

Al igual que para el modelado de los datos geoestadísticos, para el modelado de los datos reticulares, es crucial familiarizarse con la naturaleza y propiedades de los datos que se van a utilizar. Destacar que, además de la importancia de la proyección empleada para la representación de los datos, también resulta de interés la disposición de las subdivisiones del área de estudio.

Matrices de vecinos espaciales

El concepto de matriz de vecino espacial es útil para la exploración de datos de área. El elemento i-ésimo y j-ésimo de una matriz de vecinos espaciales W, denotado por w_{ij}, conecta las áreas i y j de alguna manera, $i, j \in \{1, \ldots, n\}$. W define una estructura de vecinos por toda la región de estudio y sus elementos pueden ser vistos como pesos. Asignándose más peso a las áreas j próximas al área i que a las alejadas. La definición más simple de vecino viene dada por la matriz binaria en la que $w_{ij} = 1$ si la región i y la j comparten alguna frontera,quizás un vértice, y $w_{ij} = 0$ de otro modo.

El desarrollo de la metodología para modelar en R [26] los datos reticulares se puede dividir en tres bloques principales. En primer lugar, el cálculo y representación de las SIRs o SMRs, a continuación el ajuste del modelo BYM y finalmente, la representación de los mapas de riesgo.

4.2.1. Cálculo y representación de la SMR

El primer paso del modelado está ligado al análisis descriptivo. Consiste en cargar y comprobar la estructura de los datos y el mapa. Lo óptimo sería generar objetos *data.frame* con el objeto que contiene el mapa (habitualmente `SpatialPolygons`), los datos del número de muertes observadas frente a las esperadas para cada subdivisión de área de estudio, su cociente (2.2) (SMR) y los valores

de cualquier variable que pueda ser relevante para modelar las SMRs. Destacar también que para poder comparar los resultados del modelado geoestadístico con los del modelado reticular, conviene que se emplee un mismo CRS, por lo que se transformará la proyección del sistema de coordenadas de forma análoga a como se hizo para el modelado de datos geoestadísticos. A excepción del uso de la librería "sf" [25] para la transformación del CRS, el resto de pasos necesarios para generar los *data.frame* son triviales desde el punto de vista de manejo de datos, no se requiere de librerías adicionales.

La fusión de los datos para el modelado con el objeto `SpatialPolygons` da como resultado un objeto `SpatialPolygonsDataFrame`. La fusión de los objetos se puede realizar de diversas formas; una bastante eficiente es usar la función `SpatialPolygonsDataFrame()` a la que se le introducen el mapa y los datos que se quieren fusionar y se le indica que los fusione acorde a su identificador (`match.ID = TRUE`). Los resultados del modelado usando la SMR se pueden representar usando la función `ggplot()`.

4.2.2. Ajuste del modelo BYM

Además del paquete "R-INLA" [17], para el ajuste del modelo BYM también es importante el paquete "spdep" [4].

Partiendo del modelo general (4.1), se crea la matriz de vecinos W utilizando las funciones `poly2nb()` y `nb2INLA()` del paquete "spdep" [4]. Para ello, primero se usa la función `poly2nb()` para crear un listado de vecinos basado en fronteras contiguas. Cada elemento del listado representa un área y contiene los índices de sus vecinos. A continuación, se usa la función `nb2INLA()` para convertir el listado en un archivo con la representación de la matriz de vecinos en el formato que requiere "R-INLA" [17]. Posteriormente, se lee el archivo utilizando la función `inla.read.graph()` del paquete "R-INLA" [17] y se almacena en un objeto que más tarde se usará para establecer el efecto espacial aleatorio.

Inferencia usando INLA

El modelo (4.2) incluye los efectos aleatorios u_i para modelar la variación residual espacial, y v_i para modelar ruido desestructurado. Es necesario incluir dos vectores en los datos que denoten los índices de estos efectos aleatorios. Se puede denominar `idareau` a los índices del vector u_i y `idareav` a los índices del vector para v_i. Se establecen tanto `idareau` como `idareav` igual a $1, \ldots, n$, donde n es el número de particiones de la región de estudio.

Para ajustar el predictor en `inla()`, primero se especifica la fórmula incluyendo la variable respuesta a la izquierda, y añadiendo los efectos fijos y aleatorios a la derecha. Podrían utilizarse covariables para la predicción de la variable respuesta Y_i. Los efectos aleatorios son establecidos usando la función `f()` con el parámetro igual al nombre del índice de la variable y el modelo escogido. Para ajustar u_i con el modelo BYM se usa el argumento `model = "besag"` con la matriz de vecinos W previamente diseñada. Para v_i se suele escoger la distribución de una variable aleatoria independiente e idénticamente distribuida `model = "iid"`.

Finalmente, se ajusta el modelo llamando a la función `inla()`. Utilizando la formulación del modelo (4.1), se indica la fórmula, la familia (`family="poisson"`), los datos y los conteos esperados (`E`). También se establece que el argumento `control.predictor` sea igual a `list(compute = TRUE)` para calcular las predicciones a posteriori.

4.2.3. Representación de los mapas de riesgo

De forma análoga a la representación de las predicciones de los datos geoestadísticos, utilizando la función `ggplot()`, se pueden representar los mapas del riesgo relativo medio con sus intervalos de

credibilidad al 95 %, Para ello, sólo es necesario almacenar en un objeto la salida de la función `inla()`.

4.3. Aplicación a datos reales

Para el modelado de los datos reticulares se parte de varios conjuntos de datos. Por una parte, los mencionados en el análisis descriptivo, que serían el tamaño poblacional por municipio gallego y el número de muertes por cáncer de pulmón en municipios de Galicia con más de 10000 habitantes. Por otro lado, se obtuvo el contenido medio por municipio gallego de radón residencial[1]. También destacar que utilizando las librerías "mapSpain" [10] y "tidyverse" [28] se generó un objeto `sfc_MULTIPOLYGON` en el que está almacenado el mapa de Galicia. Es decir, se tiene la región de estudio (Galicia), subdividida en áreas de estudio (municipios), sobre los que se tienen datos del número de muertes durante el 2010 y concentración media de gas radón.

Dado que el gas radón se asocia como la segunda causa de cáncer de pulmón, siendo la primera causa entre la población que nunca ha sido fumadora, y teniendo en cuenta que Galicia es una zona de alta emisión de gas radón, se consideró acertado incluir su concentración media como una variable de interés para modelar la mortalidad por cáncer de pulmón [3]. A priori se esperaría que las regiones que tuviesen mayor incidencia de cáncer de pulmón fuesen las que presentasen mayor mortalidad, por lo que se consideró la concentración como una variable acertada.

Destacar que se comprobó que existía asociación entre el gas radón en municipios gallegos y la mortalidad por cáncer de pulmón en hombres; pero la asociación no era significativa para mujeres [3]. Sin embargo, este aspecto no resultó relevante para el estudio porque, como se comentó en el análisis descriptivo, la gran mayoría de muertes que se registraron durante el 2010 en Galicia fueron de hombres.

Para comenzar, como se puede ver en la Figura 4.1, el objeto `sfc_MULTIPOLYGON` contiene el mapa de Galicia subdividido en los correspondientes municipios. Antes de su representación, se aseguró que es sistema de coordenadas fuese un sistema proyectado UTM zona 29 Norte, con la función `st_transform( 23029)` de la librería "sf" [25].

A continuación, utilizando los datos del SERGAS y del IGE, se procedió a estimar el número de muertes esperadas acorde a la expresión (2.3). Con el recuento de muertes por municipio suministrados por el SERGAS y el número de muertes esperadas estimado, se procede al cálculo de la SMR por municipio acorde a la expresión (2.2).

Llegados a este punto, se fusionan todos los datos y se puede proceder a las representaciones que se muestran en la Figura 1.7. En ella se ven las tasas de mortalidad estandarizadas calculadas para cada municipio gallego del que había datos.

El modelo BYM es de utilidad en este escenario porque permite incorporar el efecto de la covariable "concentración media de radón residencial" para predecir la mortalidad. Se propone el modelo (4.1), pero en este caso, el logaritmo del riesgo relativo, θ_i, es expresado como

$$\log(\theta_i) = \beta_0 + \beta_1 + u_i + v_i, \tag{4.5}$$

donde β_0 es el intercepto que representa el riesgo general, β_1 es coeficiente de la concentración de gas radón, u_i es una componente espacial estructurada con distribución CAR, $u_i|u_{-i} \sim N(\bar{u}_{\delta_i}, \frac{\sigma_u^2}{n_{\delta_i}})$, y v_i es un efecto espacial desestructurado definido como $v_i \sim N(0, \sigma_v^2)$.

[1] El Laboratorio de Radón de Galicia tiene tablas de medición de las concentraciones medias de radón residencial que se pueden consultar gratuitamente a través de https://radon.gal/radon-en-galicia/tablas-de-medicions/.

Figura 4.1: Mapa de Galicia subdividido en municipios y con proyección UTM zona 29N.

Para poder implementar los efectos aleatorios se crea matriz de vecinos W y a continuación se repiten los mismos pasos que en la metodología general. De este modo, se obtiene el mapa de riesgo relativo de mortalidad por cáncer de pulmón que se muestra en la Figura 4.2, con sus respectivos límites de confianza que se muestran en la Figura 4.3. Estos resultados serán comentados en el siguiente capítulo.

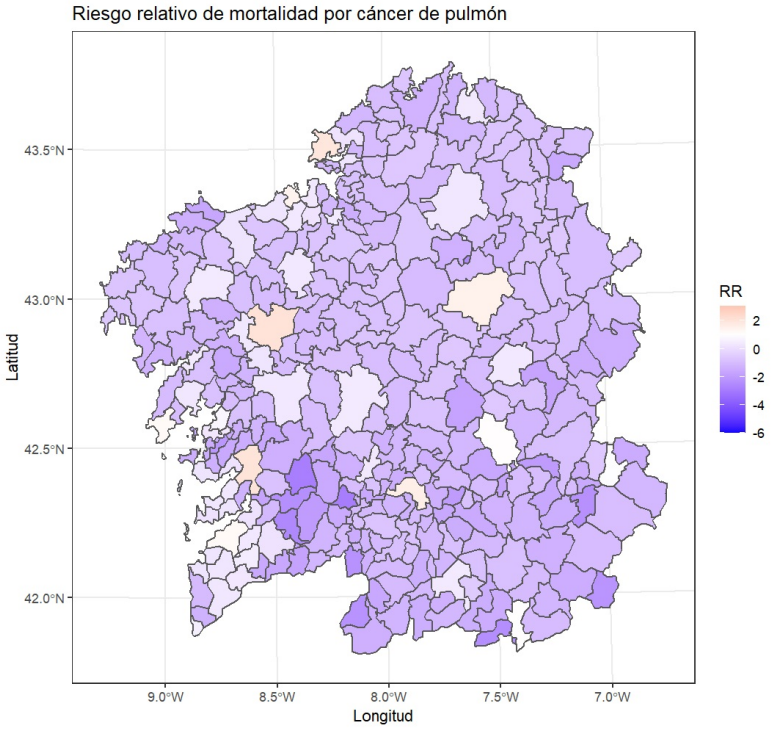

Figura 4.2: Mapa de riesgo relativo medio de muerte por cáncer de pulmón en Galicia.

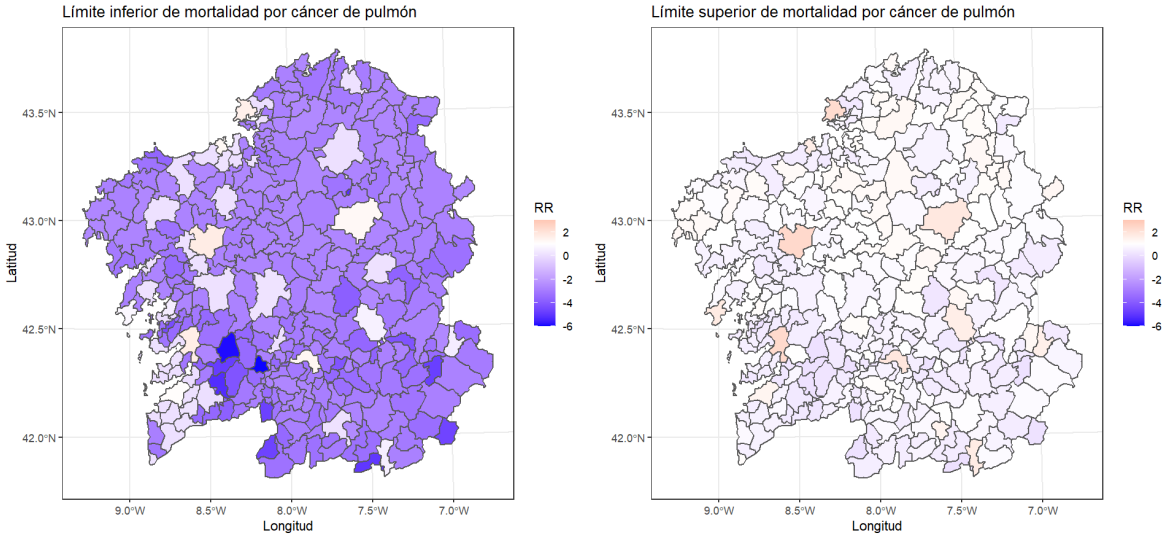

Figura 4.3: A la izquierda el límite inferior del riesgo relativo medio de mortalidad por cáncer de pulmón. A la derecha, el límite superior del riesgo relativo media de mortalidad por cáncer de pulmón.

Capítulo 5

Resultados, problemáticas y conclusiones del estudio

Como se puede inferir del título de este capítulo, éste va a constar de tres bloques centrales. Se comenzará por comentar los resultados obtenidos de forma específica para cada figura. A continuación, se comentarán algunas problemáticas a tener en cuenta del procedimiento empleado para la obtención de estos resultados y finalmente, se expondrán las conclusiones del estudio.

5.1. Resultados del estudio

Se pueden dividir los resultados en dos bloques principales. Por una parte, los mapas de predicciones de concentración de los contaminantes durante el intervalo 2000-2010 y por otro, los de riesgo de mortalidad durante el 2010.

5.1.1. Mapas de contaminación

Para comenzar, hay que destacar que el modelado se hizo en base a los datos de concentración de contaminantes log-transformados (en tejidos de *P. purum*) y por tanto, las predicciones también estarán en una escala log-transformada.

Se puede comenzar por comentar los resultados respecto al único elemento no metálico considerado en el estudio, el metaloide arsénico (As). Como se puede apreciar en la Figura 3.4, las regiones con mayor concentración media coinciden, a su vez, con las regiones con concentraciones más altas para el límite inferior y superior de predicción. También destacar que la predicción media para la log-concentración media de As en Galicia oscila entre 4,5 - 7 ng/g; dándose las concentraciones más altas para algunas de las regiones más industrializadas de la provincia de A-Coruña, como son el entorno de Ferrol, la ciudad de A Coruña, Betanzos, Carballo, los alrededores de la ría Ortigueira, la ría de Arousa o la de Muros y Noia. También conviene recalcar que, la mayoría de la región occidental de Galicia también presenta concentraciones de As mayores a las observadas en el resto del territorio. Además, son llamativas las concentraciones elevadas alrededor de los límites de la provincia de Ourense y la ciudad de Lugo. En la Figura 3.5, se pueden ver las proyecciones de la media y desviación típica a posteriori del campo gaussiano en los nodos de la malla. Se puede apreciar que la media proyectada presenta un comportamiento análogo a la del mapa de predicciones previamente comentado. Un aspecto interesante de la Figura 3.5 reside en el comportamiento de la desviación típica del campo gaussiano. Éste varía de forma característica sobre la superficie gallega, habiendo mayor desviación típica en la región norte y oeste de Galicia. A pesar de ello en estas regiones se ven puntos de baja desviación típica distribuidos

de forma regular; éstos, corresponden a los puntos de muestreo de la red de muestreos de Meirama, As Pontes y Sabón. Finalmente, también destacar que hay algunos puntos de muestreo de industrias del centro y sur de Galicia que presentan una baja desviación típica respecto a la de sus alrededores.

Como se puede observar en la Figura A.1, las predicciones de las log-concentraciones de aluminio (Al), presentan un comportamiento casi idéntico al observado para el As. Sin embargo, hay que destacar que las concentraciones de Al son notablemente superiores a las de As para cualquier punto de la superficie gallega, ya que la escala de medida para la concentración de Al es en $\mu g/g$, mientras que para el As era en ng/g. Algunas diferencias respecto al comportamiento del gradiente de concentraciones predicho para el As reside en que para el Al hay mayores concentraciones en el sur de la Provincia de Ourense y en que hay unas concentraciones más bajas en los alrededores de la ciudad de Lugo, a pesar de que su concentración es más alta en la provincia en general. En la Figura B.1 se puede ver que el comportamiento a posteriori de la media y desviación típica del campo gaussiano proyectado es muy similar al observado para el As, pero en este, caso las medias y desviaciones típicas son inferiores a las observadas para el As.

Los mapas de predicción de la Figura A.2 muestran un cambio en los gradientes de concentración predichos para el cadmio (Cd) respecto a los anteriores dos contaminantes. En este caso, se puede ver una concentración más o menos homogénea por toda la superficie gallega, con unos niveles más altos en la región más nórdica (especialmente alrededores de Ferrol), alrededor de las rías de la región occidental y en la zona occidental de la provincia de Ourense. También cabe resaltar que a diferencia de las predicciones para el As y el Al, en el caso del Cd, no son tan distantes las predicciones del límite superior e inferior. En este caso, la log-concentración media de Cd oscila entre 4,5 - 5,5 ng/g. Las proyecciones de la media y la desviación típica del campo gaussiano para el Cd (Figura B.2) también cambian respecto a las del As y el Al. En este caso, la región del norte sigue teniendo niveles medios más altos (especialmente en Ferrol y alrededores). La desviación típica máxima vuelve a disminuir, aunque las regiones con mayor desviación típica siguen manteniendo una disposición similar a la de los anteriores elementos; nuevamente, en las áreas circundantes a los puntos de muestreo previamente mencionados hay una desviación típica menor a la de otras regiones próximas.

En los mapas de predicción de la Figura A.3 se muestran las predicciones de las log-cooncentraciones de cromo (Cr) en Galicia. En este caso, la predicción de la media oscila entre los 1,5-3 $\mu g/g$. En este escenario, aunque los gradientes de coloración parezcan indicar grandes diferencias entre las predicciones del límite inferior y superior, si se tiene en cuenta la escala, el grado de variación es pequeño. Indicar que el Cr, a excepción de los niveles elevados en regiones aisladas de la provincia de A Coruña y el sur de Ourense, presenta unos gradientes similares por toda la superficie gallega. En la Figura B.3 se ven la media y la desviación típica del campo gaussiano subyacente al comportamiento de la concentración de Cr. En cuanto a la media, se tiene un campo coherente a las predicciones y en cuanto al comportamiento de la desviación típica, se puede apreciar que es análogo al comportamiento del As y el Al.

Con las predicciones para el cobre (Cu) (Figura A.4) se acentúa lo que sucedía con el Cr, siendo menos numerosas las regiones con concentraciones superiores a lo observado para la comunidad autónoma en su conjunto (nuevamente gradientes más altos se dan el la región noroeste de Galicia). Al igual que sucedía con el Cr, las diferencias entre los límites superior e inferior de predicción no son muy grandes. La predicción de la concentración media oscila entre 1,5-3 $\mu g/g$. En la Figura B.4 se ve que, a excepción de zonas aisladas, la media del campo gaussiano proyectado es bastante homogénea por toda la superficie de Galicia y que las mayores desviaciones típicas se dan alrededor de las mismas regiones que para los anteriores elementos, siendo especialmente baja en la mayor parte del centro y suroeste gallego.

De todos los contaminantes atmosféricos considerados en este estudio, el hierro (Fe), es el más abundante (Figura A.5); oscilando las predicciones de su concentración media entre 5,5-7,5 $\mu g/g$. Se

puede observar que, aunque las áreas en las que abundaban el resto de contaminantes también presentan una mayor concentración de Fe, el elemento es bastante abundante para toda la superficie gallega. Destacar que las Provincias de Coruña, el noreste de Lugo y el sur de Pontevedra y Ourense son las regiones con niveles más altos de Fe. De la Figura B.5 destacar que el comportamiento de la media del campo era el esperado, con una media especialmente alta alrededor de Cee. Destacar que la estructura de las desviaciones típicas se mantiene relativamente constante.

Respecto a las predicciones de las concentraciones de mercurio (Hg) (Figura A.6), destacar que hay bajos niveles medios de concentración para toda la superficie de Galicia (resulta relevante observar que la escala de medida de los datos era en ng/g).También resaltar que las diferencias absolutas entre la concentración media y los límites superior e inferior son muy pequeñas y que únicamente hay algunos focos con niveles superiores en el note de las provincias de A coruña y Lugo. En la Figura B.6 se puede ver que la media del campo gaussiano tiene un comportamiento acorde a las predicciones y que las desviaciones aunque siguen manteniendo una estructura similar a la observada para los restantes elementos, pasa a ser mínima (casi nula), para la mayoría de la superficie de Galicia.

En cuanto a las predicciones para el níquel (Ni) (Figura A.7), se observan concentraciones medias que oscilan entre 1,5-2,5 $\mu g/g$. Al igual que para algunos de los contaminantes previamente comentados, la mayoría de los focos de concentración elevada de Ni están ubicados en la provincia de A Coruña. La media y desviación típica del campo proyectado en la Figura B.7 muestran el comportamiento esperado teniendo en cuenta las predicciones obtenidas. Por una parte, una media de las concentraciones relativamente uniforme en la comunidad autónoma pero, con algunas regiones con valores elevados, como los alrededores de Cee y Ferrol. Destacar que la estructura general de las desviaciones típicas del campo se mantiene con pocas variaciones (la región norte y oeste de Galicia son las que presentan mayores desviaciones típicas).

Finalmente, se van a comentar los resultados de las predicciones del último elemento considerado para el estudio, el plomo (Pb) (Figura A.8). Al igual que sucedía con el Hg, los niveles medios de plomo en Galicia son reducidos (6-7 ng/g), con poca oscilación absoluta entre los límites inferior y superior de predicción y con los niveles más altos situados en la región noroeste de la comunidad. También destacar que, a excepción de los alrededores de Ferrol, no resulta sorprendente ver una concentración prácticamente uniforme de la media que se muestra en la Figura B.8. Se puede ver también que aunque las mayores varianzas se observan en la región norte y oeste de una forma similar a lo observado para los restantes elementos (en este escenario hay que tener en cuenta que son mayores las diferencias entre las regiones con mayor y menor desviación típica).

5.1.2. Mapas de riesgo de mortalidad

Por una parte están las medidas de mortalidad medias según la SMR (2.2) expuestas en el análisis descriptivo y por otro lado están los resultados que se muestran en el anterior capítulo (mapas de riesgo de mortalidad).

Mientras que en la Figura 4.2 se muestra el riesgo relativo medio de muerte por cáncer de pulmón en municipios de Galicia durante el 2010, en la Figura 4.3 se muestran el límite inferior y superior de los riesgos relativos durante ese año. No existen grandes discrepancias entre el mapa de riesgo relativo medio (Figura 4.2) y el mapa de tasas de mortalidad estandarizadas (Figura 1.7). En ambos se puede apreciar que los municipios de Ferrol, A Coruña, Santiago de Compostla, Lugo, Ribeira, Pontevedra y Ourense presentan una mortalidad superior a la esperada. Los municipios de la provincia de Pontevedra semejan ser los que tienen un riesgo relativo medio de mortalidad más bajo y en base a los mapas se puede intuir que de forma general los riesgos de mortalidad por cáncer de pulmón en Galicia eran bajos durante el 2010. Si se analizan los límites inferior y superior de la Figura 4.3, uno se puede dar cuenta

que varían ligeramente los riesgos relativos de los municipios, pero de forma general los municipios de mayor riesgo siguen siendo los mismos.

5.2. Problemáticas del estudio

En esta sección se abordarán algunas de las problemáticas a tener en cuenta para sacar conclusiones acertadas a partir de los resultados previamente comentados.

En primer lugar, hay que tener en cuenta que los datos geoestadísticos en su conjunto no han sido recolectados de forma regular ni en el tiempo ni en el espacio. Es decir, aunque existe una malla muestreada de forma regular durante el intervalo 2004-2010, el muestreo en torno a las distintas industrias se realizó de forma aislada en otros instantes de tiempo dentro del intervalo 2000-2010. En definitiva, para el ajuste se ha asumido que no se han producido grandes cambios en las emisiones por parte del tejido industrial gallego durante la primera década del siglo XXI. Es por ello que nos permitimos analizar de manera simultánea todas las observaciones recogidas en los distintos años.

Por otro lado hay que resaltar que las SMRs estimadas en el análisis descriptivo son más fiables que los resultados del modelo desarrollado en el anterior capítulo. Ello se debe a que se usaron sólo datos de municipios con más de 10000 habitantes, de modo que las SMR obtenidas son suficientemente fiables y no hay necesidad de aplicar suavizados para corregir posibles valores extremos por tamaños muestrales reducidos (como se hizo en el capítulo 4). También destacar que, al utilizar datos de municipios con más de 10000 habitantes, son pocos los municipios de los que hay datos (únicamente hay datos para 50 municipios) y por tanto, aplicar procesos de suavizado y predicción con tan pocos datos ofrece unos resultados poco informativos. Para obtener resultados más informativos habría que solicitar nuevamente datos al SERGAS, pero es un procedimiento lento y costoso.

Finalmente, es destacable una problemática que se suele dar con el análisis de datos reticulares y es el Problema de la Unidad de Área Modificable (MAUP), por la cual las conclusiones a las que se llega pueden cambiar si uno agrega los mismos datos en un nivel de agregación espacial distinto; por ejemplo si los mismos datos que se han empleado para este TFM se agregasen por provincias, es posible que las conclusiones cambiaran. El MAUP consiste en dos efectos interrelacionados. El primer efecto es la escala o "efecto de la agregación", que hace referencia a las diferentes inferencias obtenidas cuando los mismos datos son agregados en áreas cada vez más grandes. El segundo efecto es el efecto del agrupamiento o "efecto de zonación", por el cual hay una variabilidad de resultados debida a formas alternativas de las áreas que conducen a diferencias en la forma del área en la misma escala o en escalas similares [21].

5.3. Conclusiones del estudio

Asumiendo que no hubo grandes cambios en las emisiones por parte del tejido industrial gallego durante la primera década del siglo XXI y utilizando las variaciones en concentración de los distintos metales en tejidos de *P. purum* como medida de las variaciones en concentración de metales a nivel atmosférico, se puede concluir que, aunque las concentraciones de los distintos elementos variaron (siendo el As, Cd, Hg y Pb los elementos con concentraciones más bajas), el comportamiento general de los contaminantes fue similar; lo cual es lógico teniendo en cuenta que son emitidos principalmente por el tejido industrial.

Los contaminantes se pueden dividir en tres bloques atendiendo a la distribución de sus gradientes de concentración. Por una parte estaría la pareja compuesta por el Hg y el Pb, para los que no hubo unos gradientes de concentración media especialmente abundantes en ninguna parte de la región gallega, salvando algunos puntos de la zona norte y noroeste. Por otro lado estaría el Cd, elemento para

el que la región norte de la provincia de Lugo, junto al noroeste de A Coruña y Ourense presentaron las concentraciones más altas. Un fenómeno similar se produjo para el Fe, que a excepción del sureste de Lugo y norte de Ourense y Pontevedra, tuvo unas concentraciones elevadas. Finalmente, habría un grupo más heterogéneo compuesto por una serie de contaminantes (As, Al, Cr, Cu y Ni) cuyas concentraciones fueron más elevadas en el sureste y oeste de Ourense, la región occidental y norte de Galicia y el oeste de la Provincia de Lugo.

Por otro lado, respecto a los datos de mortalidad por cáncer de pulmón, no se deberían sacar más conclusiones a parte de lo ya comentado respecto a la Figura 4.2. Es decir, que los municipios de Ferrol, A Coruña, Santiago de Compostla, Lugo, Ribeira, Pontevedra y Ourense presentaron una mortalidad superior a la esperada durante el 2010.

Respecto a la posible asociación cualitativa, resulta evidente que el mayor riesgo de mortalidad por cáncer de pulmón se dio en regiones con concentraciones elevadas para la mayoría de elementos metálicos (claro ejemplo de los municipios de Ferrol, A coruña y Ourense). Pero para sacar conclusiones fiables y con menor variabilidad habría que repetir el estudio con los datos de todos los municipios posibles (sin importar que tengan menos de 10000 habitantes).

Apéndice A

Mapas de predicción

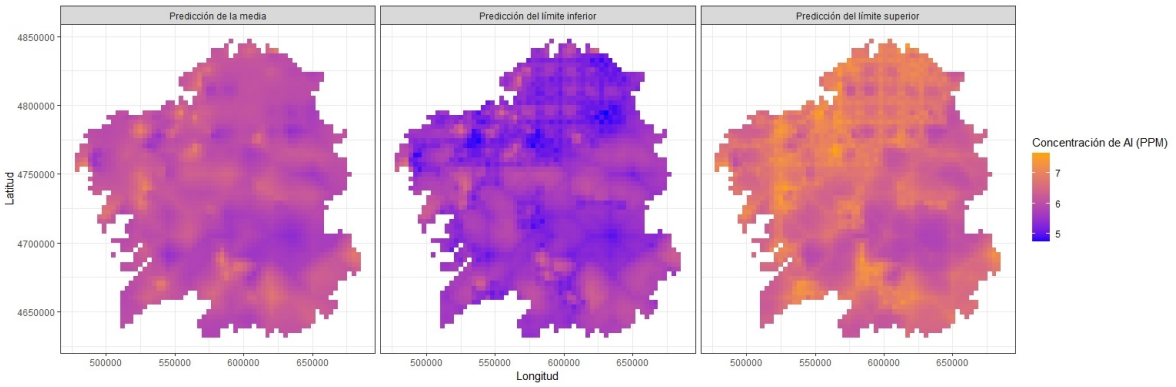

Figura A.1: Mapas de predicción de la concentración de Al. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

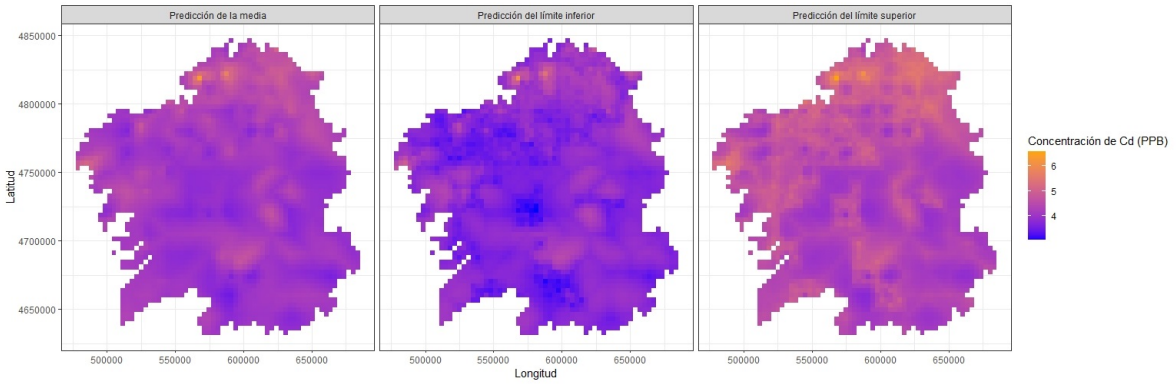

Figura A.2: Mapas de predicción de la concentración de Cd. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

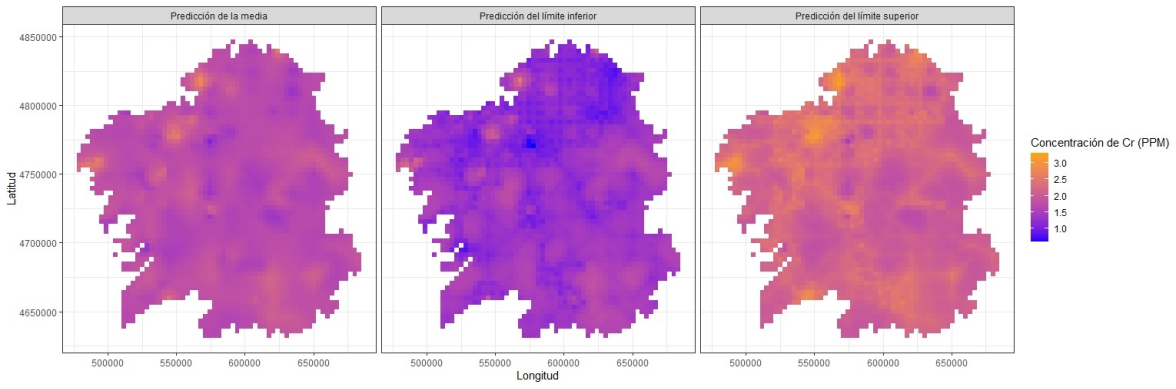

Figura A.3: Mapas de predicción de la concentración de Cr. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

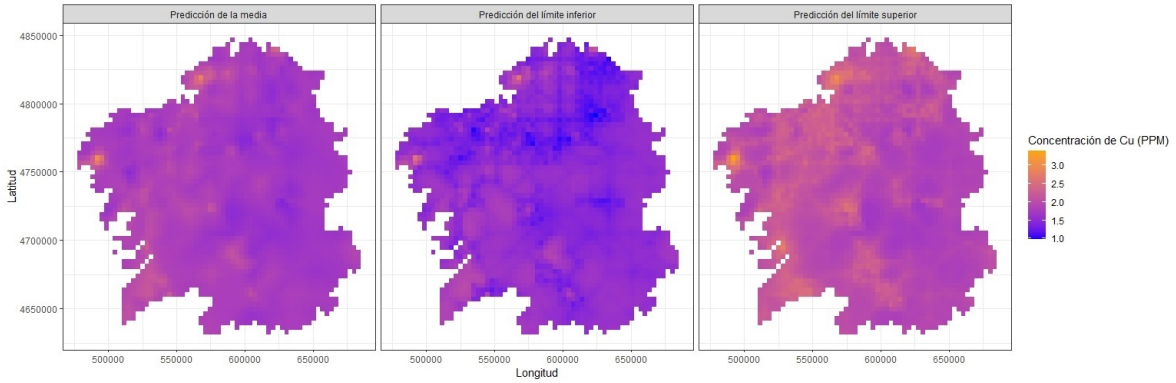

Figura A.4: Mapas de predicción de la concentración de Cu. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

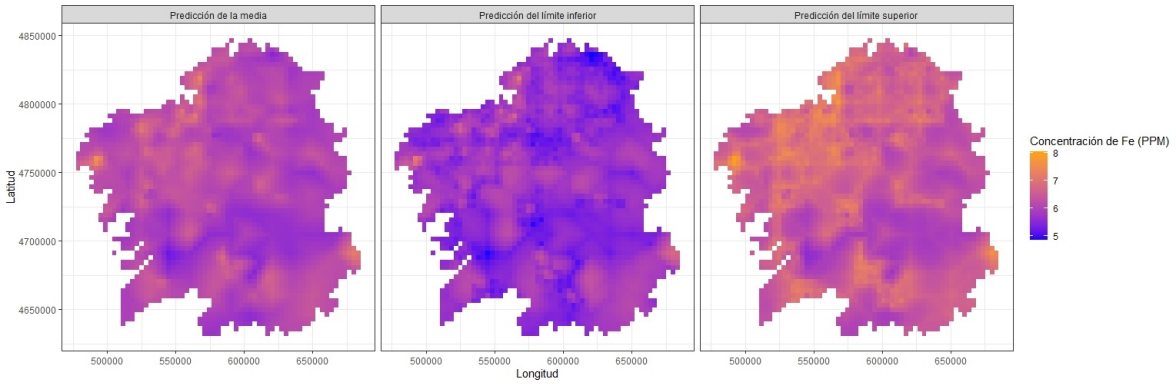

Figura A.5: Mapas de predicción de la concentración de Fe. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

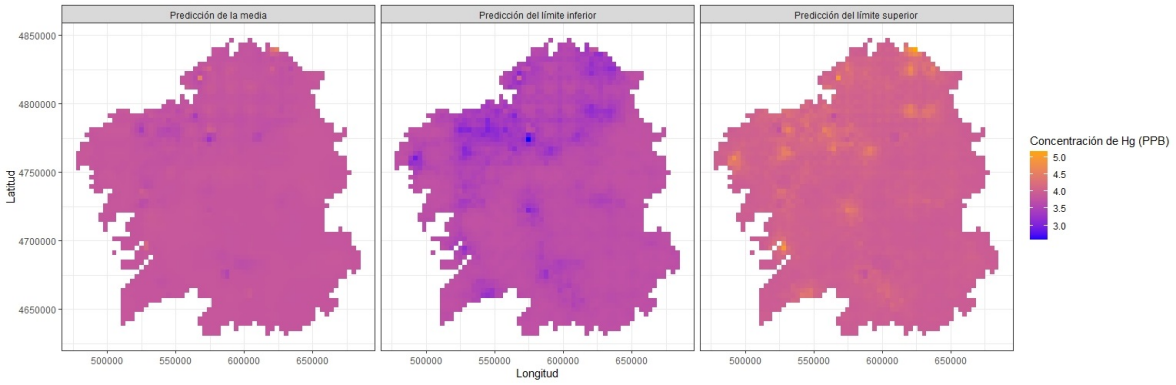

Figura A.6: Mapas de predicción de la concentración de Hg. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

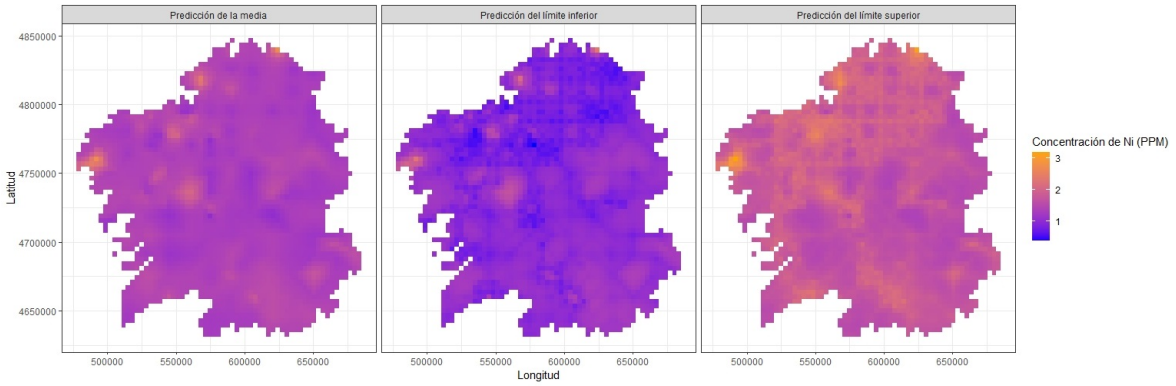

Figura A.7: Mapas de predicción de la concentración de Ni. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

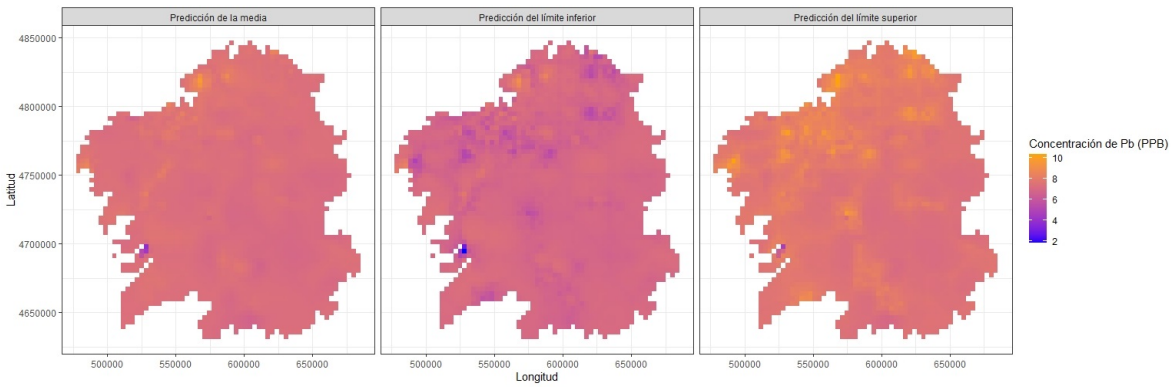

Figura A.8: Mapas de predicción de la concentración de Pb. A la izquierda la concentración media, en el centro el límite inferior (cuantil 2,5 de los valores ajustados) y a la derecha el límite superior (cuantil 97,5 de los valores ajustados).

Apéndice B

Mapas de proyección de la media y desviación típica del campo gaussiano

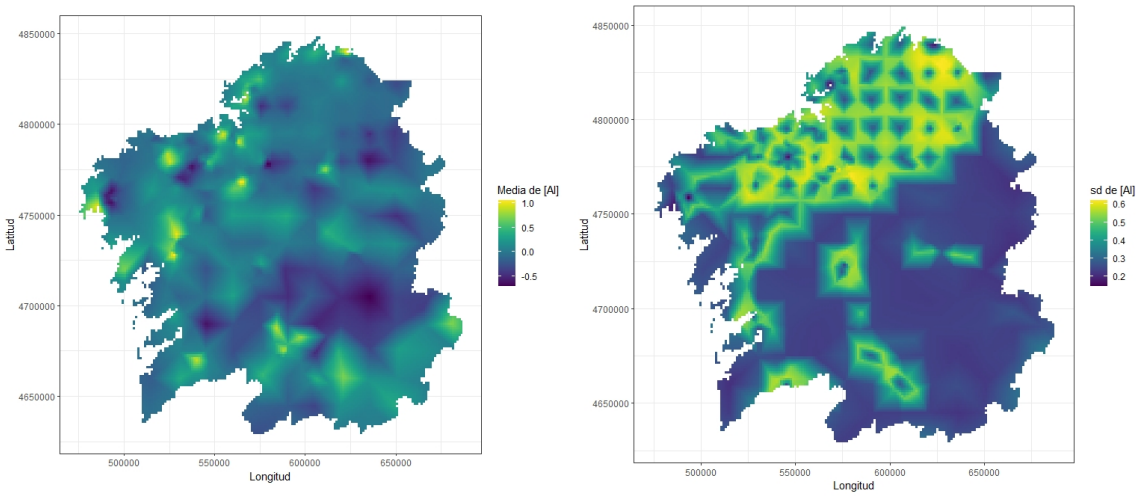

Figura B.1: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Al.

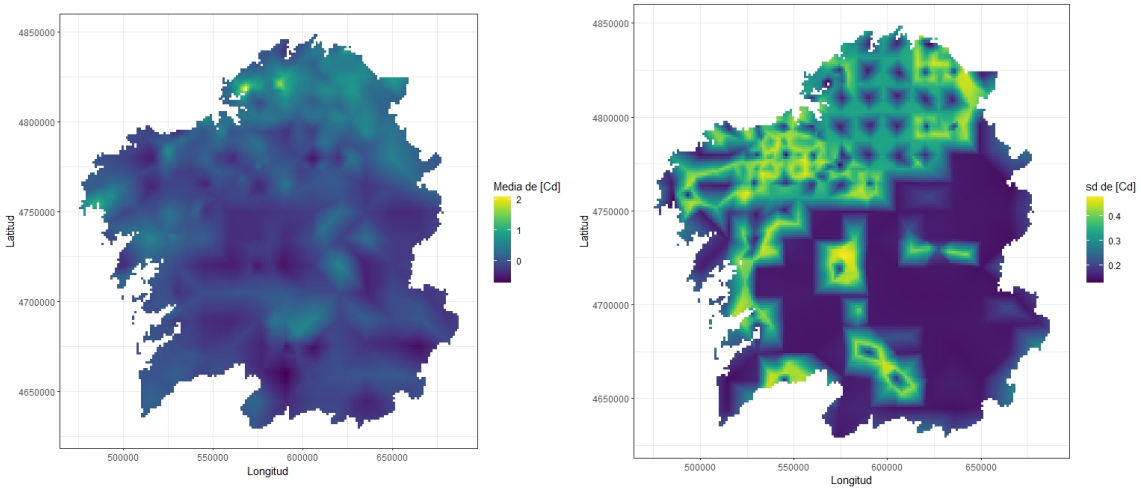

Figura B.2: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Cd.

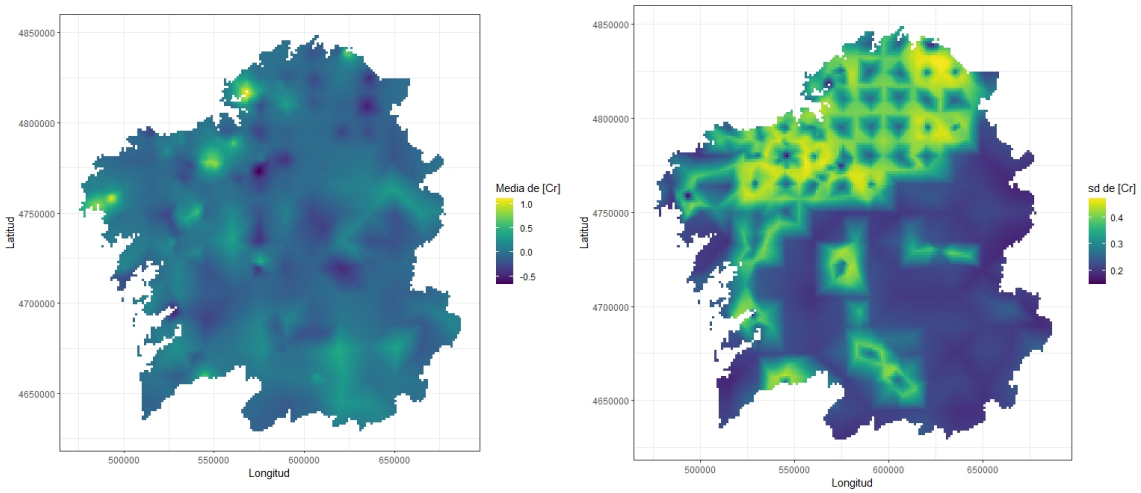

Figura B.3: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Cr.

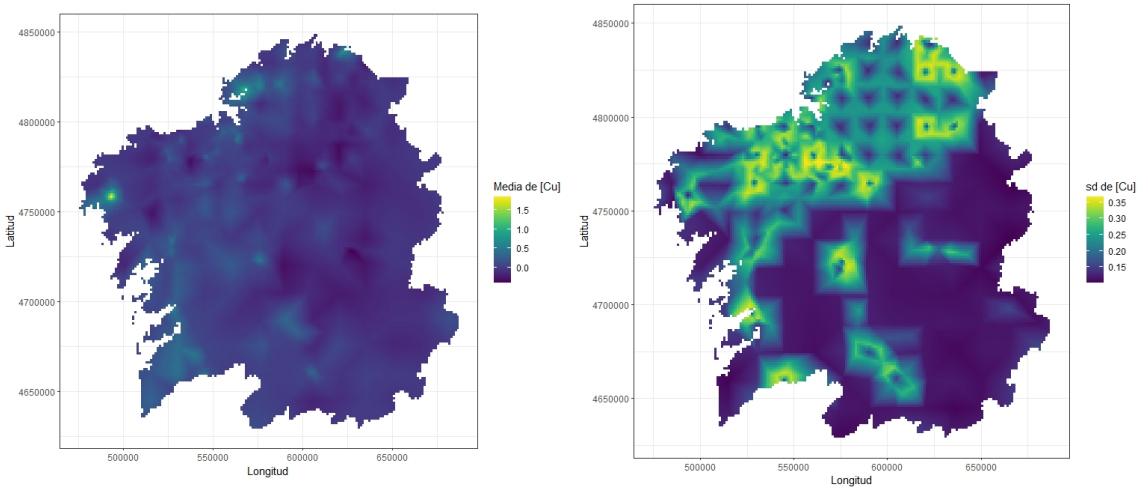

Figura B.4: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Cu.

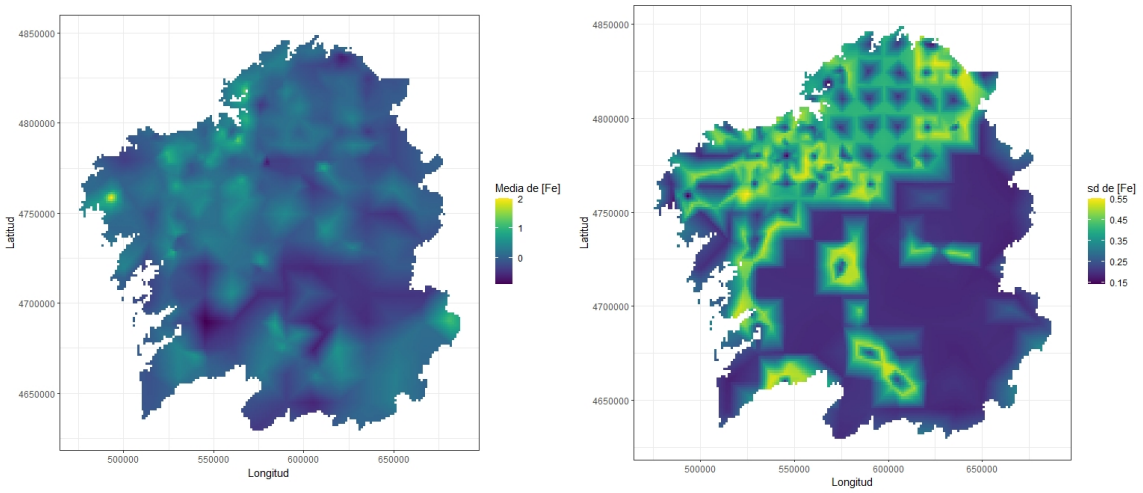

Figura B.5: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Fe.

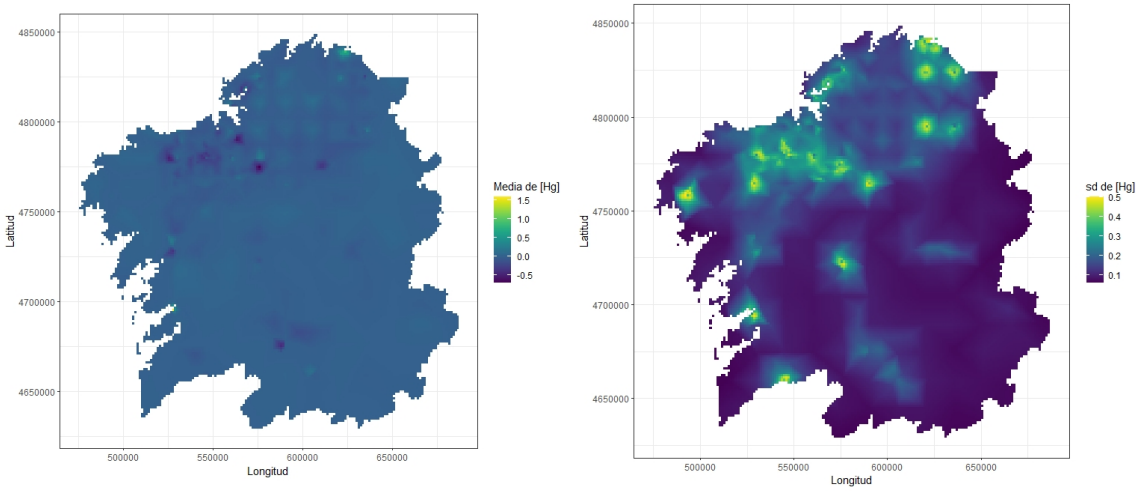

Figura B.6: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Hg.

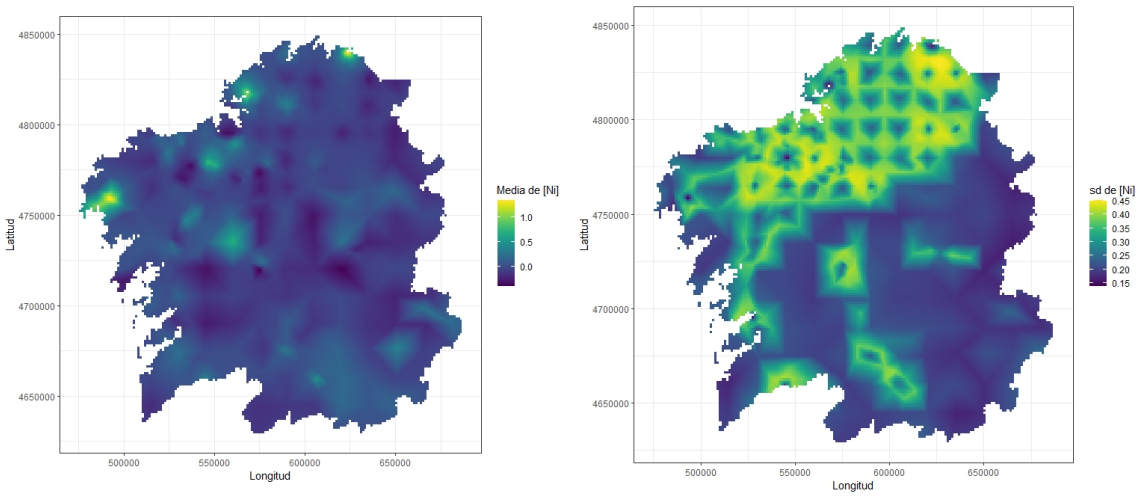

Figura B.7: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Ni.

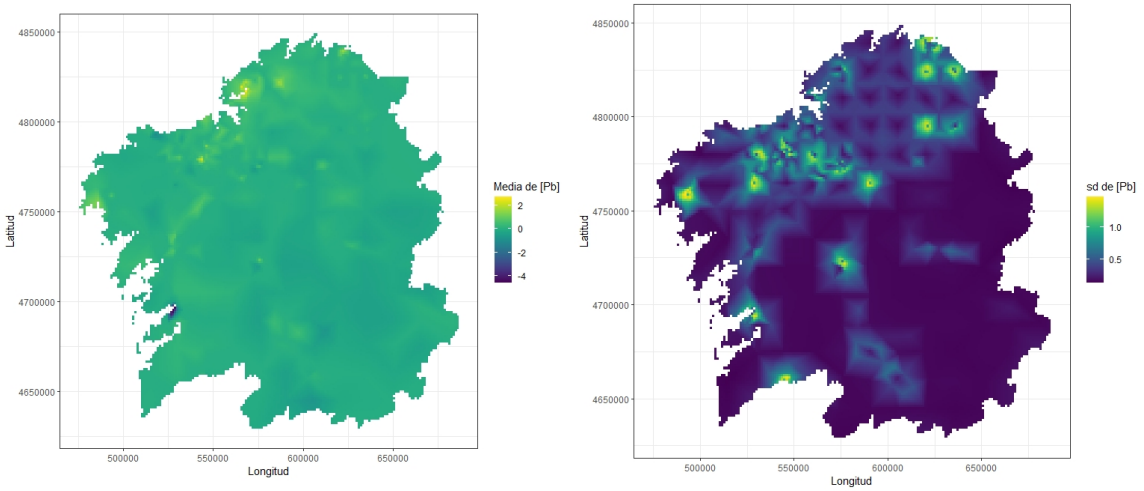

Figura B.8: Proyección en una malla de la media y desviaciones típica a posteriori del campo espacial del Pb.

Bibliografía

[1] 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.

[2] Baptiste, A. (2017) gridExtra: miscellaneous functions for "grid"graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra. Accedido el 31 de mayo de 2022.

[3] Barbosa- Lorenzo, R., Ruano-Ravina, A., Cerdeira-Caramés, S., Barros-Dios, J. M. (2015) Residential radon and lung cancer. An ecologic study in Galicia, Spain. *Medicina Clinica*, 144(7): 304-308.

[4] Bivand, R. (2022) spdep: R packages for analyzing spatial data-A comparative case study with areal data geographical analysis. R package version 1.2.4 https://doi.org/10.1111/gean.12319. Accedido el 31 de mayo de 2022.

[5] Comess, S., Donovan, G., Gatziolis, Deziel, N. C. (2021) Exposure to atmospheric metals using moss bioindicators and neonatal health outcomes in Portland, Oregon. *Enviromental Pollution*, 284: 117343.

[6] 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.

[7] Garnier, S., Ross, N., Rudis, R., Camargo, A. P., Sciaini, M., Scherer, C. (2021) viridis: colorblind friendly color maps for R. R package version 0.6.2. https://sjmgarnier.github.io/viridis/. Accedido el 31 de mayo de 2022.

[8] Gavrilov, I., Pusev, R. (2014) normtest: tests for normality. R package version 1.1. https://CRAN.R-project.org/package=normtest. Accedido el 31 de mayo de 2022.

[9] Gross, J., Ligges, U. (2015) nortest: tests for normality. R package version 1.0-4. https://CRAN.R-project.org/package=nortest. Accedido el 31 de mayo de 2022.

[10] Hernangómez, D. (2022) mapSpain: administrative boundaries of Spain. R package version 0.6.1. https://ropenspain.github.io/mapSpain/. Accedido el 31 de mayo de 2022.

[11] Hijmans, R. J. (2022) raster: geographic data analysis and modeling. R package version 3.5-15. https://CRAN.R-project.org/package=raster. Accedido el 31de mayo de 2022.

[12] INE (2010) Poboación segundo sexos e grupos quinquenais de idade. https://www.ige.gal/igebdt/esqv.jsp?ruta=verTabla.jsp?OP=1&B=1&M=&COD=100&R=9913[all]&C=1[all];0[2010]&F=2:0&S=&SCF=. Accedido el 18 de enero de 2022.

[13] Komsta, L., Novomestky, F. (2022) moments: moments, cumulants, skewness, kurtosis and related tests. R package version 0.14.1. https://CRAN.R-project.org/package=moments. Accedido el 31 de mayo de 2022.

[14] Kurina-Franca, G. (2021) *Modelos Bayesianos para Datos Geoestadísticos. Mapeo Digital de Suelos con R-INLA*. Tesis, Universidad Nacional de Córdoba.

[15] Lequy, E., Siemiatycki, J., Leblond, S., Meyer, C., Zhivin, S.. Vienneau, D., Hoogh, K., Goldber, M., Zins, M., Jacquemin, B. (2019) Long-term exposure to atmospheric metals assessed by mosses and mortality in France. *Enviroment International*, 129: 145-153.

[16] Li, L., Zheng, B., Liu, L. (2010) Biomonitoring and Bioindicators Used for River Ecosystems: Definitions, Approaches and Trends. *Procedia Enviromental Sciences*, 2: 1510-1524.

[17] Lindgren, F., Rue, H., Lindstrom, J. (2011) An explicit link between gaussian fields and gaussian Markov random fields: the stochastic partial differential equation approach (with discussion). *Journal of the Royal Statistical Society B*, 73(4): 423-498.

[18] Machado, A., García, N., García, C., Acosta, L., Córdova, A., Linares, M., Giraldoth, D., Velásquez, H. (2008) Contaminación por metales (Pb, Zn, Ni y Cr) en aire, sedimentos viales y suelo en una zona de alto tráfico vehicular. *Revista Internacional de Contaminación Ambiental*, 24(4): 171-182.

[19] Méndez-Valderrey, J. L. Pseudoscleropodium purum. ISSN 1887-5068. https://www.asturnatura.com/especie/pseudoscleropodium-purum.html. Accedido el 31 de mayo de 2022.

[20] Monsalve Graterol, N. C. (2013) *Modelos Jerárquicos Bayesianos Espaciales en Epidemiología Agrícola*. Tesis, Universitat de Valéncia.

[21] Moraga, P. (2019) *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny*. Chapman & Hall/CRC Biostatistics Series.

[22] Organización Mundial de la Salud (2021) El radón y sus efectos sobre la salud. https://www.who.int/es/news-room/fact-sheets/detail/radon-and-health. Accedido el 30 de mayo de 2022

[23] Pebesma, E.J., Bivand R. S. (2005) Classes and methods for spatial data in R. R package version 1.4.5. https://cran.r-project.org/doc/Rnews/. Accedido el 30 de mayo de 2022.

[24] Pebesma, E. (2021) lwgeom: bindings to selected 'liblwgeom' functions for simple features. R package version 0.2-8. https://CRAN.R-project.org/package=lwgeom. Accedido el 30 de mayo de 2022.

[25] Pebesma, E. (2018) Simple features for R: standardized support for spatial vector data. R package version 1.0.7. https://doi.org/10.32614/RJ-2018-009. Accedido el 30 de mayo de 2022.

[26] R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing. Vienna,Austria. https://www.R-project.org/. Accedido en 30 de mayo de 2022.

[27] Schlutow, A., Schröder, W., Nickel, S. (2021) Atmospheric deposition and element accumulation in moss sampled across Germany 1990-2015: trends and relevance for ecological integrity and human health. *Atmosphere*, 12:193.

[28] Wickham, H., Averick, M., Bryan, J., Chang, W., D'Agostino-McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Milton-Bache, S., Müller, K., Ooms, J., Robinson, D., Paige-Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H. (2019) tidyverse: welcome to the tidyverse. R package version 1.3.1.https://doi.org/10.21105/joss.01686. Accedido el 31 de mayo de 2022.

[29] Wickham, H., François, R., Henry, L., Müller, K. (2022) dplyr: a grammar of data manipulation. R package version 1.0.9. https://CRAN.R-project.org/package=dplyr. Accedido el 31 de mayo de 2022.

[30] Wickham, H. (2016) ggplot2: elegant graphics for data analysis. R package version 3.3.6. https://ggplot2.tidyverse.org. Accedido el 31 de mayo de 2022.

[31] Wickham, H. (2019) stringr: simple, consistent wrappers for common string operations. R package version 1.4.0. https://CRAN.R-project.org/package=stringr. Accedido el 31 de mayo de 2022.

[32] Wilke, C. O. (2020). cowplot: streamlined plot theme and plot annotations for 'ggplot2'. R package version 1.1.1. https://CRAN.R-project.org/package=cowplot. Accedido el 31 de mayo de 2022.

[33] Wolterbeek, H. Th., Verburg, T. G. (2003) Atmospheric metal deposition in a moss data correlation study with mortality and disease in the Netherlands. *The Science of Total Enviroment*, 319: 53-64.