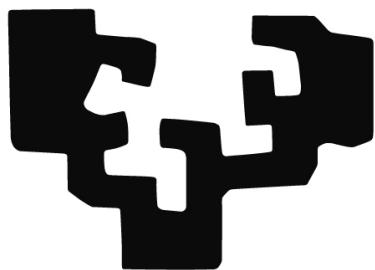


eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea

Máster Universitario en Modelización e  
Investigación Matemática, Estadística y  
Computación 2021/2022

*Trabajo Fin de Máster*  
**Modelos Espaciales y  
Espaciotemporales en Disease  
Mapping**

---

Ander Bodegas Díez

Tutores  
Aritz Adin Urtasun  
Jaione Etxeberria Andueza  
Leioa, septiembre de 2022

## **Resumen**

El presente Trabajo Fin de Máster tiene por objeto realizar una revisión de la representación cartográfica de enfermedades o *disease mapping*, en inglés, el campo de la epidemiología espacial que se centra en la estimación de la distribución geográfica de una enfermedad y su evolución en el tiempo. Se describirán tanto las medidas clásicas de estimación del riesgo como algunos de los modelos espaciales y espaciotemporales más utilizados en la literatura. Estos modelos se explicarán desde el punto de vista teórico, enmarcándolos dentro del ámbito de la estadística Bayesiana, un enfoque basado en el Teorema de Bayes que permite obtener distribuciones de los parámetros del modelo condicionado a los datos. La aplicabilidad de los mismos se ilustrará empleando datos reales de cáncer de pulmón en Gran Bretaña en el periodo 2002-2019. Finalmente, se realizará un estudio comparativo entre las tasas de incidencia y mortalidad por cáncer de pulmón en hombres y mujeres, desvelando así patrones de riesgo espaciales y temporales que pueden ser de gran interés en el ámbito de la salud pública. Todo esto se llevará a cabo empleando la técnica INLA (siglas en inglés de *integrated nested Laplace approximations*), un algoritmo determinista para inferencia Bayesiana que permite obtener resultados precisos en un tiempo menor que los métodos de simulación basados en algoritmos MCMC (*Markov chain Monte Carlo*). La inferencia y estimación de los modelos se realizará utilizando el paquete R-INLA, a través del software estadístico libre R.

## **Abstract**

The purpose of this Master's Thesis is to review the techniques used in Disease Mapping, the field of spatial epidemiology focused on the estimation of the geographical distribution of a disease and its evolution in time. Both classical risk estimation measures and some of the most typically used spatial and spatio-temporal models will be described. These models will be explained from a theoretical point of view, framing them within the scope of Bayesian statistics, an approach based on Bayes' Theorem that allows to obtain distributions of the model's parameters conditioned to the data. The applicability of these models will be illustrated using real lung cancer data in Great Britain in the period 2002-2019. Finally, a comparative study between lung cancer incidence and mortality rates in men and women will be carried out, revealing spatial and temporal risk patterns that may be interesting in the field of public health. All this will be carried out using the INLA (integrated nested Laplace approximations) technique, a deterministic algorithm for Bayesian inference that allows to obtain accurate results in a shorter time than other simulation-based methods as MCMC (Markov chain Monte Carlo) algorithms. Model fitting and inference will be carried out using the R-INLA package, through the free statistical software R.

# Índice general

<b>1. Introducción</b>	<b>1</b>
<b>2. Marco teórico</b>	<b>3</b>
2.1. <i>Disease mapping</i> . . . . .	3
2.2. Medidas clásicas de estimación de riesgo . . . . .	5
2.2.1. Método de estandarización directo . . . . .	6
2.2.2. Método de estandarización indirecto . . . . .	6
2.3. Estadística Bayesiana . . . . .	8
2.3.1. Inferencia Bayesiana . . . . .	9
2.4. Modelos jerárquicos . . . . .	11
2.5. Modelos espaciales . . . . .	14
2.5.1. Datos de área . . . . .	14
2.5.2. Matriz de precisión e independencia condicionada . . . . .	16
2.5.3. Modelos espaciales para el <i>disease mapping</i> . . . . .	20
2.5.4. Modelo CAR intrínseco . . . . .	21
2.5.5. Modelo BYM . . . . .	22
2.5.6. Modelo de Leroux . . . . .	22
2.6. Modelos espaciotemporales . . . . .	24

2.6.1.	Modelos que incorporan estructura temporal	24
2.6.2.	Matriz de precisión e independencia condicionada	25
2.6.3.	Modelos espaciotemporales para el <i>disease mapping</i>	28
2.6.4.	Modelos aditivos	28
2.6.5.	Modelos con interacción	29
2.7.	Comparación de modelos	30
2.8.	<i>Integrated nested Laplace approximations</i> (INLA)	31
<b>3.</b>	<b>Ilustración con datos reales</b>	<b>32</b>
3.1.	Fuente de datos	32
3.2.	Análisis descriptivo	34
3.3.	Modelos espaciales	35
3.4.	Modelos espaciotemporales	42
3.5.	Análisis de la incidencia y mortalidad de cáncer de pulmón en Gran Bretaña (2002-2019)	45
<b>4.</b>	<b>Conclusiones</b>	<b>51</b>
<b>A.</b>	<b>Clasificación Internacional de Enfermedades CIE-10</b>	<b>55</b>
<b>B.</b>	<b>Tablas de comparación de modelos espaciotemporales</b>	<b>56</b>
<b>C.</b>	<b>Mapas por año de incidencia y mortalidad de cáncer de pulmón</b>	<b>58</b>

# Índice de Tablas

2.1. Tipos de interacciones espaciotemporales . . . . .	30
3.1. Estructura de la base de datos objeto de estudio. . . . .	34
3.2. Comparación de modelos espaciales. . . . .	39
3.3. Comparación de modelos espaciotemporales. . . . .	43
A.1. Descripción de la décima versión de la Clasificación Internacional de Enfermedades. . . . .	55
B.1. Comparación de modelos espaciotemporales para los datos de incidencia en hombres. . . . .	56
B.2. Comparación de modelos espaciotemporales para los datos de mortalidad en hombres. . . . .	57
B.3. Comparación de modelos espaciotemporales para los datos de incidencia en mujeres. . . . .	57
B.4. Comparación de modelos espaciotemporales para los datos de mortalidad en mujeres. . . . .	57

# Índice de Figuras

2.1. (a) Mapa original de John Snow mostrando los casos de cólera (indicados mediante rectángulos apilados) en la epidemia de Londres de 1854 (John Snow). (b) Mapa de la incidencia de COVID-19 acumulada a 14 días en las provincias españolas a 3 de marzo de 2022 (COVID). . . . .	4
2.2. Organización de las observaciones en $J$ niveles o grupos. . . . .	11
2.3. Estructura de un modelo agrupado ( <i>pooling model</i> ). . . . .	12
2.4. Estructura de un modelo no agrupado ( <i>no-pooling model</i> ). . . . .	13
2.5. Estructura de un modelo parcialmente agrupado ( <i>partial-pooling model</i> ) o de un modelo jerárquico ( <i>hierarchical model</i> ). . . . .	14
2.6. Matriz de estructura espacial de las provincias de Castilla-La Mancha con las siguientes enumeraciones: Toledo (1), Ciudad Real (2), Guadalajara (3), Cuenca (4) y Albacete (5). . . . .	16
2.7. Matriz de estructura temporal de primer orden de los años 2002-2006 con las siguientes enumeraciones: 2002 (1), 2003 (2), 2004(3), 2005 (4) y 2006 (5). . .	25
2.8. Matriz de estructura temporal de segundo orden de los años 2002-2006 con las siguientes enumeraciones: 2002 (1), 2003 (2), 2004 (3), 2005 (4) y 2006 (5). . .	26
3.1. Subdivisión final de los territorios: 106 <i>clinical commissioning groups</i> ingleses (rojo), 22 <i>local authorities</i> galeses (verde) y 14 <i>health boards</i> escoceses (azul). . .	33

3.2. Casos observados, población y tasa cruda por cien mil habitantes de cáncer de pulmón en mujeres (año 2019) . . . . .	35
3.3. Grafo de vecindad de las regiones bajo estudio. . . . .	37
3.4. Aproximación de las distribuciones posteriores y medianas (azul) del valor base ( $\eta$ ), del efecto espacial correspondiente a la región de Barnsley ( $u_1$ ) y del hiperparámetro ( $\tau_u$ ). . . . .	38
3.5. Aproximación de las distribuciones posteriores y medianas (azul) del hiperparámetro de Leroux ( $\lambda$ ) y del predictor lineal correspondiente a la región de Barnsley ( $\eta_1$ ). . . . .	40
3.6. Mapa de tasas ajustadas y probabilidades de exceso obtenidas con el modelo de Leroux. . . . .	41
3.7. Aproximación de las distribuciones posteriores y medianas (azul) del hiperparámetro de Leroux ( $\lambda$ ) y de los hiperparámetros espacial ( $\tau_u$ ), temporal ( $\tau_\phi$ ) y de interacción ( $\tau_\delta$ ). . . . .	43
3.8. Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con el modelo Leroux + PA1 + TipolII. . . . .	44
3.9. Patrón espacial (izquierda) y temporal (derecha) obtenidos con el modelo Leroux + PA1 + TipolII. . . . .	45
3.10. Patrones temporales obtenidos para cada conjunto de datos . . . . .	47
3.11. Patrones espaciales obtenidos para cada conjunto de datos . . . . .	48
C.1. Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de incidencia en hombres. . . . .	59
C.2. Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de incidencia en mujeres. . . . .	60

C.3. Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de mortalidad en hombres. . . . .	61
C.4. Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de mortalidad en mujeres. . . . .	62

# **Capítulo 1**

## **Introducción**

Este Trabajo Fin de Máster se enmarca en el ámbito de la Representación Cartográfica de Enfermedades, un ámbito particular de la estadística espacial conocido en inglés como *Disease Mapping*, cuyo objetivo es proporcionar estimaciones a nivel de área geográfica de diferentes indicadores de salud como la incidencia, la supervivencia o la mortalidad por distintas enfermedades. Este tipo de investigaciones resultan altamente interesantes, ya que, a menudo, los factores individuales no son suficientes para explicar las causas que provocan una enfermedad y, por tanto, los factores contextuales de la zona de residencia deben tenerse en cuenta. Para la correcta representación de una enfermedad en un mapa, los casos, la población en riesgo, su tamaño y su estructura por grupos de edad han de tenerse en cuenta, ya que regiones más pobladas y/o poblaciones más envejecidas son más susceptibles de tener un mayor número de enfermos. Con el fin de comparar los efectos de una enfermedad entre regiones o áreas es necesario emplear procedimientos que permitan la comparación entre poblaciones, evitando o minimizando los efectos de los factores que pudiesen distorsionar o confundir la comparación, como por ejemplo la edad. Por esta razón, la mayoría de estudios emplean riesgos o tasas estandarizadas por edad. A menudo, estas medidas son excesivamente variables entre áreas, y en

la práctica, deben ser suavizadas mediante el uso de modelos estadísticos que las estabilicen.

Los modelos espaciales y espaciotemporales expuestos en este trabajo se han desarrollado en el ámbito de la estadística Bayesiana, un enfoque que entiende los parámetros del modelo como variables aleatorias con una cierta distribución, la cual se ve actualizada al observar los datos. Esta nueva distribución condicionada a los datos es fruto del Teorema de Bayes, y en la mayoría de casos requiere el uso de métodos numéricos para su cálculo. Uno de los algoritmos más conocidos es el método *Markov chain Monte Carlo* (MCMC), el cual se basa en la simulación para aproximar distribuciones. No obstante, a medida que aumenta la complejidad del modelo (principalmente debido a la dimensionalidad de los datos o al elevado número de parámetros a estimar), el algoritmo MCMC se puede volver muy lento, y es por eso que han surgido otros métodos más eficientes, como es el caso de la técnica de inferencia Bayesiana aproximada INLA (Rue et al. (2009) [1]). El algoritmo *Integrated Nested Laplace Approximations* (INLA) está especialmente diseñado para los modelos Gaussianos latentes, una familia de modelos entre los cuales se encuentran los expuestos en este trabajo. A diferencia de los métodos MCMC, INLA es un algoritmo determinista y no basado en simulaciones, el cual proporciona resultados precisos en un tiempo menor.

El principal objetivo de este trabajo es exponer los conceptos básicos del *disease mapping*, así como describir los modelos espaciales y espaciotemporales más empleados en el área de la epidemiología, situándolos en el contexto de la estadística bayesiana. Además, estos modelos se ajustarán a datos reales de cáncer de pulmón en la isla de Gran Bretaña durante el periodo 2002-2019, con el propósito de estudiar la distribución geográfica de esta enfermedad y su evolución en el tiempo. Todo el trabajo se desarrollará en R, empleando el paquete R-INLA (<https://www.r-inla.org/>), especialmente diseñado para la inferencia Bayesiana de modelos latentes Gaussianos mediante el método INLA.

# **Capítulo 2**

## **Marco teórico**

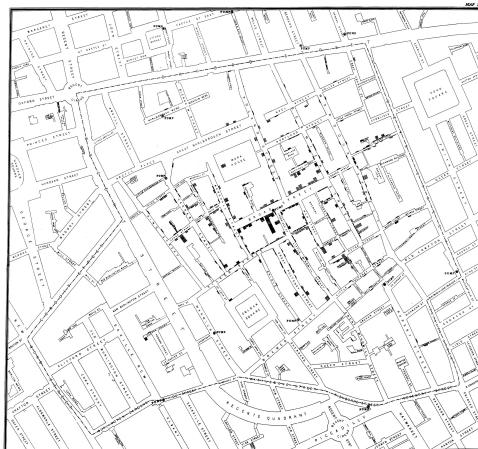
En este capítulo se desarrollarán las ideas introducidas anteriormente, poniendo en contexto y dando un sentido teórico a los modelos que más tarde se pondrán en práctica. Se supondrá un conocimiento básico sobre conceptos de Teoría de la Probabilidad y Estadística, como por ejemplo medidas descriptivas, variables aleatorias y sus distribuciones de probabilidad más comunes.

### **2.1. *Disease mapping***

A menudo, los factores individuales no son suficientes para explicar las causas que provocan una enfermedad, por lo que se deben tener en cuenta otros factores como la zona de residencia. Esta información puede ser de gran utilidad a la hora de explicar o modelizar los datos de la enfermedad, ya que puede existir una dependencia espacial entre zonas cercanas, además de otros posibles efectos o factores regionales que puedan estar influyendo en el comportamiento de los mismos. En el caso de los datos epidemiológicos, es común que exista cierta dependencia espacial, ya que, a menudo, la incidencia o mortalidad de las enfermedades pueden estar influenciadas por factores geográficos, y la información se recopila habitualmente de

forma agregada a nivel regional (unidades administrativas como municipios, zonas básicas de salud, provincias, etc.).

El doctor John Snow (1813-1858) es considerado fundador de la epidemiología moderna, ya que fue la primera persona de la que se tiene conciencia que empleó un mapa para estudiar la propagación de una enfermedad. En particular, Snow representó los casos de cólera del barrio londinense de Soho en el mapa que se puede ver en la Figura 2.1(a), descubriendo que el origen de la enfermedad se debía a una bomba de agua contaminada, ya que alrededor de ella se concentraban la mayoría de los casos [2]. Tras el éxito del doctor Snow, los mapas han seguido siendo empleados tanto para analizar como para reflejar el estado de una enfermedad en los diferentes países y regiones. El conjunto de técnicas que se han ido desarrollando entorno al uso de mapas en el ámbito de la salud pública se conoce como *disease mapping*. Un ejemplo reciente es el de la pandemia de COVID-19, donde los mapas de incidencia como el de la Figura 2.1(b) han sido una herramienta clave para los medios de comunicación.



(a) Mapa de John Snow



(b) Mapa de la incidencia de COVID-19

Figura 2.1: (a) Mapa original de John Snow mostrando los casos de cólera (indicados mediante rectángulos apilados) en la epidemia de Londres de 1854 ([John Snow](#)). (b) Mapa de la incidencia de COVID-19 acumulada a 14 días en las provincias españolas a 3 de marzo de 2022 ([COVID](#)).

## 2.2. Medidas clásicas de estimación de riesgo

En el ejemplo del doctor Snow, la representación en un mapa de la residencia de los afectados fue suficiente para determinar la causa de la enfermedad. Sin embargo, en la actualidad ya no se representan sólo los casos. La población en riesgo, su tamaño y su estructura por grupos de edad han de tenerse en cuenta, ya que poblaciones más envejecidas son más susceptibles de tener un mayor número de enfermos. Con el fin de comparar los efectos de una enfermedad entre regiones o áreas, es necesario emplear procedimientos que permitan la comparación entre poblaciones evitando o minimizando los efectos de los factores que pudiesen distorsionar o confundir la comparación, por ejemplo la edad. Por esta razón, la mayoría de estudios emplean riesgos o tasas estandarizadas por edad. Consideremos que el país o lugar geográfico bajo estudio está formado por  $S$  regiones o áreas, numeradas con  $i \in \{1, \dots, S\}$ . Además, supongamos que se tiene información de las  $S$  regiones durante  $T$  instantes de tiempo, numerados con  $t \in \{1, \dots, T\}$ . Para cada región  $i$  e instante  $t$ , denotemos por  $O_{it}$  el número de casos observados y por  $N_{it}$  el número de individuos en riesgo. Un nuevo indicador que resuelve el problema anterior es la denominada *tasa cruda* (TC), que se define como:

$$TC_{it} = \frac{O_{it}}{N_{it}} \cdot 100,000 \quad (2.1)$$

Este indicador representa la proporción de individuos afectados por una enfermedad en cada región e instante, y se suele medir en número de casos por cada cien mil habitantes. Si se observa de nuevo la Figura 2.1(b), se puede ver que el mapa representa la tasa cruda en cada una de las regiones, pudiendo en este caso comparar unas con otras ya que ahora el indicador tiene en cuenta cada una de las poblaciones en riesgo. En este caso, la tasa cruda se ha calculado empleando el número de nuevos casos observados durante las últimas dos semanas, lo que se

conoce como incidencia acumulada.

Pese a haber resuelto el problema de las diferencias entre las poblaciones en riesgo, la tasa cruda sigue siendo un indicador muy pobre. Por ejemplo, no tiene en cuenta que la población de una región pueda estar más envejecida que otra, lo cual puede ser el causante de un mayor número de enfermos. Con el propósito de resolver este problema, aparecen los llamados métodos de estandarización.

### 2.2.1. Método de estandarización directo

Este primer método se basa en una cierta población de referencia, como por ejemplo la población estándar europea, la cual recoge la proporción de individuos en cada grupo de edad teniendo en cuenta los diferentes países europeos. Supongamos que la población bajo estudio se divide en  $J$  grupos de edad, numerados con  $j \in \{1, \dots, J\}$ . De esta manera, para cada región  $i$  y periodo  $t$ ,  $O_{itj}$  y  $N_{itj}$  indican el número de casos observados y el número de individuos en riesgo en el grupo de edad  $j$ , respectivamente. Para calcular la tasa estandarizada ( $TE_{it}$ ), solamente debemos conocer la población en riesgo ( $P_j$ ) correspondiente al  $j$ -ésimo grupo de edad.

$$TE_{it} = \frac{\sum_{j=1}^J P_j \cdot \frac{O_{itj}}{N_{itj}}}{\sum_{j=1}^J P_j} \cdot 100,000 \quad (2.2)$$

Este nuevo indicador permite comparar regiones con grupos de edad distintos, ya que la comparación se hace con respecto a la misma población de referencia.

### 2.2.2. Método de estandarización indirecto

El segundo método de estandarización emplea exclusivamente información de las regiones bajo estudio, combinando la de cada una de ellas para estimar el número de casos esperados

en cada región. Manteniendo la notación anterior, el número de casos esperados se define como:

$$e_{it} = \sum_{j=1}^J N_{itj} \cdot \frac{O_j}{N_j} \quad (2.3)$$

Donde  $O_j = \sum_{i=1}^S \sum_{t=1}^T O_{itj}$  y  $N_j = \sum_{i=1}^S \sum_{t=1}^T N_{itj}$ . Finalmente, el *ratio de incidencia estandarizado* (RIE) se calcula como:

$$RIE_{it} = \frac{O_{it}}{e_{it}} \quad (2.4)$$

A diferencia de los anteriores indicadores, el RIE aporta una información diferente. Si el ratio es mayor que 1, esto quiere decir que se han observado más casos de los esperados, lo que indica un mayor riesgo de la enfermedad en esa región y periodo de tiempo en comparación con el total del territorio bajo estudio durante el periodo de análisis completo. Por el contrario, si el ratio es menor que 1, esto indicará que el riesgo de la enfermedad en la región es menor respecto al marco de referencia. No obstante, que el RIE sea bajo no tiene por qué significar un buen estado de la enfermedad, ya que este indicador es exclusivamente comparativo.

Los métodos de estandarización pueden servir para solucionar tanto el problema de la diferencia entre las poblaciones en riesgo como el problema de la distribución de edades. Sin embargo, cuando la población de las regiones es muy pequeña o cuando la enfermedad es rara y contabiliza pocos casos, estos indicadores se pueden volver extremadamente variables. Por esta razón, es necesario emplear modelos estadísticos que incorporen información que permitan suavizar las tasas o riesgos introduciendo dependencia (correlación) espacial, temporal y espaciotemporal en los datos. Estos modelos comúnmente se formulan dentro de un marco Bayesiano jerárquico con dos enfoques principales: un enfoque empírico Bayesiano (enfoque clásico o frecuentista) y uno totalmente Bayesiano (*fully Bayesian approach* en inglés).

## 2.3. Estadística Bayesiana

El concepto de probabilidad ha tenido varias interpretaciones a lo largo de la historia, todas ellas con el objetivo de cuantificar la ocurrencia de un suceso durante un experimento aleatorio, usualmente asignándole un número entre 0 y 1. Por un lado, el enfoque objetivista trata de asignar esta cantidad desde un punto de vista objetivo, dejando de lado toda creencia u opinión personal y basándose solamente en las propiedades matemáticas tanto del suceso como del experimento. El principal enfoque de este tipo es el frecuentista, el cual define la probabilidad de un suceso como la frecuencia relativa de éxitos a largo plazo. Cuando el experimento se puede repetir tantas veces como sea necesario, por ejemplo, lanzar una moneda, el enfoque frecuentista parece adecuado. Sin embargo, existen experimentos que no poseen esta propiedad, como por ejemplo si alguien se pregunta cuál es la probabilidad de que un determinado partido gane las elecciones. Por otro lado, el enfoque subjetivista o Bayesiano emplea la probabilidad para reflejar el grado de incertidumbre que tiene acerca de la ocurrencia de un suceso. De este modo, para un mismo experimento, cada persona puede asignar su propia probabilidad a cada suceso, reflejando su creencia particular. Ahora es posible responder a la pregunta de las elecciones, pudiendo asignar una probabilidad en base a opiniones expertas o estudios sobre otras elecciones.

Cada enfoque ha desarrollado sus propias técnicas y métodos, y dependiendo del estudio que se desee llevar a cabo puede ser más conveniente el empleo de uno u otro. En este caso, tanto los modelos como los métodos numéricos que se utilizarán se han desarrollado dentro del ámbito de la estadística Bayesiana.

### 2.3.1. Inferencia Bayesiana

Los modelos estadísticos suelen estar definidos en base a unos parámetros, el valor de los cuales se ajusta en función de los datos que se observan. A menudo se tienen ciertas hipótesis o creencias acerca de estos parámetros, las cuales se quieren contrastar mediante la observación de datos. El enfoque Bayesiano entiende estos parámetros no como cantidades fijas sino como variables aleatorias con una cierta distribución de probabilidad, y la inferencia consistirá en estudiar cómo cambia esta distribución al observar los datos. Sea  $Y$  una variable aleatoria con función de probabilidad  $p$  (o función de densidad de probabilidad si  $Y$  es continua), si esta se modela mediante un cierto parámetro (o parámetros)  $\theta$ , entonces la función de verosimilitud se define como:

$$L(\theta | y) := p(Y = y | \theta) \quad (2.5)$$

donde  $y \in \mathbb{R}$  son los datos observados. Es decir,  $L(\theta | y)$  representa la probabilidad de que  $Y$  tome el valor  $y$  en función del valor de  $\theta$ . Por ejemplo,  $Y$  puede reflejar el número de muertes observadas por una cierta enfermedad y el interés reside en estudiar la tasa de mortalidad  $\theta$ . Por simplicidad, a partir de ahora se denotará la verosimilitud con  $p(y | \theta)$ . Puesto que se ha adoptado el enfoque Bayesiano, el parámetro  $\theta$  es una cantidad desconocida que se entiende como una variable aleatoria con una cierta distribución de probabilidad. La primera distribución que se le asigna es independiente de los datos observados y se conoce como distribución *a priori*, denotada con  $p(\theta)$ . Esta función debe reflejar la información que se tiene en un principio sobre el parámetro  $\theta$ , pudiendo ser más o menos concreta. Bajo una estructura jerárquica, esta distribución puede depender a su vez de otros parámetros, llamados hiperparámetros, los cuales se estudiarán más adelante.

El siguiente paso es estudiar cómo varía la distribución del parámetro  $\theta$  tras haber obser-

vado los datos. Dadas las dos componentes (verosimilitud y distribución a priori), la inferencia se realiza a través del Teorema de Bayes:

$$p(\theta | y) = \frac{p(y | \theta) \cdot p(\theta)}{p(y)} \quad (2.6)$$

obteniendo lo que se conoce como distribución *a posteriori*, denotada con  $p(\theta | y)$ . En el denominador de la Ecuación 2.6 aparece la distribución marginal de  $y$ , la cual se suele considerar una constante de normalización al no depender de  $\theta$ . Así pues, la Ecuación 2.6 muchas veces se presenta como:

$$p(\theta | y) \propto p(y | \theta) \cdot p(\theta) \quad (2.7)$$

donde el signo de igualdad ( $=$ ) es sustituido por el signo de proporcionalidad ( $\propto$ ). Si  $D_\theta$  denota el soporte de  $\theta$ , entonces la distribución marginal de  $y$  puede ser calculada mediante:

$$p(y) = \sum_{\theta \in D_\theta} p(y | \theta) \cdot p(\theta) \quad \text{o} \quad p(y) = \int_{\theta \in D_\theta} p(y | \theta) \cdot p(\theta) \cdot d\theta \quad (2.8)$$

si bien  $\theta$  es discreta o continua, respectivamente.

La ventaja del enfoque bayesiano es que ofrece toda la distribución *a posteriori* del parámetro  $\theta$ , pudiendo calcular la mediana o cualquier otro indicador que se considere oportuno. Del mismo modo, se pueden construir los llamados intervalos de credibilidad (IC), que pese a su similitud con los intervalos de confianza, aportan una información diferente. Bajo el enfoque frecuentista, un intervalo de confianza del  $(100 - \alpha)\%$  sugiere que si se repite el mismo experimento un número elevado de veces, entonces alrededor del  $\alpha\%$  de estas, el valor de  $\theta$  quedará fuera del intervalo. Por el contrario, un intervalo de credibilidad del  $(100 - \alpha)\%$  simplemente indica que  $p(\theta \in IC | y) = 1 - \frac{\alpha}{100}$ , desligándolo completamente del concepto de repetibilidad.

## 2.4. Modelos jerárquicos

A menudo, se dispone de información estructurada en diferentes grupos o niveles, como pueden ser rangos de edad, períodos de tiempo, hospitales o regiones geográficas. Cuando se pretende modelizar este tipo de datos, puede ser interesante caracterizar cada grupo mediante un parámetro propio, reflejando sus propias características y dando flexibilidad al modelo. Sin embargo, cuando alguno de estos grupos contiene muy pocos datos, la incertidumbre acerca de su parámetro puede llegar a ser muy alta, además de que se deja de lado información procedente de otras observaciones. Los modelos jerárquicos tratan de resolver este problema permitiendo el intercambio de información entre parámetros aunque estos sean diferentes, un concepto que se conoce como intercambiabilidad o *exchangeability*.

Supongamos que se tienen  $n$  observaciones agrupadas en  $J$  niveles o grupos, cada uno de los cuales numerados con  $j \in \{1, \dots, J\}$  y formados por  $n_j$  elementos, de manera que  $n = n_1 + \dots + n_J$ . Así, para cada grupo  $j$ , su observación  $i$ -ésima se representará mediante  $y_{ij}$ , con  $i \in \{1, \dots, n_j\}$ . En la Figura 2.2 se puede visualizar la organización de los datos en los diferentes grupos.

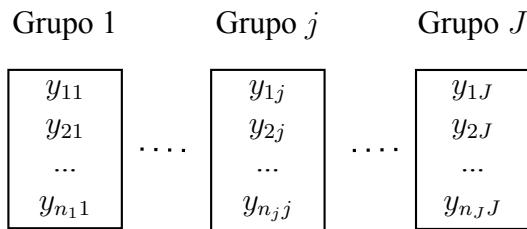


Figura 2.2: Organización de las observaciones en  $J$  niveles o grupos.

A la hora de modelizar los datos, en un primer extremo aparece el modelo que considera el mismo parámetro  $\theta$  para todos los grupos, conocido como modelo agrupado o *pooling model*. Esto equivale a pasar por alto el hecho de que cada grupo o nivel puede tener sus propias características, haciendo que sus observaciones presenten un comportamiento diferente al de

las del resto de grupos. Por ejemplo, si se desea estudiar una cierta enfermedad y se dispone de información de pacientes de diversos hospitales, entonces este modelo no refleja las diferencias entre los centros, como podrían ser unas peores instalaciones o falta de material o de profesionales. En la Figura 2.3 se puede observar la estructura de un modelo agrupado, en el que se emplea un mismo parámetro  $\theta$  para modelizar todos los grupos.

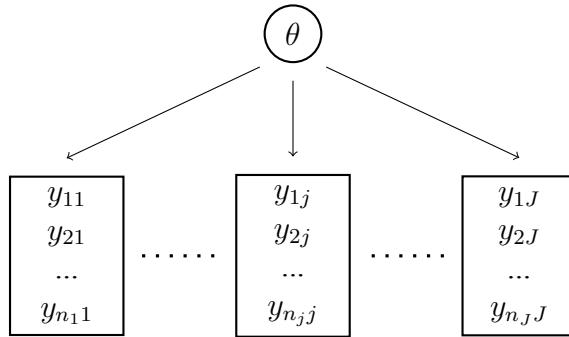


Figura 2.3: Estructura de un modelo agrupado (*pooling model*).

En el otro extremo, aparece el modelo que asigna un parámetro diferente a cada grupo, conocido como modelo no agrupado o *no-pooling model*. Es decir, ahora cada grupo  $j$  estará caracterizado por su propio parámetro  $\theta_j$  *a priori* independiente del resto. En este caso, se da más flexibilidad al modelo, pero la inferencia en cada parámetro se realizará teniendo en cuenta exclusivamente los datos de su grupo, y esto puede ser un problema cuando alguno de ellos carezca de suficiente información. En estos casos, la variabilidad de las distribuciones a posteriori de los parámetros será muy alta, lo que equivale a una gran incertidumbre. Por ejemplo, esto puede ocurrir cuando uno de los hospitales posea pocos pacientes con la enfermedad bajo estudio por ser pequeño o pertenecer a una zona poco poblada. En la Figura 2.4 se tiene la estructura de un modelo no agrupado, en el que cada grupo se modeliza mediante un parámetro propio independiente del resto.

Los planteamientos anteriores han presentado varios problemas, ya sea por falta de flexibilidad o por una independencia total entre los parámetros que hace perder fuerza al modelo

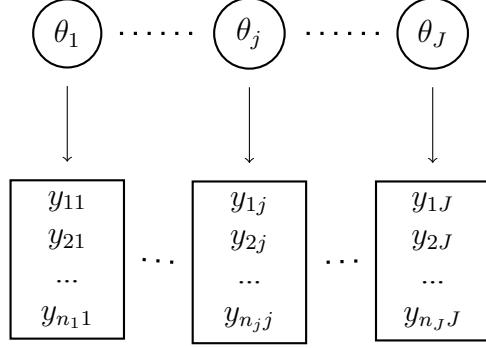


Figura 2.4: Estructura de un modelo no agrupado (*no-pooling model*).

cuando la información es escasa. Una posible solución, la cual es la base de los modelos jerárquicos, es introducir un nuevo parámetro (o parámetros)  $\tau$ , normalmente conocido como hiperparámetro. La función de  $\tau$  ya no será describir los datos, sino que se encargará de controlar los propios parámetros  $\theta_j$ , desvelando la estructura jerárquica que da nombre a esta clase de modelos. La idea es considerar los parámetros  $\theta_j$  como elementos de un vector aleatorio  $\Theta := [\theta_1, \dots, \theta_J]^T$ , cuya precisión (inversa de la varianza) dependa del hiperparámetro  $\tau$ . Este podrá ser fijado (modelo parcialmente agrupado o *partial-pooling model*) o bien considerado una variable aleatoria con una cierta distribución *a priori* (modelo jerárquico o *hierarchical model*), lo que suele ser habitual. De esta manera, cada  $\theta_j$  deberá ajustarse a los datos de su grupo, pero también tendrá que compartir información con el resto de parámetros para ajustarse a su distribución conjunta. En la Figura 2.5 podemos observar la estructura de este tipo de modelos.

Habitualmente, se asume que el vector aleatorio  $\Theta$  sigue una distribución normal multivariante de media cero, y dependiendo de cómo se construya la matriz de precisión tendremos un tipo de modelos u otro. A estos parámetros también se los conoce como efectos aleatorios o *random effects* y el caso más básico es considerarlos variables independientes e idénticamente distribuidas (i.i.d.), es decir,

$$\Theta = [\theta_1, \dots, \theta_J]^T \cong N_J(\mu, (\tau I_J)^{-1}) \quad (2.9)$$

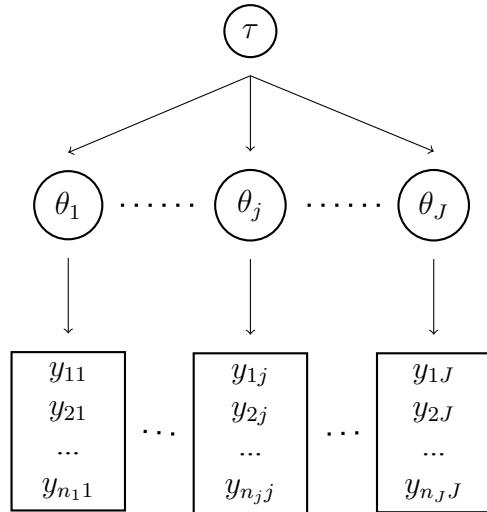


Figura 2.5: Estructura de un modelo parcialmente agrupado (*partial-pooling model*) o de un modelo jerárquico (*hierarchical model*).

donde  $I_J$  denota la matriz identidad de orden  $J$ . Sin embargo, la matriz de precisión se puede construir de múltiples maneras, pudiendo introducir estructuras espaciales o temporales, cuya combinación da lugar a los modelos que trataremos en este trabajo.

## 2.5. Modelos espaciales

En la sección anterior se han introducido los modelos jerárquicos, así como el tipo de efectos aleatorios más básico, los i.i.d. No obstante, estos no tienen en cuenta la estructura espacial que puede existir entre grupos, así que en esta sección se verá cómo se puede introducir esta información y se hará un repaso de los modelos espaciales más comunes.

### 2.5.1. Datos de área

Cuando se habla de datos espaciales, se está haciendo referencia a información que está relacionada de alguna manera con una posición o región geográfica. Sin embargo, dependiendo de cómo sea esta relación, pueden existir diferentes tipos de datos, y conviene aclarar cuáles son los que se van a tratar en este trabajo. Según [3], existen tres tipos de datos espaciales:

1. Datos de área (*lattice data*): cuando el dominio  $D$  de la variable de interés  $Y$  es fijo. Este puede ser regular (cuadrículas) o irregular (distritos, provincias, etc.), teniendo un número finito de divisiones con fronteras bien definidas. Un ejemplo son los datos de incidencia de COVID-10 representado en la Figura 2.1(b).
2. Datos geoestadísticos (*point-reference data*): cuando la variable de interés  $Y$  varía de forma continua en su dominio fijo  $D$ , observándose en algunos puntos  $s \in D$ .
3. Datos de procesos puntuales (*point-pattern data*): cuando el dominio  $D$  en sí mismo es aleatorio, es decir, la localización donde se ha observado el dato es aleatoria. Un ejemplo son los datos representados por el doctor Snow en su mapa de la Figura 2.1(a).

Así pues, según el tipo de información de la que se disponga, se deberá tratar de una manera u otra. En particular, en este trabajo el interés reside en los datos de área, especialmente cuando las fronteras son irregulares. La existencia de estas delimitaciones permite crear los denominados grafos de vecindad, en los que los nodos se corresponden con las regiones bajo estudio y las aristas conectan dos nodos si estos son vecinos, entendiendo por vecinos que comparten frontera. A su vez, los grafos pueden ser representados mediante las matrices de adyacencia, cuyas entradas  $(i, j)$  son 1 si las regiones  $i$  y  $j$  son vecinas y 0 en caso contrario.

Volviendo a la notación de las Sección 2.2 en la que se tenían  $S$  regiones bajo estudio, diremos que dos regiones  $i$  y  $j$  son vecinas, denotándolo con  $i \sim j$ , si sus fronteras coinciden en al menos un punto. Además, denotaremos con  $\mathcal{V}(i)$  el número de vecinos de cada región  $i$ , es decir,  $\mathcal{V}(i) := \#\{j \mid i \sim j\}$ . Gracias a esto, es posible crear lo que se conoce como matriz de estructura espacial  $R$ , que se define como:

$$R_{ij} = \begin{cases} \mathcal{V}(i) & \text{si } i = j \\ -1 & \text{si } i \sim j \\ 0 & \text{si otro caso} \end{cases} \quad (2.10)$$

En la Figura 2.6 se puede observar la matriz de estructura espacial de Castilla-La Mancha. Como se puede ver, esta matriz no es más que una matriz de adyacencia negativa en la que aparece el número de vecinos de cada nodo (provincia) en la diagonal.

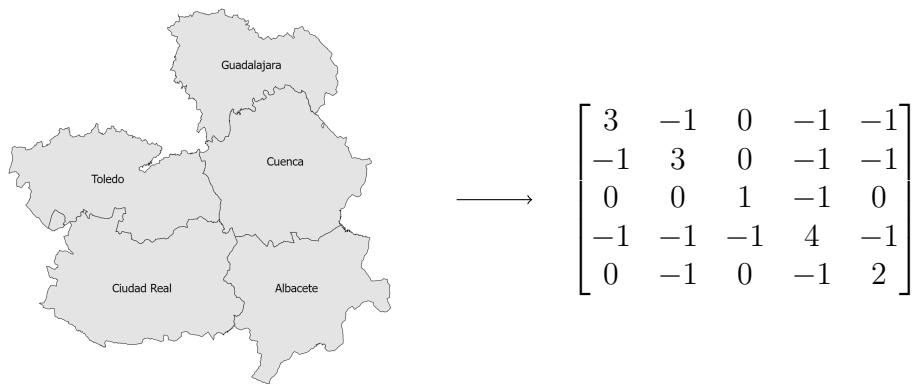


Figura 2.6: Matriz de estructura espacial de las provincias de Castilla-La Mancha con las siguientes enumeraciones: Toledo (1), Ciudad Real (2), Guadalajara (3), Cuenca (4) y Albacete (5).

### 2.5.2. Matriz de precisión e independencia condicionada

La idea de los modelos espaciales es emplear la matriz de estructura espacial del grafo de vecindad para construir la matriz de precisión del vector de efectos aleatorios. Esto se debe a que la matriz de precisión da información sobre la dependencia condicionada entre las componentes del vector aleatorio, de manera que si la entrada  $(i, j)$  de la matriz es cero, significa que las componentes  $i$  y  $j$  son independientes dado el resto de componentes. Al construir la matriz de precisión a partir de la matriz de estructura espacial, estaremos diciendo que los efectos aleatorios de dos regiones no vecinas serán independientes dado el resto de regiones, ya que

la entrada será cero según la Ecuación 2.10. A continuación, se explicará más en detalle lo comentado.

Sea  $X$  un vector aleatorio  $n$ -dimensional que sigue distribución normal multivariante con vector de medias  $\mu$  y matriz de varianzas y covarianzas  $\Sigma$ :

$$X \sim N_n(\mu, \Sigma) \quad (2.11)$$

es conocido que su función de densidad de probabilidad viene dada por:

$$p(x) := p(X = x) = ((2\pi)^n |\Sigma|)^{-1/2} \exp(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)) \quad (2.12)$$

Si ahora se define la matriz de precisión  $\Phi$  como la inversa de  $\Sigma$ , es decir,

$$\Phi := \Sigma^{-1} \quad (2.13)$$

entonces la Ecuación 2.12 se puede reescribir en términos de  $\Phi$ :

$$p(x) = ((2\pi)^{-n} |\Phi|)^{1/2} \exp(-\frac{1}{2}(x - \mu)^T \Phi (x - \mu)) \quad (2.14)$$

El vector  $X$  puede ser dividido en dos subvectores  $X_1$  y  $X_2$  de dimensiones  $n_1$  y  $n_2$ , respectivamente ( $n_1 + n_2 = n$ ). El vector de medias y la matriz de varianzas y covarianzas por bloques quedarían:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_1 & C \\ C^T & \Sigma_2 \end{bmatrix} \quad (2.15)$$

Por supuesto, los subvectores heredan la distribución normal multivariante, de manera que  $X_i \sim N_{n_i}(\mu_i, \Sigma_i)$ . Ahora bien, gracias al álgebra lineal, se puede demostrar que la matriz

de precisión por bloques sería de la siguiente manera:

$$\Phi = \begin{bmatrix} \Phi_1 & P \\ P^T & \Phi_2 \end{bmatrix} \quad (2.16)$$

donde

- $\Phi_1 = (\Sigma_1 - C\Sigma_2^{-1}C^T)^{-1}$

- $P = -\Phi_1 C \Sigma_2^{-1}$

- $\Phi_2 = (\Sigma_2 - C^T \Sigma_1^{-1} C)^{-1}$

Una vez se obtiene esto, el término de dentro de la exponencial de la Ecuación 2.14 se puede escribir como  $-Q(x)/2$ , donde:

$$Q(x) := (x - \mu)^T \Phi (x - \mu) = (x_1 - \mu_1)^T \Phi_1 (x_1 - \mu_1) + (2.17) \\ + 2(x_1 - \mu_1)^T P (x_2 - \mu_2) + (x_2 - \mu_2)^T \Phi_2 (x_2 - \mu_2)$$

y por tanto, se tendrá que  $p(x) \propto \exp(-Q(x)/2)$ . Ahora bien, condicionando  $X_1$  a  $X_2$ , puesto que el término  $(x_2 - \mu_2)^T \Phi_2 (x_2 - \mu_2)$  sería ahora una constante, entonces  $p(x_1 | x_2) \propto \exp(-Q(x_1 | x_2)/2)$ , donde

$$Q(x_1 | x_2) := (x_1 - \mu_1)^T \Phi_1 (x_1 - \mu_1) + 2(x_1 - \mu_1)^T P (x_2 - \mu_2) \quad (2.18)$$

De esta manera, la distribución condicionada de  $X_1$  a  $X_2$  vendrá dada por

$$X_1 | X_2 \sim N_{n_1}(\mu_1 - \Phi_1^{-1} P (x_2 - \mu_2), \Sigma_1) \quad (2.19)$$

ya que

$$\begin{aligned} p(x_1 | x_2) &\propto \exp(-Q(x_1 | x_2)/2) \propto \\ &\propto \exp((x_1 - (\mu_1 - \Phi_1^{-1}P(x_2 - \mu_2)))^T \Phi_1 (x_1 - (\mu_1 - \Phi_1^{-1}P(x_2 - \mu_2)))) \end{aligned} \quad (2.20)$$

Por último, consideremos que  $X_1$  está formado sólo por la primera componente y que  $X_2$  contiene al resto, lo que se puede indicar con  $X_{-1}$ . Si se denotan con  $\phi_{ij}$  los elementos de la matriz de precisión  $\Phi$ , según el razonamiento anterior se tiene que

$$X_1 | X_{-1} \sim N\left(\mu_1 - \frac{1}{\phi_{11}} P(x_{-1} - \mu_{-1}), \frac{1}{\phi_{11}}\right) \quad (2.21)$$

pero ahora, dado que  $P$  es el vector  $[\phi_{12}, \phi_{13}, \dots, \phi_{1n}]^T$ , entonces

$$X_1 | X_{-1} \sim N\left(\mu_1 - \sum_{j=2}^n \frac{\phi_{1j}}{\phi_{11}} (x_j - \mu_j), \frac{1}{\phi_{11}}\right) \quad (2.22)$$

Por supuesto, esto se puede extender para cualquier componente  $i$ -ésima del vector  $X$ , de manera que la ecuación general para la distribución de la componente  $i$ -ésima condicionada al resto de componentes sería:

$$X_i | X_{-i} \sim N\left(\mu_i - \sum_{j \neq i} \frac{\phi_{ij}}{\phi_{ii}} (x_j - \mu_j), \frac{1}{\phi_{ii}}\right) \quad (2.23)$$

Llegados a este punto, se puede ver bien que si  $\phi_{ij} = 0$ , entonces las componentes  $i$  y  $j$  son independientes condicionadas al resto, ya que el valor de  $x_j$  no afectará para nada a la distribución de  $X_i | X_{-i}$  [4]. Esta es la clave de los modelos espaciales, combinar este resultado con la matriz de estructura espacial, consiguiendo que la dependencia entre los efectos aleatorios de las regiones dependa del grafo de vecindad, y por tanto de la geografía de las mismas.

### 2.5.3. Modelos espaciales para el *disease mapping*

Volviendo al ámbito de la epidemiología espacial, si se desea realizar un estudio sobre una cierta enfermedad, es posible hacerlo a través de dos indicadores: el riesgo relativo (casos observados entre casos esperados) o la tasa de mortalidad o incidencia (casos observados entre población en riesgo). Puesto que el uso de uno u otro es análogo, denotaremos ambos con  $r_i$  para cada región  $i$ , así como con  $e_i$  el número de casos esperados y con  $N_i$  la población en riesgo. Bajo estas condiciones, es habitual considerar que, condicionado al riesgo/tasa  $r_i$ , el número de casos observados  $O_i$  sigue una distribución de Poisson de parámetro  $\lambda_i$ . Puesto que  $\lambda_i$  coincide con la media de la variable aleatoria, según el indicador empleado se tendrá:

$$O_i \mid r_i \sim \begin{cases} Poi(\lambda_i = e_i r_i) & \text{con el riesgo relativo} \\ Poi(\lambda_i = N_i r_i) & \text{con la tasa} \end{cases} \quad (2.24)$$

Así pues, el objetivo será modelizar  $r_i$  mediante efectos aleatorios que tengan una estructura jerárquica basada en la matriz del grafo de vecindad. En particular,  $r_i$  se enlazará con un predictor lineal  $\eta_i$  mediante la función logaritmo, obteniendo el siguiente modelo logarítmico lineal de Poisson:

$$\log(r_i) = \eta_i \implies \begin{cases} \log(\lambda_i) = \eta_i + \log(e_i) & \text{con el riesgo relativo} \\ \log(\lambda_i) = \eta_i + \log(N_i) & \text{con la tasa} \end{cases} \quad (2.25)$$

Como se había dicho, trabajar con un indicador u otro es análogo, ya que tanto  $\log(e_i)$  como  $\log(N_i)$  se consideran variables de exposición u *offset*, y las diferencias entre unos modelos u otros dependerán de cómo se especifique el predictor lineal  $\eta_i$ . Por lo general, los modelos espaciales se basan en distribuciones condicionadas autoregresivas (CAR), donde, en el modelo

más simple,  $\eta_i$  se construye de la siguiente manera:

$$\eta_i = \eta + u_i \quad (2.26)$$

donde  $\eta$  representa una valor base y el vector  $u = [u_1, \dots, u_S]^T$  está estructurado espacialmente.

A continuación, se expondrán los modelos CAR más empleados.

#### 2.5.4. Modelo CAR intrínseco

Una de las distribuciones CAR más simples es la llamada distribución CAR intrínseca o iCAR [5]. Su uso es muy frecuente en el mapeo de enfermedades y tiene la siguiente forma:

$$\eta_i = \eta + u_i \quad (2.27)$$

Se asume que el vector aleatorio  $u$  sigue una distribución normal multivariante de media 0 y matriz de precisión  $\tau_u R$ , donde  $\tau_u$  es un hiperparámetro con una cierta distribución *a priori* y  $R$  es la matriz de estructura espacial definida en la Ecuación 2.10. La media de la distribución se asume nula, ya que estos efectos aleatorios simplemente se encargarán de oscilar alrededor del valor base  $\eta$ , mientras que la matriz de precisión se construye a partir de  $R$  para introducir la dependencia espacial. La distribución marginal de cada componente  $u_i$  condicionada al resto de parámetros se puede obtener mediante la Ecuación 2.23:

$$u_i | u_{-i} \sim N \left( 0 - \sum_{j \neq i} \frac{\tau_u R_{ij}}{\tau_u R_{ii}} (u_j - 0), \frac{1}{\tau_u R_{ii}} \right) \quad (2.28)$$

pero ahora, teniendo en cuenta la construcción de  $R$ ,

$$u_i | u_{-i} \sim N \left( \sum_{j \sim i} \frac{u_j}{\mathcal{V}(i)}, \frac{1}{\tau_u \mathcal{V}(i)} \right) \quad (2.29)$$

Como se puede ver, las componentes con mayor número de vecinos tendrán una varianza menor, ya que absorberán más información de su entorno. Del mismo modo, la media de cada componente condicionada al resto será la media de sus vecinos.

### 2.5.5. Modelo BYM

En la práctica, el modelo iCAR puede resultar muy restrictivo, ya que asume una dependencia espacial completa entre todas las regiones vecinas. Esto se puede mejorar con el modelo BYM [5], el cual añade un segundo componente espacial a la Ecuación 2.27:

$$\eta_i = \eta + u_i + v_i \quad (2.30)$$

donde  $v_i$  son las componentes de un vector aleatorio  $v = [v_1, \dots, v_S]$  con una distribución normal multivariante de media 0 y matriz de precisión  $\tau_v I_S$ . Es decir, los efectos aleatorios son i.i.d., el caso más simple que se vio dentro de los modelos jerárquicos. A partir de ahora, se hará referencia a los efectos aleatorios  $u_i$  y  $v_i$  como efectos espaciales estructurados y no estructurados, respectivamente.

### 2.5.6. Modelo de Leroux

El modelo de Leroux [6] es similar al modelo BYM, solo que modeliza los efectos estructurados y no estructurados de un modo más compacto:

$$\eta_i = \eta + \xi_i \quad (2.31)$$

En este caso, los efectos  $\xi_i$  son las componentes de un vector aleatorio  $\xi = [\xi_1, \dots, \xi_S]$  de media cero y cuya matriz de precisión es una combinación lineal entre  $\tau_\xi R$  y  $\tau_\xi I_S$ , siendo  $\tau_\xi$  un hiperparámetro que controla la variabilidad. Esta matriz tiene pues la siguiente forma:

$$\tau_\xi(\lambda R + (1 - \lambda)I_S) \quad (2.32)$$

En este caso,  $\lambda$  es otro hiperparámetro que varía entre 0 y 1. Cabe destacar que si  $\lambda = 0$ , entonces la matriz de precisión sería  $\tau_\xi I_S$  (parámetros i.i.d.), mientras que si  $\lambda = 1$  la matriz sería  $\tau_\xi R$  y estaríamos en el caso del modelo iCAR. La distribución marginal de cada componente  $\xi_i$  condicionada al resto de parámetros se puede obtener mediante la Ecuación 2.23:

$$\xi_i | \xi_{-i} \sim N \left( 0 - \sum_{j \neq i} \frac{\tau_\xi \lambda R_{ij}}{\tau_\xi(\lambda R_{ii} + 1 - \lambda)} (\xi_j - 0), \frac{1}{\tau_\xi(\lambda R_{ii} + 1 - \lambda)} \right) \quad (2.33)$$

pero ahora, teniendo en cuenta la construcción de  $R$ ,

$$\xi_i | \xi_{-i} \sim N \left( \sum_{j \sim i} \frac{\lambda \xi_j}{\lambda \mathcal{V}(i) + 1 - \lambda}, \frac{1}{\tau_\xi(\lambda \mathcal{V}(i) + 1 - \lambda)} \right) \quad (2.34)$$

De nuevo, se puede ver que coincide con la Ecuación 2.27 (modelo iCAR) cuando  $\lambda = 1$ , compartiendo algunas de sus propiedades como que la varianza sea menor cuanto mayor sea el número de vecinos. El modelo de Leroux permite desplazar la importancia de los efectos estructurados a no estructurados en función de los datos, lo que lo convierte en un modelo muy flexible.

## 2.6. Modelos espaciotemporales

Es posible que, además de contar con información espacial, esta información esté disponible en diferentes períodos de tiempo, como pueden ser años, meses, etc. Las dependencias temporales pueden ser determinantes a la hora de realizar un estudio, ya que habitualmente el comportamiento de los datos es similar entre instantes de tiempo cercanos o con las mismas características. En esta sección se verá cómo introducir estas estructuras temporales en los modelos, además de cómo combinarlas con las espaciales.

### 2.6.1. Modelos que incorporan estructura temporal

Una serie temporal se puede definir como una sucesión de datos medidos en determinados instantes de tiempo y ordenados cronológicamente. En este caso, serán de interés aquellos datos que estén medidos en intervalos de tiempo de igual tamaño, como pueden ser meses, años, etc. Del mismo modo que en los datos de área se podía definir para cada región el conjunto de sus vecinos, en las series temporales podemos definirlo para cada instante de tiempo. En este caso, diremos que dos instantes  $t_i$  y  $t_j$  son vecinos si son próximos en el tiempo. La cercanía se puede medir en el número de intervalos de tiempo que los separan, apareciendo diferentes conjuntos de vecinos según como se defina esta proximidad. Diremos que los instantes  $t_i$  y  $t_j$  son vecinos de  $m$ -ésimo orden si  $0 < |t_i - t_j| \leq m$ , denotando por  $\mathcal{V}_m(t_i)$  el número de vecinos  $m$ -ésimos del instante  $t_i$ . Así pues, del mismo modo que se construye la matriz de estructura espacial  $R$  (ver Ecuación 2.10), se puede construir la matriz de estructura temporal de  $m$ -ésimo orden  $T_m$ :

$$(T_m)_{ij} = \begin{cases} \mathcal{V}_m(t_i) & \text{si } i = j \\ -1 & \text{si } 0 < |t_i - t_j| \leq m \\ 0 & \text{si otro caso} \end{cases} \quad (2.35)$$

En la práctica, es habitual trabajar con  $m = 1, 2$  para no complicar demasiado los modelos. En las Figuras 2.7 y 2.8 se pueden observar las matrices de estructura temporales de estos dos órdenes para los años 2002-2006, así como el grafo que conecta cada año con sus vecinos en cada caso.

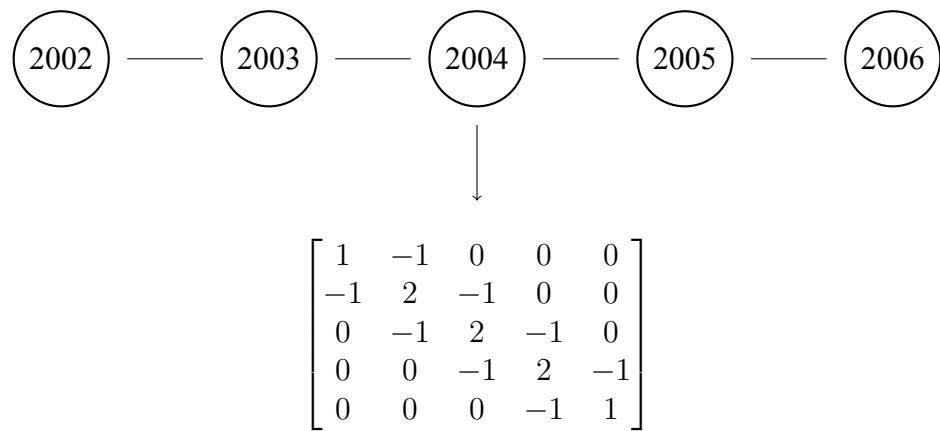


Figura 2.7: Matriz de estructura temporal de primer orden de los años 2002-2006 con las siguientes enumeraciones: 2002 (1), 2003 (2), 2004(3), 2005 (4) y 2006 (5).

### 2.6.2. Matriz de precisión e independencia condicionada

La idea de los modelos espaciotemporales es análoga a la explicada en la sección anterior: construir la matriz de precisión del vector de parámetros mediante las matrices de estructura espaciales y temporales. Cuando la matriz de precisión tiene la forma  $\tau T_m$ , las componentes del vector aleatorio presentan una estructura que hace que estos modelos se conozcan como paseos aleatorios (PA) o *random walks* (RW). En primer lugar, sea  $x = [x_1, \dots, x_n]$  un vector cualquiera, entonces por cómo está definida la matriz de estructura temporal  $T_m$  (ver Ecuación 2.35), se cumple que:

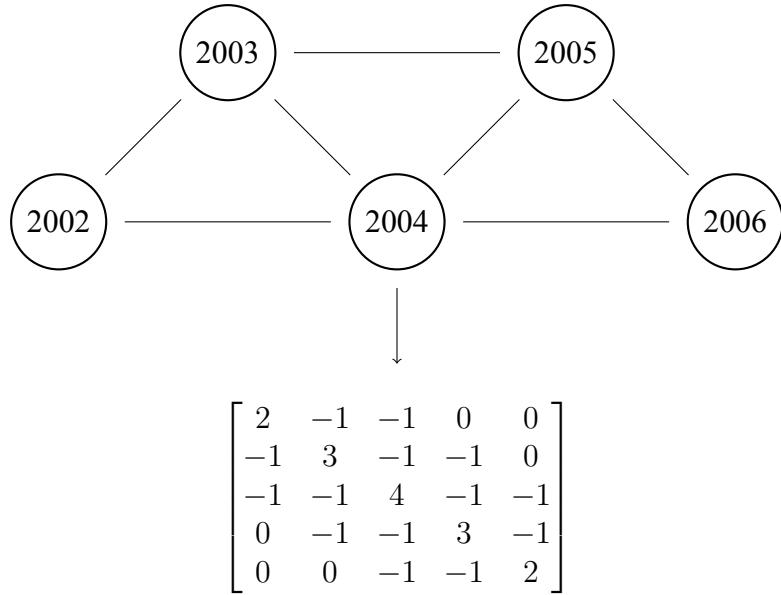


Figura 2.8: Matriz de estructura temporal de segundo orden de los años 2002-2006 con las siguientes enumeraciones: 2002 (1), 2003 (2), 2004 (3), 2005 (4) y 2006 (5).

$$x^T T_1 x = \sum_{i=2}^n (x_i - x_{i-1})^2, \quad x^T T_2 x = \sum_{i=3}^n ((x_i - x_{i-1}) - (x_{i-1} - x_{i-2}))^2 \quad (2.36)$$

Si se define el incremento  $\Delta x_i := \Delta^1 x_i := x_i - x_{i-1}$  y después, por recurrencia,  $\Delta^m x_i := \Delta^{m-1} x_i - \Delta^{m-1} x_{i-1}$ , entonces se tendrá que, de forma general:

$$x^T T_m x = \sum_{i=m+1}^n (\Delta^m x_i)^2 \quad (2.37)$$

Si ahora se toma un vector aleatorio  $X$  con distribución normal multivariante de media 0 y matriz de precisión  $\tau T_m$ , entonces su función de densidad de probabilidad será:

$$p(x) := p(X = x) = ((2\pi)^{-n} |\tau T_m|)^{1/2} \exp(-\frac{1}{2} x^T \tau T_m x) \quad (2.38)$$

Ahora bien, por la Ecuación 2.37,

$$p(x) \propto \exp(x^T \tau T_m x) = \exp(\tau \sum_{i=m+1}^n (\Delta^m x_i)^2) = \prod_{i=m+1}^n \exp(\tau(\Delta^m x_i)^2) \quad (2.39)$$

Es decir, la función de densidad de probabilidad de  $X$  es proporcional a

$$\prod_{i=m+1}^n \exp(\tau(\Delta^m x_i)^2) \quad (2.40)$$

lo cual es proporcional a la función de densidad de probabilidad del vector aleatorio  $\Delta^m = [\Delta^m x_i, \dots, \Delta^m x_{n-m}]$  de media 0 y matriz de precisión  $\tau I_{n-m}$ . Esto significa que se están considerando i.i.d. los incrementos  $\Delta^m x_i$ , con distribución normal de media cero y varianza  $1/\tau$ . En el caso de  $n = 1$  se tiene que:

$$\Delta x_i \sim N(0, 1/\tau) \iff x_i - x_{i-1} \sim N(0, 1/\tau) \quad (2.41)$$

lo que equivale a:

$$x_i = x_{i-1} + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1/\tau) \quad (2.42)$$

Es decir, el valor en el instante  $i$ -ésimo es el valor en el instante  $(i-1)$ -ésimo más una cierta cantidad aleatoria. Esta es la razón por la que se conocen como paseos aleatorios a los modelos que emplean estos efectos, ya que recuerda a un paseo por una ciudad en la que, en cada esquina, se elige de forma aleatoria la dirección en la que continuar la marcha. Pese a no parecer muy complejo, muchos fenómenos tienen un comportamiento similar a la de un paseo aleatorio, y por ello estos modelos tienen una gran importancia en el estudio de series temporales.

### 2.6.3. Modelos espaciotemporales para el *disease mapping*

El planteamiento de los modelos es totalmente análogo al caso espacial, cambiando únicamente la notación para incluir los diferentes instantes de tiempo. Así pues, para cada región  $i$  e instante  $t$ , dado el riesgo  $r_{it}$ , el número de casos observados  $O_{it}$  sigue una distribución de Poisson de parámetro  $\lambda_{it}$ . Puesto que  $\lambda_{it}$  coincide con la media de la variable aleatoria, según el indicador empleado se tendrá:

$$O_{it} \mid r_{it} \sim \begin{cases} Poi(\lambda_{it} = e_{it}r_{it}) & \text{con el riesgo relativo} \\ Poi(\lambda_{it} = N_{it}r_{it}) & \text{con la tasa} \end{cases} \quad (2.43)$$

Así pues, el objetivo será modelizar  $r_{it}$  mediante parámetros que tengan una estructura jerárquica basada en la matriz de estructura espacial, temporal, o una combinación de ambas. En particular,  $r_{it}$  se enlazará con un predictor lineal  $\eta_{it}$  mediante la función logaritmo, obteniendo el siguiente modelo logarítmico lineal de Poisson:

$$\log(r_{it}) = \eta_{it} \implies \begin{cases} \log(\lambda_{it}) = \eta_{it} + \log(e_{it}) & \text{con el riesgo relativo} \\ \log(\lambda_{it}) = \eta_{it} + \log(N_{it}) & \text{con la tasa} \end{cases} \quad (2.44)$$

### 2.6.4. Modelos aditivos

Este tipo de modelos simplemente consisten en combinar los modelos vistos en la Sección 2.5 con algún tipo de paseo aleatorio, habitualmente el de primer o segundo orden. Además de esto, también es posible añadir efectos temporales i.i.d., es decir, de la misma forma que se había hecho para cada región pero ahora para cada instante de tiempo. Así pues, el predictor lineal  $\eta_{it}$  se puede construir de varias maneras combinando los siguientes tipos de efectos espaciales y temporales, además del valor base  $\eta$ :

$$\left\{ \begin{array}{ll} u \sim N_S(0, (\tau_u R)^{-1}), & \text{espacial estructurado} \\ v \sim N_S(0, (\tau_v I_S)^{-1}), & \text{espacial no estructurado} \\ \xi \sim N_S(0, (\tau_\xi(\lambda R + (1 - \lambda)I_S))^{-1}), & \text{espacial Leroux} \\ \phi \sim N_T(0, (\tau_\phi T_m)^{-1}), & \text{temporal estructurado (PA } m\text{-ésimo)} \\ \gamma \sim N_T(0, (\tau_\gamma I_T)^{-1}), & \text{temporal no estructurado} \end{array} \right. \quad (2.45)$$

### 2.6.5. Modelos con interacción

Con los efectos espaciales y temporales expuestos hasta ahora, es posible hacer que los modelos puedan tener en cuenta diferentes regiones o instantes de tiempo. Sin embargo, los efectos espaciales se mantienen invariantes en el tiempo y viceversa, lo cual se puede solucionar mediante la introducción de efectos que dependan del espacio y del tiempo simultáneamente: las interacciones. Una manera de pensar en las interacciones es la siguiente: si se tienen datos de  $S$  regiones en  $T$  instantes de tiempo, en vez de considerar dos vectores de efectos de tamaño  $S$  y  $T$  (espacial y temporal, respectivamente), ahora se construirá un único vector de efectos  $\Delta := [\delta_{it}]$  de tamaño  $S \cdot T$ , donde cada uno de ellos se corresponderá con cada uno de los  $S \cdot T$  grupos formados por las combinaciones espaciotemporales. La matriz de precisión de este nuevo vector se construirá combinando las matrices de estructura de cada tipo y la matriz identidad, según las correlaciones que se deseen introducir. En particular, la matriz de precisión de las interacciones será el producto de Kronecker de dos matrices  $A = (a_{ij})$  y  $B = (b_{ij})$ , una operación matricial denotada por  $A \otimes B$ :

$$A \otimes B := \begin{bmatrix} a_{11}B & \dots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \dots & a_{mn}B \end{bmatrix} \quad (2.46)$$

Combinando matrices de precisión estructuradas y no-estructuradas en el espacio/tiempo, Knorr-Held [7] definió cuatro tipos de interacciones espacio-temporales (ver Tabla 2.1).

Interacción	Matriz	Correlación espacial	Correlación temporal
Tipo I	$I_S \otimes I_T$	✗	✗
Tipo II	$I_S \otimes T_m$	✗	✓
Tipo III	$R \otimes I_T$	✓	✗
Tipo IV	$R \otimes T_m$	✓	✓

Tabla 2.1: Tipos de interacciones espaciotemporales

Los efectos aleatorios expuestos hasta ahora, en combinación con las interacciones, ofrecen una serie de modelos muy potentes y con una gran flexibilidad. El próximo paso será introducir medidas de selección de modelos, para poder elegir entre unos y otros el más conveniente en cada situación.

## 2.7. Comparación de modelos

En la práctica, es necesario disponer de criterios que permitan evaluar y seleccionar los modelos que mejor se ajusten a los datos. A continuación, se describirán los principales criterios de comparación de modelos Bayesianos que se han empleado en este trabajo.

En primer lugar, el *Deviance Information Criterion* o DIC [8] es la medida de ajuste más usada en los modelos Bayesianos. El DIC se calcula como la suma de la media *a posteriori* de la desviación (una medida de la bondad del ajuste) y el número de parámetros efectivos (una medida de la complejidad del modelo). Por otro lado, el más reciente *Watanabe-Akaike Information Criterion* o WAIC [9], es un método para estimar la precisión de la predicción

puntual fuera de la muestra de un modelo Bayesiano ajustado. Gelman et al. (2004) [10] se recomienda el uso del WAIC por delante del DIC, ya que a diferencia de éste, el WAIC es invariante a las parametrizaciones y funciona también con modelos singulares. Finalmente, el *Logaritmic Score* o LS [11] es una regla de puntuación para comparar modelos en términos de su rendimiento predictivo. En los tres casos, valores más bajos apuntan a un mejor ajuste del modelo.

## 2.8. *Integrated nested Laplace approximations (INLA)*

El método INLA propuesto recientemente por Rue et al. (2009) [1] es un algoritmo determinista para inferencia Bayesiana basado en aproximaciones de Laplace anidadas integradas. Este método está especialmente diseñado para modelos Gaussianos latentes, una subclase de modelos de regresión aditivos estructurados que son lo suficientemente flexibles para usarse en muchos tipos de aplicaciones. Muchos de estos modelos admiten propiedades de independencia condicional que conducen a matrices de precisión dispersas, e INLA aprovecha esto para acelerar el cálculo. Esto permite hacer inferencia Bayesiana sin ejecutar largos y complejos algoritmos de *Markov chain Monte Carlo* (MCMC). En [1] o [3] se pueden encontrar los detalles sobre el funcionamiento de INLA.

La metodología INLA está implementada en el software estadístico gratuito R, a través de la librería R-INLA. La documentación del paquete, diversos manuales y un foro de discusión están disponibles en el sitio web de R-INLA: <http://www.r-inla.org/>.

# Capítulo 3

## Ilustración con datos reales

En este capítulo se ilustrará empleando datos reales la aplicabilidad de los modelos presentados en la sección anterior. Para ello, se emplearán datos de incidencia y mortalidad de cáncer de pulmón en la isla de Gran Bretaña en el periodo 2002-2019. El código y los datos para poder reproducir los resultados de este capítulo están disponibles en [Github](#), pudiendo trabajar con otros tipos de cáncer más allá del cáncer de pulmón.

### 3.1. Fuente de datos

Los datos de cáncer de Gran Bretaña recogen, para cada región, año, sexo y tipo de cáncer, el número de nuevos casos observados y el número de muertes correspondiente. El territorio bajo estudio se corresponde al de tres de las cuatro naciones que componen el Reino Unido: Inglaterra, Gales y Escocia, las cuales abarcan toda la isla de Gran Bretaña, además de pequeñas islas adyacentes. Pese a que las tres naciones pertenecen a un mismo país, el sistema nacional de salud de cada una de ellas ([NHS England](#), [NHS Wales](#) y [NHS Scotland](#)) funciona de manera independiente, por lo que los datos se han recopilado por separado y se han unido en una misma base de datos. Además de esto, los datos de población para cada año, región y

sexo se han recopilado para cada nación (*Population England*, *Population Wales*, y *Population Scotland*) y se han añadido a la misma. En cuanto a las regiones bajo estudio, existen diversas divisiones del territorio en cada una de las tres naciones, escogiendo en este caso las siguientes: Inglaterra se ha dividido a nivel de *clinical commissioning group* (106 regiones), Gales a nivel de *local authority* (22 regiones) y Escocia a nivel de *health board* (14 regiones), un total de 142 regiones (ver Figura 3.1).

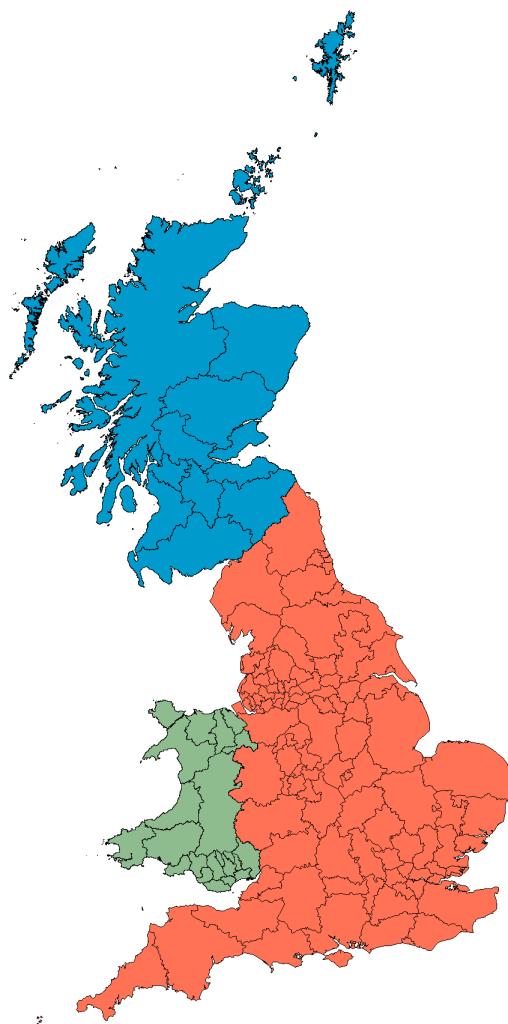


Figura 3.1: Subdivisión final de los territorios: 106 *clinical commissioning groups* ingleses (rojo), 22 *local authorities* galeses (verde) y 14 *health boards* escoceses (azul).

En la Tabla 3.1 se puede ver el encabezado de la base de datos final, en la que tanto las regiones como los tipos de cáncer vienen codificados, en el segundo caso según la déci-

ma versión de la Clasificación Internacional de Enfermedades (CIE-10). Además de esto, la descripción de cada código se puede ver en la Tabla A.1 del Apéndice A.

Región	Año	Sexo	CIE-10	Casos	Muertes	Población
E38000006	2002	Mujer	C91-C95	14	4	112309
E38000006	2002	Mujer	C43	13	2	112309
E38000006	2002	Mujer	C67	19	10	112309
E38000006	2002	Mujer	C50	145	37	112309
E38000006	2002	Mujer	C53	12	8	112309
E38000006	2002	Mujer	C18-C20	62	36	112309

Tabla 3.1: Estructura de la base de datos objeto de estudio.

Durante el resto del capítulo se trabajará con los datos de cáncer de pulmón (CIE-10 C33-C34), empleando los de incidencia en mujeres (nuevos casos observados) para ilustrar los modelos descritos en el Capítulo 2. Este tipo de cáncer es uno de los más comunes en la región, con una media de 43002 nuevos casos y 33782 muertes por año entre ambos sexos en el periodo 2002-2019.

## 3.2. Análisis descriptivo

Antes de aplicar los modelos, se tratará de representar la distribución geográfica de la enfermedad mediante las medidas clásicas propuestas en el Capítulo 2. A modo ilustrativo, se han tomado los datos observados en el año 2019, representando directamente los casos observados en cada región (Figura 3.2 izquierda). Como se comentó, esto no refleja bien la distribución geográfica de la enfermedad, ya que algunas regiones tienen mayor población que otras y eso puede hacer que aumente el número de casos (Figura 3.2 centro). La principal solución a este problema es calcular las tasas crudas por cien mil habitantes, tal y como se ha indicado en la Ecuación 2.1 (Figura 3.2 derecha). Los patrones de casos observados y población son prácticamente idénticos, mientras que cambia totalmente al calcular la tasa cruda.

De haber tenido los datos desglosados por grupos de edad, se podría haber calculado o

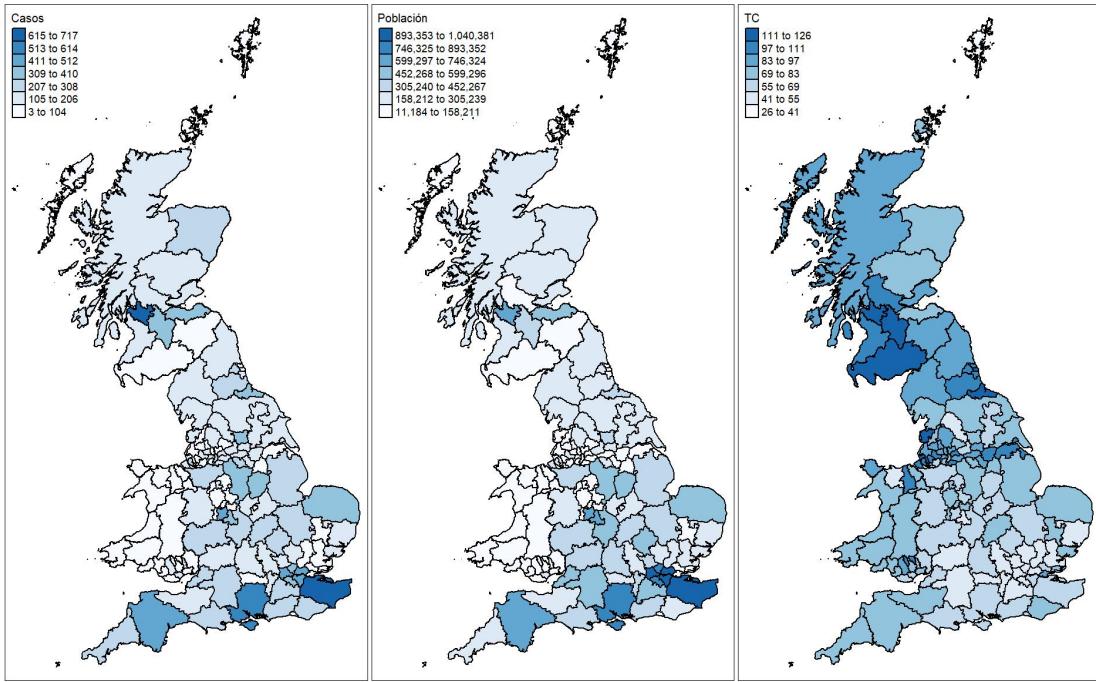


Figura 3.2: Casos observados, población y tasa cruda por cien mil habitantes de cáncer de pulmón en mujeres (año 2019).

bien la tasa cruda estandarizada mediante alguna población estándar, o bien los casos esperados y con ellos el ratio de incidencia estandarizado. Esto hubiera corregido posibles sesgos por edad, ya que poblaciones más envejecidas podrían presentar mayores tasas.

### 3.3. Modelos espaciales

En esta sección se ilustrarán algunos de los modelos espaciales presentados en la Sección 2.5, analizando las tasas de incidencia por cáncer de pulmón en mujeres durante el año 2019. Recapitulando, se está asumiendo que los casos observados en cada región,  $O_i$ , se distribuyen según una distribución de Poisson condicionados a la tasa  $r_i$ . La media de la distribución,  $\lambda_i$ , en este caso se asume que es igual a  $r_i$  por la población, esto es,  $\lambda_i = N_i r_i$ . Por último, esta media se enlaza mediante la función logaritmo con un predictor lineal  $\eta_i$ , el cual estará compuesto por diversos efectos espaciales junto con un valor base  $\eta$ :

$$\eta_i = \eta + \text{parámetros espaciales} \quad (3.1)$$

Estos efectos espaciales se agrupan en vectores, cuya distribución *a priori* es diferente según el modelo empleado y, utilizando al método INLA, será posible obtener una aproximación de las distribuciones marginales *a posteriori* de cada efecto. Del mismo modo, también se podrán obtener estas mismas distribuciones para los diferentes hiperparámetros que presente cada modelo. Puesto que los modelos de Poisson mixtos descritos en el capítulo anterior sufren de problemas de identificación entre los efectos fijos y aleatorios, es necesario imponer restricciones de suma a cero adecuados sobre los efectos aleatorios espaciales (ver, por ejemplo, [12]).

Las matrices de estructura espaciales se construirán según la Ecuación 2.10, en la que dos regiones se consideran vecinas si comparten frontera. No obstante, debido a la geografía de la zona, existen conexiones que deberían ser consideradas pese a no cumplir esta condición. Este es el caso de las regiones insulares, las cuales se han conectado a las regiones pertenecientes a la isla de Gran Bretaña (la mayor de todas) con las que mayor relación tienen. Además de esto, existen regiones que, pese a no compartir frontera, quedan conectadas por grandes puentes que sortean las bahías o canales que las separan. La cartografía no es tan detallada como para considerar estas conexiones, así que se han añadido manualmente. En total han sido 9 las conexiones extra que se han considerado para la construcción de la matriz de estructura espacial pese a no cumplir la condición de compartir frontera. El grafo de vecindad final con el que se trabajará se puede ver en la Figura 3.3.

En primer lugar, se ajustará un modelo iCAR, ya que es el más sencillo dentro de los que emplean la matriz de estructura espacial. Como se ha visto en la Sección 2.5, el predictor lineal de este modelo se construye del siguiente modo:

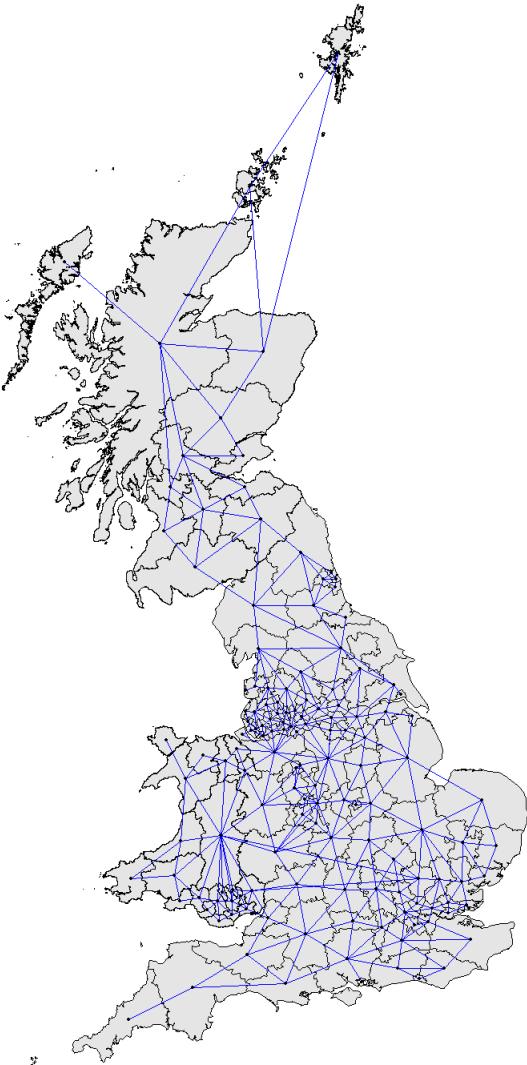


Figura 3.3: Grafo de vecindad de las regiones bajo estudio.

$$\eta_i = \eta + u_i \quad (3.2)$$

donde el vector de efectos  $u := [u_1, \dots, u_S]$  sigue una distribución normal multivariante de media cero y matriz de precisión en base a la matriz de estructura espacial,  $R$ . Esto es,

$$u := [u_1, \dots, u_S] \sim N_S(0, (\tau_u R)^{-1}) \quad (3.3)$$

Por último, el hiperparámetro  $\tau_u$  también tiene una distribución a priori, la cual se ha

establecido en este caso como no informativa. Concretamente, se asume una distribución uniforme  $[0, \infty[$  para la desviación típica. Es decir, este tipo de distribuciones *a priori* reflejan un total desconocimiento del hiperparámetro antes de observar los datos. Así pues, tras estimar el modelo con la estrategia de estimación Bayesiana INLA, se obtienen las aproximaciones de las distribuciones posteriores de todos los parámetros, incluyendo las del valor base  $\eta$ , los efectos espaciales  $u_i$  y el hiperparámetro  $\tau_u$ . En la Figura 3.4 se pueden ver estas distribuciones junto con cada una de las medianas.

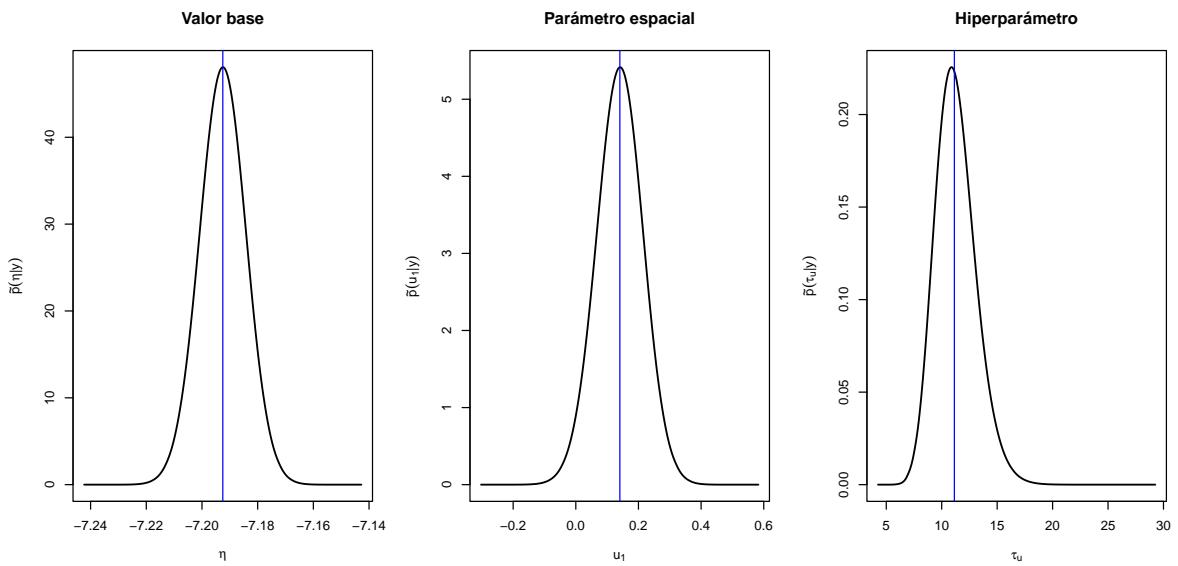


Figura 3.4: Aproximación de las distribuciones posteriores y medianas (azul) del valor base ( $\eta$ ), del efecto espacial correspondiente a la región de Barnsley ( $u_1$ ) y del hiperparámetro ( $\tau_u$ ).

El valor base  $\eta$  tiene una mediana de  $-7.19$ , cuya exponencial es de  $7.52 \cdot 10^{-4}$ , lo que quiere decir que la tasa ajustada base sobre la que oscilarán todas las regiones será de 75.2 casos por cien mil habitantes, algo que encaja con lo visto en tercer mapa de la Figura 3.2. Así pues, las regiones con efectos espaciales cuya mediana esté por debajo/encima de 0 tendrán una tasa ajustada menor/mayor que 75.2. Por ejemplo, en el caso de la primera región de la base de datos, Barnsley, la mediana de su parámetro espacial ( $u_1$ ) es de  $0.14 > 0$ , lo que quiere decir que la tasa ajustada en esa región estará por encima de 75.2, concretamente, 86.6. La variabilidad que

existe entre los parámetros espaciales viene explicada por el hiperparámetro  $\tau_u$ , la precisión. A mayor precisión, menor variabilidad y viceversa, teniendo en este caso  $\tau_u$  una mediana de 11.14.

Además del modelo iCAR, en la Sección 2.5 se han propuesto otros modelos espaciales. A continuación, se ajustarán todos ellos y se seleccionará el más conveniente utilizando distintos criterios de selección de modelos Bayesianos, como el DIC, el WAIC y el LS. En la Tabla 3.2 se pueden ver los indicadores para cada uno de los cuatro modelos ajustados. Los resultados de los modelos que emplean parámetros espacialmente estructurados son muy parecidos, observándose una mayor diferencia con los del modelo i.i.d., lo cual indica que existe cierta correlación espacial en los datos. Así pues, finalmente se seleccionará el modelo de Leroux para continuar el estudio, ya que ofrece una mayor flexibilidad.

Modelo	DIC	WAIC	LS
i.i.d.	1201.42	1172.06	669.51
iCAR	1181.41	1163.90	630.05
BYM	1181.82	1162.75	630.35
Leroux	1181.63	1161.67	630.01

Tabla 3.2: Comparación de modelos espaciales.

Además de la precisión  $\tau_\xi$ , el modelo de Leroux presenta un segundo hiperparámetro,  $\lambda$ , el cual controla el grado de dependencia espacial que se introduce en los efectos espaciales. Valores cercanos a 1 indicarán una estructura espacial mayor y viceversa, tal y como se puede ver en la Ecuación 2.32. La aproximación de la distribución *a posteriori* de este hiperparámetro se puede ver en la Figura 3.5 (izquierda), teniendo una media de 0.91 en este caso, por lo que los efectos espaciales tenderán a ser similares a los estimados en un modelo iCAR.

Una vez seleccionado el modelo final, el último paso es representar las tasas ajustadas de cada región en un mapa. Además de esto, también es posible obtener aproximaciones de las distribuciones *a posteriori* de estas tasas ajustadas, o lo que es lo mismo, de los predictores

lineales  $\eta_i$ . Gracias a esto, será posible calcular la probabilidad de que las tasas ajustadas superen un cierto umbral, obteniendo lo que se conocen como probabilidades de exceso o *exceedance probabilities*. Normalmente, es interesante que este umbral represente algún tipo de referencia, como por ejemplo una tasa cruda a nivel europeo o mundial. No obstante, por falta de datos, en este trabajo se usará como tasa de referencia el promedio de tasas crudas en el total del territorio. Para el año 2019, este promedio es de 68.13 casos por cien mil habitantes, pudiendo observar la probabilidad de que la tasa ajustada de cada región lo supere. En la Figura 3.5 (derecha) se puede ver la distribución del predictor lineal para la región de Barnsley, en la que se ha pintado el área bajo la curva que representa la probabilidad de que éste supere el umbral comentado, 0.93. Finalmente, en la Figura 3.6 se puede observar el mapa de tasas ajustadas (izquierda) y el de probabilidades de exceso empleando el umbral de 68.13 casos por cien mil habitantes.

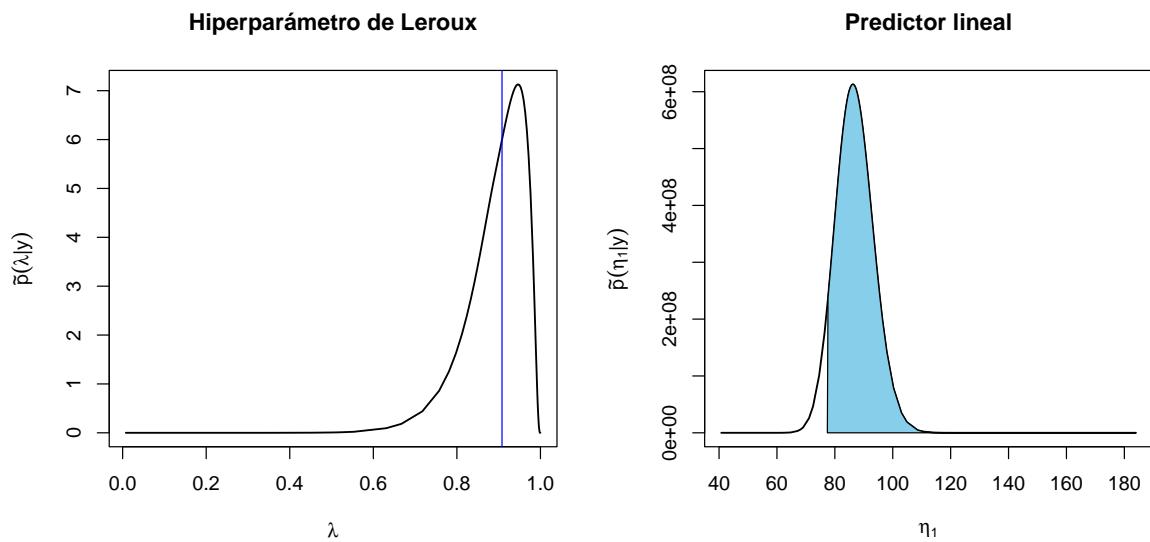


Figura 3.5: Aproximación de las distribuciones posteriores y medianas (azul) del hiperparámetro de Leroux ( $\lambda$ ) y del predictor lineal correspondiente a la región de Barnsley ( $\eta_1$ ).

Hay que recordar que en el mapa de tasas ajustadas se está resumiendo toda una distribución con la mediana, mientras que en el mapa de probabilidades se está teniendo en cuenta toda la distribución, ya que ahora también entra en juego cómo esté de sesgada. Si se com-

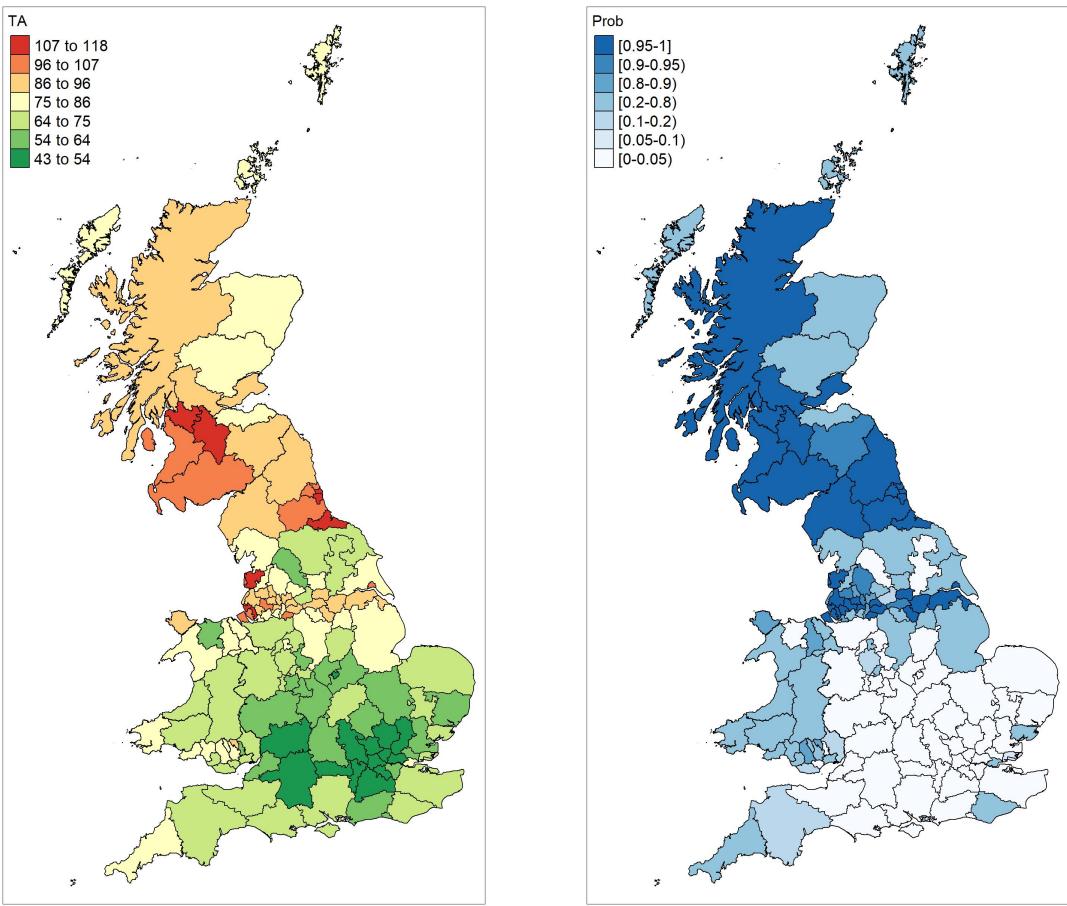


Figura 3.6: Mapa de tasas ajustadas y probabilidades de exceso obtenidas con el modelo de Leroux.

paran estos resultados con los obtenidos en el análisis descriptivo, se puede ver que las tasas crudas y las tasas ajustadas son muy similares, lo que nos indica que, probablemente, en este caso los indicadores clásicos son unos buenos estimadores de la distribución espacial de la tasa de incidencia. Sin embargo, en el siguiente capítulo veremos la necesidad de utilizar modelos espaciotemporales que suavicen las tasas crudas induciendo correlación espacial y/o temporal en los datos.

### 3.4. Modelos espaciotemporales

El planteamiento es similar al utilizado en el caso espacial, añadiendo un segundo subíndice,  $t$ , que representará el tiempo. Así pues, el predictor lineal  $\eta_{it}$  se construirá ahora con efectos espaciales y temporales, además del valor base  $\eta$  y, en algunos casos, un término de interacción espaciotemporal:

$$\eta_{it} = \eta + \text{parámetros espaciales} + \text{parámetros temporales} + \text{interacción} \quad (3.4)$$

Al igual que los espaciales, los efectos temporales también se agrupan en vectores aleatorios, cuya distribución puede ser i.i.d. o presentar la estructura de un paseo aleatorio, normalmente de primer o segundo orden. Lo mismo ocurre con las interacciones, cuyas matrices de precisión se construyen según se explica en la Tabla 2.1. Las distribuciones *a posteriori* de cada uno de los parámetros e hiperparámetros se pueden obtener del mismo modo que en el caso espacial.

A continuación, se ilustrarán los modelos empleando datos de cáncer de pulmón en mujeres en el periodo 2002-2019, ajustando varios modelos y seleccionando el mejor en base a los criterios de selección de modelos Bayesianos DIC, WAIC y LS. En concreto, los modelos propuestos serán los resultado de combinar, el modelo de Leroux para los efectos aleatorios espaciales, junto con un paseo aleatorio de orden 1 y 2 (PA1 y PA2) para los efectos aleatorios temporales, además de los cuatro tipos de interacciones. En la Tabla 3.3 se pueden ver los indicadores obtenidos para los 10 modelos propuestos, presentando el modelo Leroux + PA1 + TipoII el mejor ajuste (valores más bajos).

Así pues, una vez elegido el modelo con el mejor ajuste, las aproximaciones de las distribuciones *a posteriori* de todos los hiperparámetros se pueden ver en la Figura 3.7. El valor base  $\eta$  tiene una mediana de -7.33, cuya exponencial es de  $6.56 \cdot 10^{-4}$ , así que en este caso

Modelo	DIC	WAIC	LS
Leroux + PA1	20111.89	20177.72	10089.68
Leroux + PA1 + TipoI	19798.92	19804.59	9989.47
<b>Leroux + PA1 + TipoII</b>	<b>19592.54</b>	<b>19626.67</b>	<b>9831.70</b>
Leroux + PA1 + TipoIII	19813.52	19874.87	9985.88
Leroux + PA1 + TipoIV	19601.15	19638.49	9833.77
Leroux + PA2	20110.36	20172.18	10086.80
Leroux + PA2 + TipoI	19794.78	19803.30	9987.36
Leroux + PA2 + TipoII	19652.53	19703.65	9864.12
Leroux + PA2 + TipoIII	19813.71	19876.43	9985.47
Leroux + PA2 + TipoIV	19601.15	19638.49	9833.77

Tabla 3.3: Comparación de modelos espaciotemporales.

todas las tasas ajustadas para cada par región-año estarán alrededor de 65.6 casos por cien mil habitantes. El parámetro  $\lambda$  tiene en este caso una mediana de 0.93, por lo que de nuevo los efectos espaciales tenderán a ser similares a los estimados en un modelo iCAR.

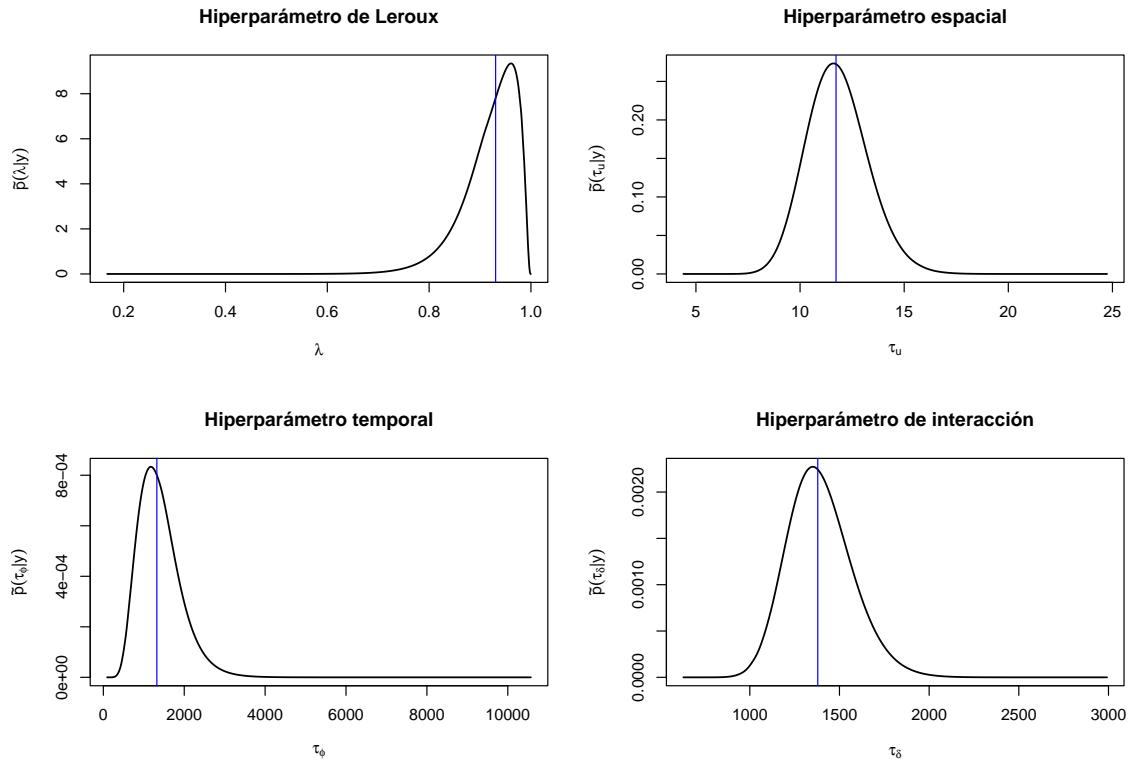


Figura 3.7: Aproximación de las distribuciones posteriores y medianas (azul) del hiperparámetro de Leroux ( $\lambda$ ) y de los hiperparámetros espacial ( $\tau_u$ ), temporal ( $\tau_\phi$ ) y de interacción ( $\tau_\delta$ ).

El último paso será representar las tasas ajustadas para cada año y región, además de las probabilidades de exceso. En este caso, los umbrales se calcularán para cada año del mismo

modo que en el caso espacial, empleando como tasa de referencia final el promedio de los umbrales de todos los años, siendo en este caso de 68.13 casos por cien mil habitantes. Por cuestiones de espacio, se presentan los mapas alternos de dos en dos años: desde 2002 hasta 2018. En la Figura 3.8 se pueden ver los mapas de tasas ajustadas (izquierda) y de probabilidades de exceso (derecha).

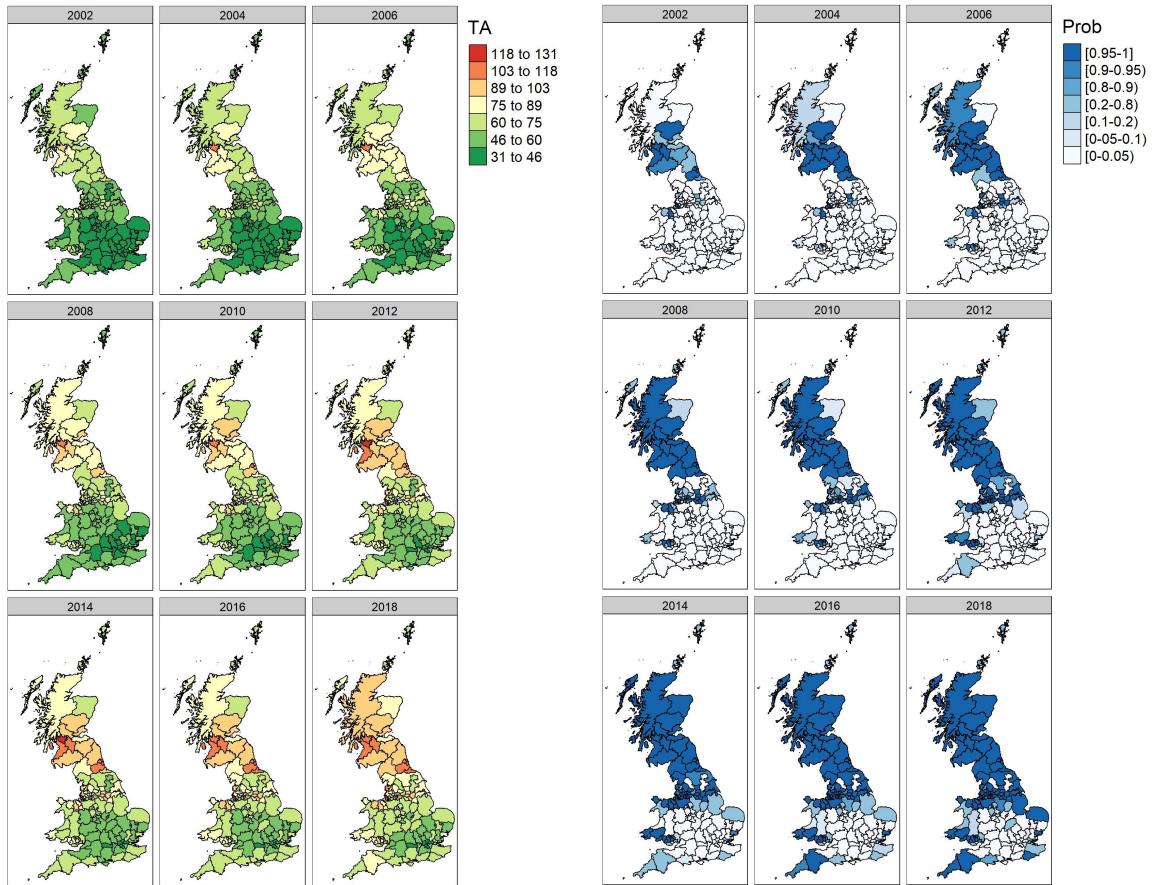


Figura 3.8: Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con el modelo Leroux + PA1 + TipolII.

Una de las ventajas de los modelos espaciotemporales es que, una vez impuestas las restricciones de suma a cero que permiten resolver los problemas de identificación entre los efectos fijos y aleatorios del modelo, se pueden definir los denominados patrones espaciales (comunes para todo el periodo) y temporales (comunes para todas las regiones). Para ello, dado el predictor lineal del modelo Leroux + PA1 + TipolII:

$$\eta_{it} = \eta + \xi_i + \phi_t + \delta_{it} \quad (3.5)$$

el patrón espacial se corresponderá con la parte  $\eta + \xi_i$ , mientras que el temporal lo hará con  $\eta + \phi_t$ . En la Figura 3.9 se han representado ambos patrones, añadiendo en el caso temporal el intervalo de credibilidad (IC) al 95 %.

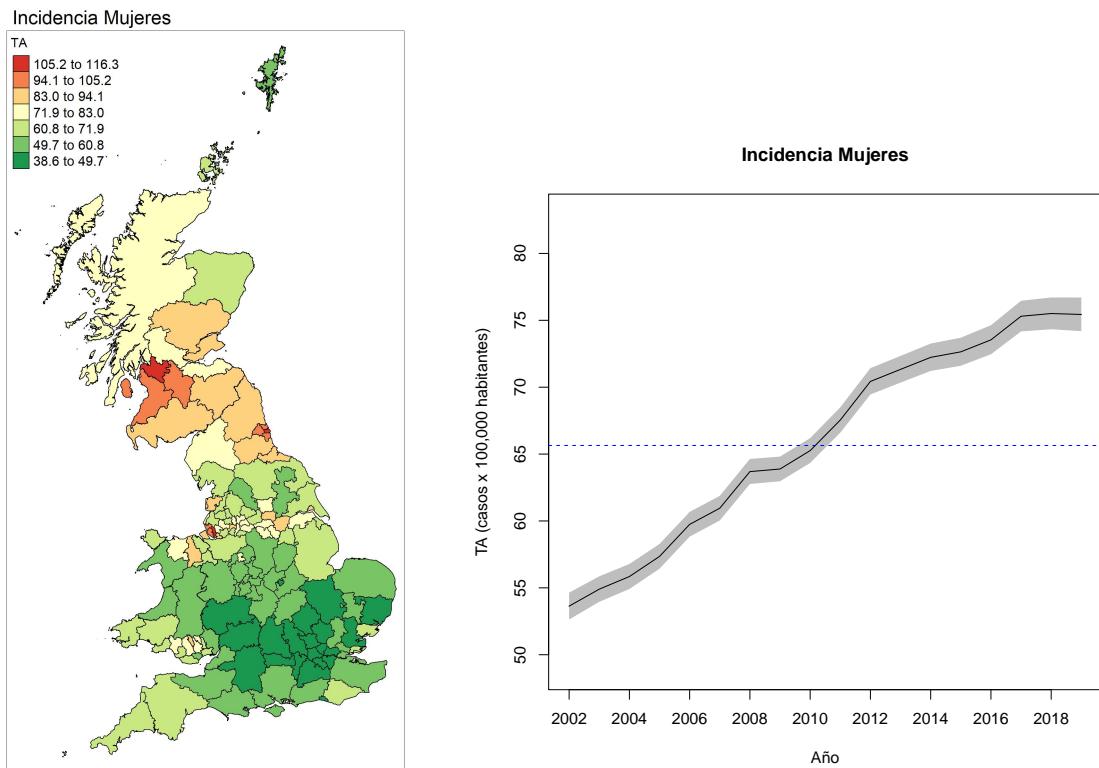


Figura 3.9: Patrón espacial (izquierda) y temporal (derecha) obtenidos con el modelo Leroux + PA1 + TipolII.

### 3.5. Análisis de la incidencia y mortalidad de cáncer de pulmón en Gran Bretaña (2002-2019)

En esta última sección, se llevará a cabo un análisis comparativo por sexo de la incidencia y mortalidad de cáncer de pulmón (CIE-10 C33-C34) en Gran Bretaña. Para ello, se tomarán

los datos de hombres y mujeres en el periodo 2002-2019, aplicando modelos espaciotemporales y comparando sus resultados. En primer lugar, para cada uno de los cuatro conjuntos de datos, se ajustarán los 10 modelos espaciotemporales propuestos en la Sección 3.4, seleccionando el más conveniente en base a los criterios de selección de modelos Bayesianos. En las Tablas B.1, B.2, B.3 y B.4 del Apéndice B se pueden ver los resultados obtenidos para los cuatro conjuntos de datos analizados. Para todos ellos, los criterios de selección de modelos apuntan al modelo Leroux + PA1 + TipoII.

Una vez seleccionados los modelos, el primer paso será comparar algunos de los efectos e hiperparámetros. Por ejemplo, las medianas de los valores base para los datos de incidencia de hombres y mujeres son, respectivamente, -7.09 y -7.33, lo que se corresponde con unas tasas ajustadas globales de 83.4 y 65.6, por lo que parece haber una mayor incidencia global en el sexo masculino. En cuanto a la mortalidad, parece también inclinarse hacia este sexo, con unas medianas de -7.3 y -7.58 en hombres y mujeres, respectivamente. Esto se corresponde con unas tasas ajustadas globales de 67.2 y 50.8, menores que las de incidencia, tal y como cabría esperar. En cuanto a los hiperparámetros de Leroux, todos son cercanos a 1. Concretamente, las medianas para los datos de hombres y mujeres son, respectivamente, 0.92 y 0.93 para la incidencia y 0.92 y 0.93 para la mortalidad. De nuevo los efectos espaciales tenderán a ser similares a los estimados en un modelo iCAR.

En un segundo paso, se mostrarán los cuatro patrones espaciales y temporales, ya que de esta manera será más fácil comparar los resultados para cada conjunto de datos. Así pues, en la Figura 3.11 se pueden ver los patrones espaciales, mientras que en la Figura 3.10 se pueden ver los temporales. Además de esto, en el Apéndice C se pueden ver los mapas por año de tasas ajustadas y probabilidades de exceso para cada conjunto de datos. Como umbral se han usado las tasas crudas medias en el periodo 2002-2019, siendo de 85.12 y 68.13 casos por cien

mil habitantes para la incidencia y de 68.77 y 52.54 para la mortalidad de hombres y mujeres, respectivamente.

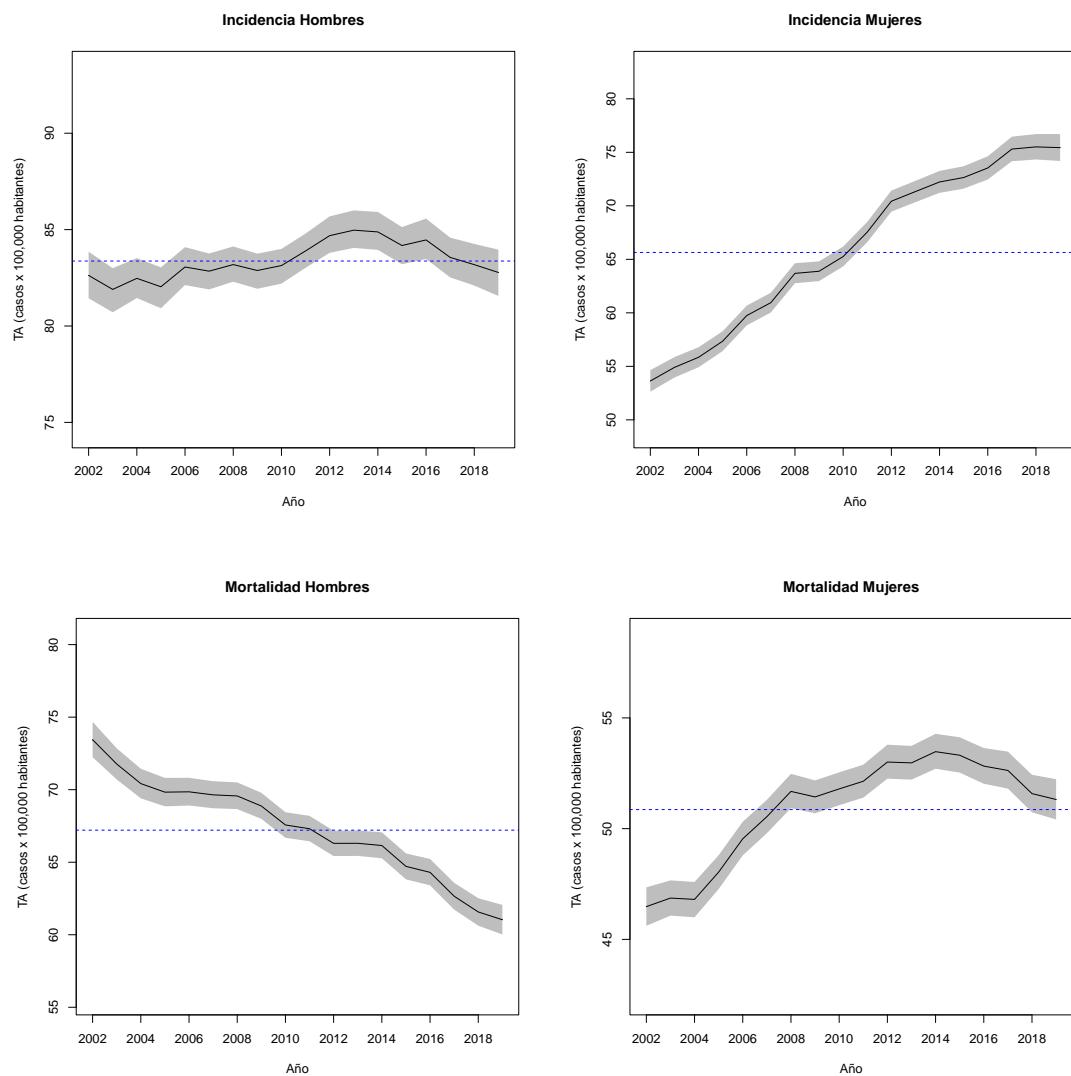


Figura 3.10: Patrones temporales obtenidos para cada conjunto de datos

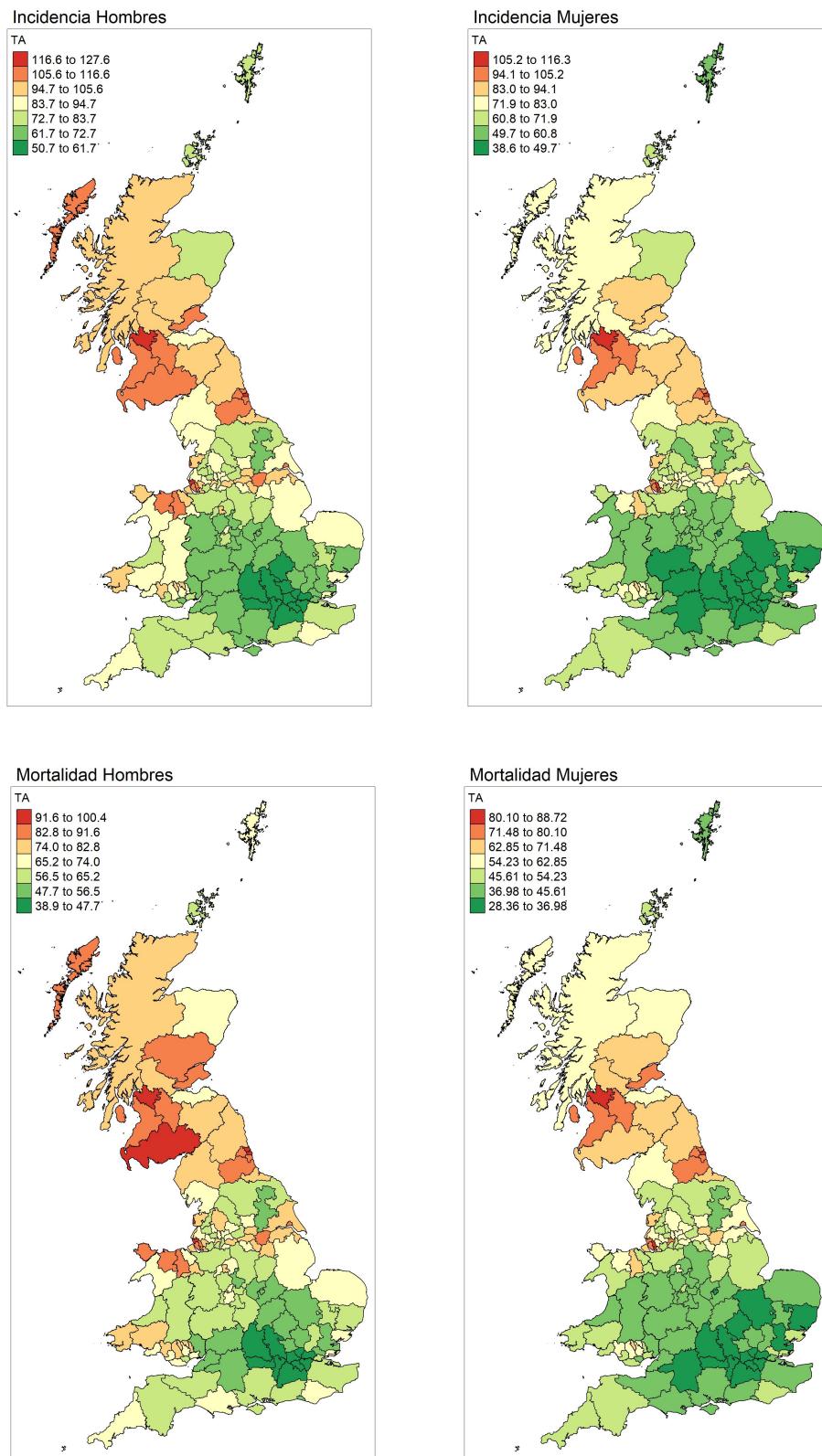


Figura 3.11: Patrones espaciales obtenidos para cada conjunto de datos

La distribución geográfica en los cuatro casos es muy similar, observándose las zonas de tasas más altas principalmente en los alrededores de Glasgow (Escocia), Newcastle upon Tyne (Inglaterra) y Liverpool (Inglaterra). Por el contrario, las zonas con menores tasas se aglomeran alrededor de Londres (Inglaterra). El consumo de tabaco es la principal causa a día de hoy de cáncer de pulmón, además de otros factores como la contaminación del aire [13]. En el atlas de 2007 sobre consumo de tabaco en Escocia [14], se puede ver que las zonas de mayor consumo se concentran en las regiones centrales de la nación, especialmente en las de Greater Glasgow and Clyde y Lanarkshire. Esto puede estar relacionado con los resultados obtenidos, ya que estas regiones presentan unas tasas ajustadas mayores que el resto. Además del consumo de tabaco, la contaminación del aire provocada por la actividad industrial ha podido propiciar el desarrollo del cáncer de pulmón en algunas regiones. Por ejemplo, la ribera del río Clyde, y en especial la ciudad de Glasgow (Escocia), ha sido históricamente una importante zona de fabricación de barcos, entre otras cosas [15]. Lo mismo ocurre con la ciudad de Newcastle upon Tyne [16] y Liverpool [17], en el norte de Inglaterra. Además de la contaminación del aire, puede que el desarrollo del cáncer de pulmón en estas zonas esté también relacionado con el amianto, un mineral que ha sido muy empleado en la industria naval durante los dos últimos siglos. El amianto, también conocido como uralita o asbestos, es un tipo de mineral con muy buenas propiedades aislantes, además de ser resistente y económico. Por esta razón, su uso se popularizó principalmente a partir de la Revolución Industrial, desconociendo sus riesgos para la salud. Cuando este material está en mal estado, comienza a descomponerse en microfibras, las cuales pueden ser inhaladas y acumularse en los pulmones, pudiendo ocasionar diversas enfermedades como el mesotelioma y el cáncer de pulmón [18]. De hecho, un estudio revela que, en 2016, el cáncer de pulmón ha sido una de las principales causas de muerte relacionadas con el amianto en el Reino Unido [19].

En cuanto al patrón temporal, se observan diferencias claras entre los conjuntos de datos.

Por un lado, la incidencia en hombres parece mantenerse más o menos estable durante todo el periodo, mientras que en el caso de las mujeres crece en todo momento, casi alcanzando en los últimos años los niveles en hombres. Por el otro lado, la mortalidad presenta una tendencia descendente en el caso de los hombres, mientras que ocurre lo contrario en el caso de las mujeres, aunque a partir del 2013 parece estar comenzando a descender.

Finalmente, se puede concluir que, durante el periodo 2002-2019 en Gran Bretaña, las zonas con mayores tasas se concentran en algunas de las principales zonas históricamente industriales de la isla. Las tasas en hombres han sido mayores que en mujeres tanto en incidencia como en mortalidad durante todo el periodo. No obstante, la tendencia en los datos de mujeres parece más preocupante, ya que, aunque la mortalidad parece estar comenzando a descender en los últimos años, la incidencia continúa aumentando.

# **Capítulo 4**

## **Conclusiones**

Los modelos espaciales y espaciotemporales han demostrado ser una gran herramienta en el ámbito de la representación cartográfica de enfermedades. Pese a que en ocasiones las medidas clásicas de estimación de riesgo pueden ser suficientes para estudiar una enfermedad, cuando se analizan enfermedades raras o regiones muy desagregadas el uso de modelos estadísticos que incorporan dependencia espacial y/o temporal resulta imprescindible. Estos modelos permiten obtener patrones espaciales y temporales, además de calcular las probabilidades de que las tasas ajustadas superen ciertos umbrales de referencia. El estudio del cáncer que se ha llevado a cabo en este trabajo, no obstante, podría haber sido de mayor calidad de haber tenido a disposición un mayor detalle en los datos. Por ejemplo, aumentando el número de regiones en el territorio y recogiendo los datos por grupos de edad, el análisis del cáncer de pulmón podría haber sido más detallado, además de no haber estado influenciado por la edad. Aún así, los resultados obtenidos muestran unos patrones claros, pudiendo ser de gran interés en el ámbito de la salud pública.

# Bibliografía

- [1] Rue, H., Martino, S. and Chopin, N. “Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations”. En: *Statistical Society: Series B (Statistical Methodology)* 71.2 (2009), págs. 319-392.
- [2] Francisco Doménech. *John Snow y el origen de una nueva medicina en los tiempos del cólera*. URL: <https://www.bbvaopenmind.com/ciencia/grandes-personajes/john-snow-origen-de-una-nueva-medicina-en-tiempos-del-colera/>. Consultado el 3 de marzo de 2022.
- [3] M. Blangiardo y M. Cameletti. “Spatial and Spatio-Temporal Bayesian Models with R-INLA”. En: *John Wiley Sons* (2015).
- [4] Scotland C. Leman. *Normal Theory and the Precision Matrix*. URL: <https://www.apps.stat.vt.edu/leman/VTCourses/Precision.pdf>. Consultado el 24 de junio de 2022.
- [5] Besag, J., York, J., and Mollié, A. “Bayesian image restoration, with two applications in spatial statistics.” En: *Annals of the Institute of Satitstical Mathematics* 43.1 (1991).
- [6] Leroux, B. G., Lei, X. and Breslow, N. “Estimation of disease rates in small areas. A new mixed model for spatial dependence”. En: *Statistical Models in Epidemiology, the Environment, and Clinical Trials* 116 (1999), págs. 179-191.

- [7] Knorr-Held, L. “Bayesian modelling for inseparable space-time variation in disease risk”. En: *Statistics in Medicine* 19.17–18 (2000), págs. 2555-2567.
- [8] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. “Bayesian measures of model complexity and fit”. En: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64.4 (2010), págs. 583-639.
- [9] Watanabe, S. “Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory”. En: *Journal of Machine Learning Research* 11.Dec (2010), págs. 3571-3594.
- [10] Gelman, A., Hwang, J., and Vehtari, A. “Understanding predictive information criteria for bayesian models”. En: *Statistics and Computing* 24.6 (2014), págs. 997-1016.
- [11] Gneiting, T. and Raftery, A. E. “Strictly proper scoring rules, prediction, and estimation”. En: *Journal of the American Statistical Association* 102.477 (2007), págs. 359-378.
- [12] Goicoa, T., Adin, A., Ugarte, M. D., and Hodges, J. S. “In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results”. En: *Stochastic Environmental Research and Risk Assessment* 32.3 (2018), págs. 749-770.
- [13] Danaei G, Vander Hoorn S, Lopez AD, Murray CJ, Ezzati M. “Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors”. En: *Lancet* 366.9499 (2005), págs. 1784-1793.
- [14] NHS Health Scotland, ISD Scotland and ASH Scotland. “An atlas of tobacco smoking in Scotland”. En: *NHS Health Scotland, Edinburgh* (2007).
- [15] Britannica, The Editors of Encyclopaedia. “River Clyde”. En: *Encyclopedia Britannica* (2019). URL: <https://www.britannica.com/place/River-Clyde>. Consultado el 10 de septiembre de 2022.

- [16] Britannica, The Editors of Encyclopaedia. “Newcastle upon Tyne”. En: *Encyclopedia Britannica* (2022). URL: <https://www.britannica.com/place/Newcastle-upon-Tyne-England>. Consultado el 10 de septiembre de 2022.
- [17] Lilly Edwards. “Liverpool’s Industrial Heritage”. En: *Encyclopedia Britannica* (2019). URL: <https://www.clickliverpool.com/features/31760-liverpools-industrial-heritage/>. Consultado el 10 de septiembre de 2022.
- [18] MAPFRE Servicio de Prevención. *Análisis retrospectivo de la exposición de trabajadores del sector de la construcción naval al amianto y de su relación causa-efecto con patologías del aparato respiratorio*. MAPFRE, Servicio de Prevención, 2008. URL: <https://books.google.es/books?id=a00uXwAACAAJ>.
- [19] Thompsons Solicitors. *Asbestos statistics and facts UK*. URL: <https://www.thompsons.law/support/legal-guides/asbestos-statistics-and-facts-uk>. Consultado el 10 de septiembre de 2022.

# Apéndice A

## Clasificación Internacional de Enfermedades CIE-10

CIE-10	Descripción
C00-C97 excl. C44	Todos los cánceres excepto el cáncer de piel no melanoma
C15	Cáncer de esófago
C16	Cáncer de estómago
C18-C20	Cáncer de colon y recto (colorrectal)
C22	Cáncer de hígado y conductos biliares intrahepáticos
C25	Cáncer de páncreas
C33-C34	Cáncer de tráquea, bronquios y pulmón
C43	Cáncer de piel tipo melanoma (melanoma)
C50	Cáncer de mama
C53	Cáncer de cérvix
C61	Cáncer de próstata
C67	Cáncer de vejiga
C91-C95	Leucemia

Tabla A.1: Descripción de la décima versión de la Clasificación Internacional de Enfermedades.

## Apéndice B

### Tablas de comparación de modelos espaciotemporales

En este apéndice se incluyen las tablas con los indicadores para los 10 modelos espaciotemporales propuestos. Cada una de ellas se corresponde con uno de los cuatro conjuntos de datos de cáncer de pulmón (CIE-10 C33-C34) considerados (indicencia y mortalidad en hombres y en mujeres).

Modelo	DIC	WAIC	LS
Leroux + PA1	20654.79	20719.62	10360.52
Leroux + PA1 + TipoI	20327.69	20334.74	10257.32
<b>Leroux + PA1 + TipoII</b>	<b>20128.37</b>	<b>20160.96</b>	<b>10098.48</b>
Leroux + PA1 + TipoIII	20345.04	20414.95	10257.08
Leroux + PA1 + TipoIV	20143.32	20181.36	10107.08
Leroux + PA2	20658.42	20721.40	10361.35
Leroux + PA2 + TipoI	20323.84	20334.48	10257.20
Leroux + PA2 + TipoII	20192.60	20236.22	10130.91
Leroux + PA2 + TipoIII	20344.91	20416.20	10257.59
Leroux + PA2 + TipoIV	20143.32	20181.36	10107.08

Tabla B.1: Comparación de modelos espaciotemporales para los datos de incidencia en hombres.

Modelo	DIC	WAIC	LS
Leroux + PA1	19859.95	19910.61	9955.99
Leroux + PA1 + TipoI	19628.09	19634.56	9882.73
<b>Leroux + PA1 + TipoII</b>	<b>19396.72</b>	<b>19404.50</b>	<b>9714.93</b>
Leroux + PA1 + TipoIII	19689.12	19761.03	9913.06
Leroux + PA1 + TipoIV	19421.40	19435.36	9730.13
Leroux + PA2	19855.73	19903.37	9952.28
Leroux + PA2 + TipoI	19627.83	19633.64	9880.50
Leroux + PA2 + TipoII	19437.07	19461.43	9739.46
Leroux + PA2 + TipoIII	19687.15	19758.72	9910.77
Leroux + PA2 + TipoIV	19421.40	19435.36	9730.13

Tabla B.2: Comparación de modelos espaciotemporales para los datos de mortalidad en hombres.

Modelo	DIC	WAIC	LS
Leroux + PA1	20111.89	20177.72	10089.68
Leroux + PA1 + TipoI	19798.92	19804.59	9989.47
<b>Leroux + PA1 + TipoII</b>	<b>19592.54</b>	<b>19626.67</b>	<b>9831.70</b>
Leroux + PA1 + TipoIII	19813.52	19874.87	9985.88
Leroux + PA1 + TipoIV	19601.15	19638.49	9833.77
Leroux + PA2	20110.36	20172.18	10086.80
Leroux + PA2 + TipoI	19794.78	19803.30	9987.36
Leroux + PA2 + TipoII	19652.53	19703.65	9864.12
Leroux + PA2 + TipoIII	19813.71	19876.43	9985.47
Leroux + PA2 + TipoIV	19601.15	19638.49	9833.77

Tabla B.3: Comparación de modelos espaciotemporales para los datos de incidencia en mujeres.

Modelo	DIC	WAIC	LS
Leroux + PA1	18989.72	19023.53	9512.38
Leroux + PA1 + TipoI	18854.42	18860.64	9470.31
<b>Leroux + PA1 + TipoII</b>	<b>18588.42</b>	<b>18572.43</b>	<b>9297.19</b>
Leroux + PA1 + TipoIII	18862.87	18904.68	9474.85
Leroux + PA1 + TipoIV	18594.48	18587.26	9303.91
Leroux + PA2	18988.18	19020.44	9510.77
Leroux + PA2 + TipoI	18850.24	18858.45	9468.76
Leroux + PA2 + TipoII	18618.20	18626.41	9320.45
Leroux + PA2 + TipoIII	18860.17	18902.02	9473.28
Leroux + PA2 + TipoIV	18594.48	18587.26	9303.91

Tabla B.4: Comparación de modelos espaciotemporales para los datos de mortalidad en mujeres.

## **Apéndice C**

### **Mapas por año de incidencia y mortalidad de cáncer de pulmón**

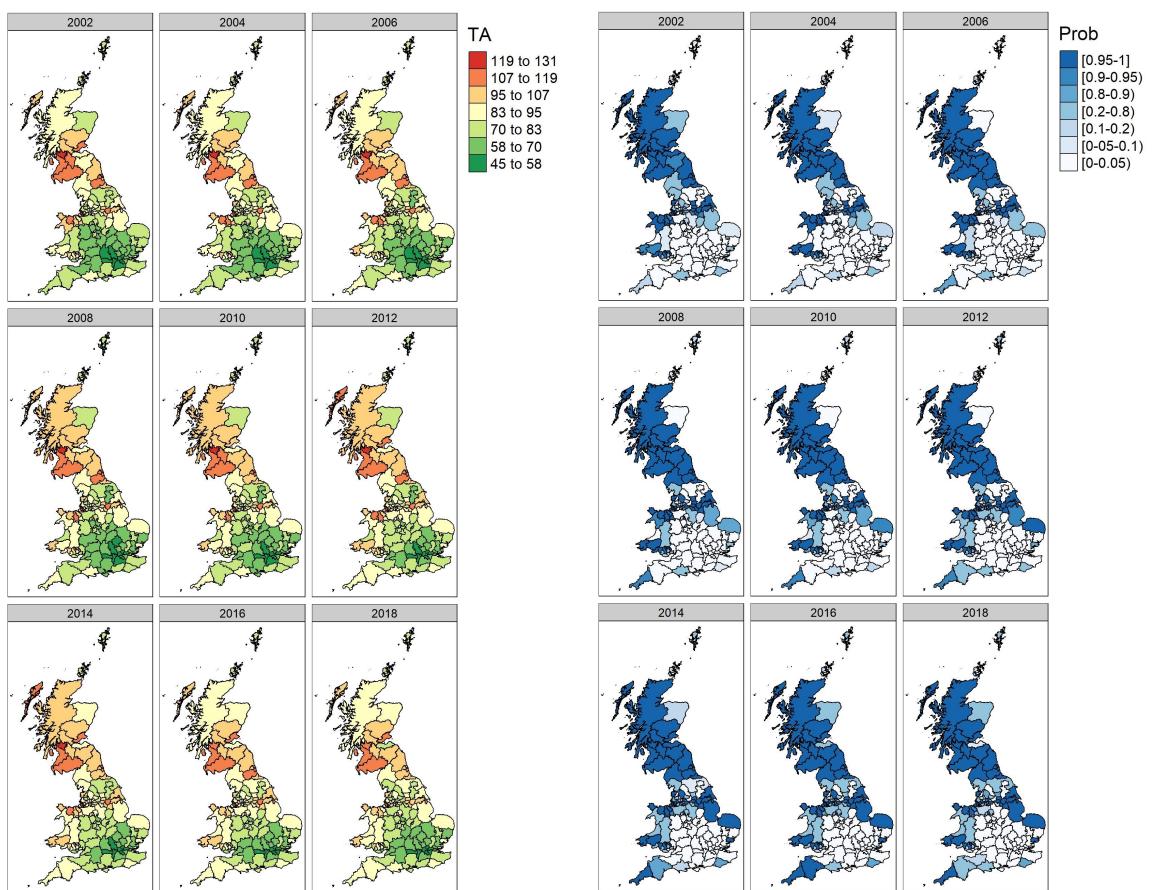


Figura C.1: Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de incidencia en hombres.

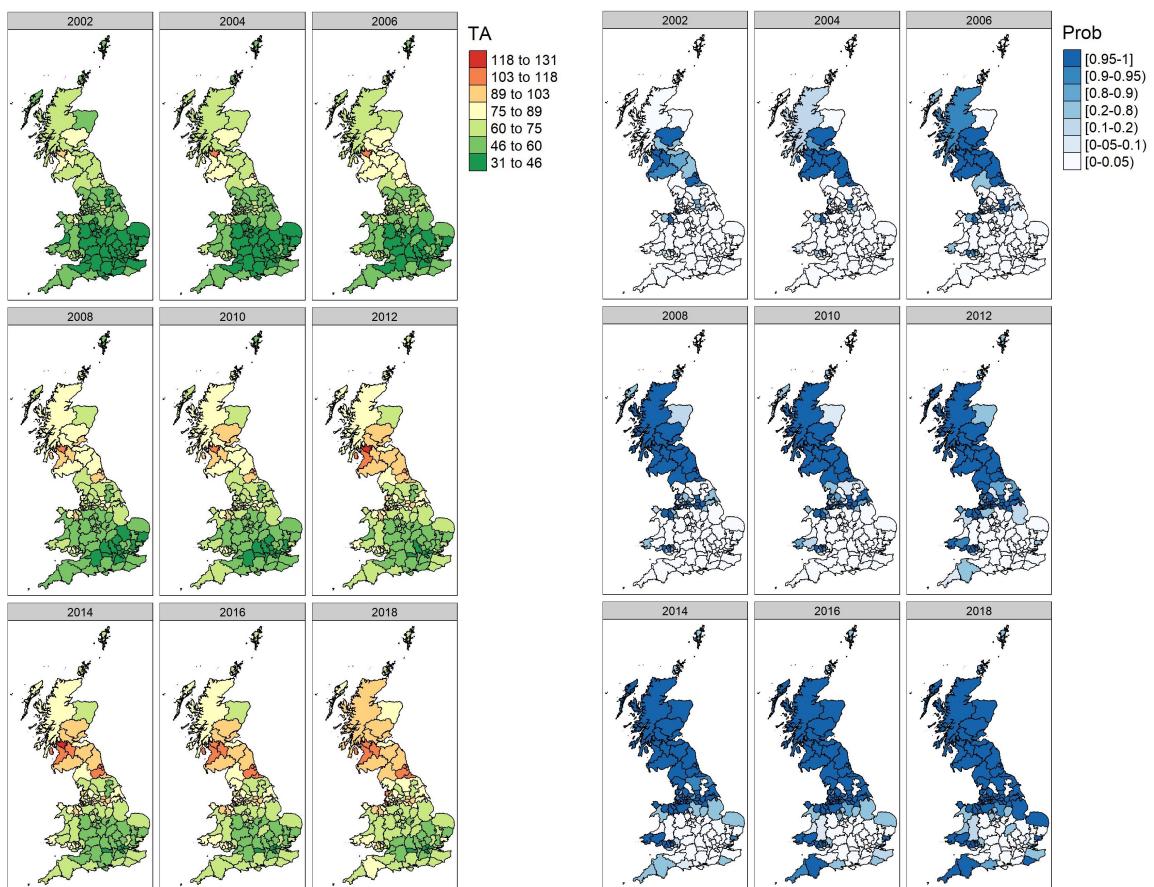


Figura C.2: Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de incidencia en mujeres.

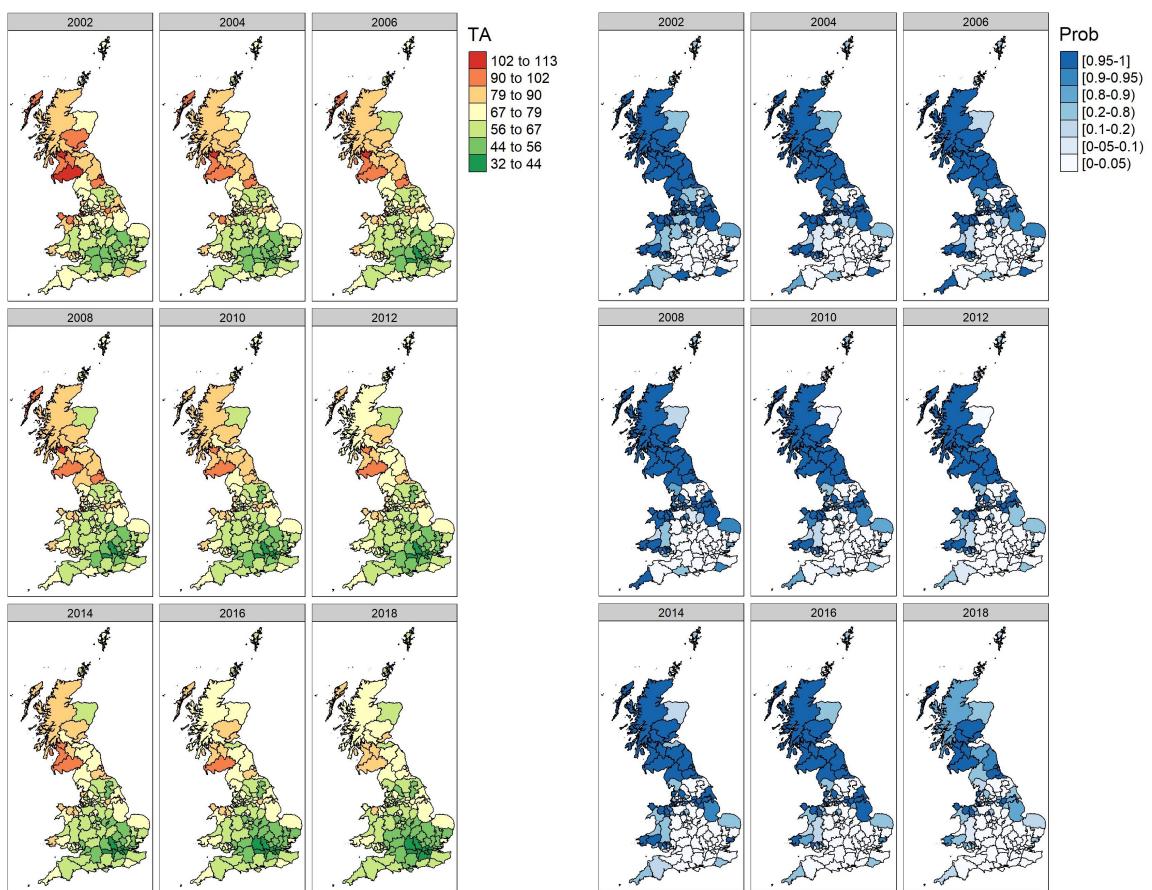


Figura C.3: Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de mortalidad en hombres.

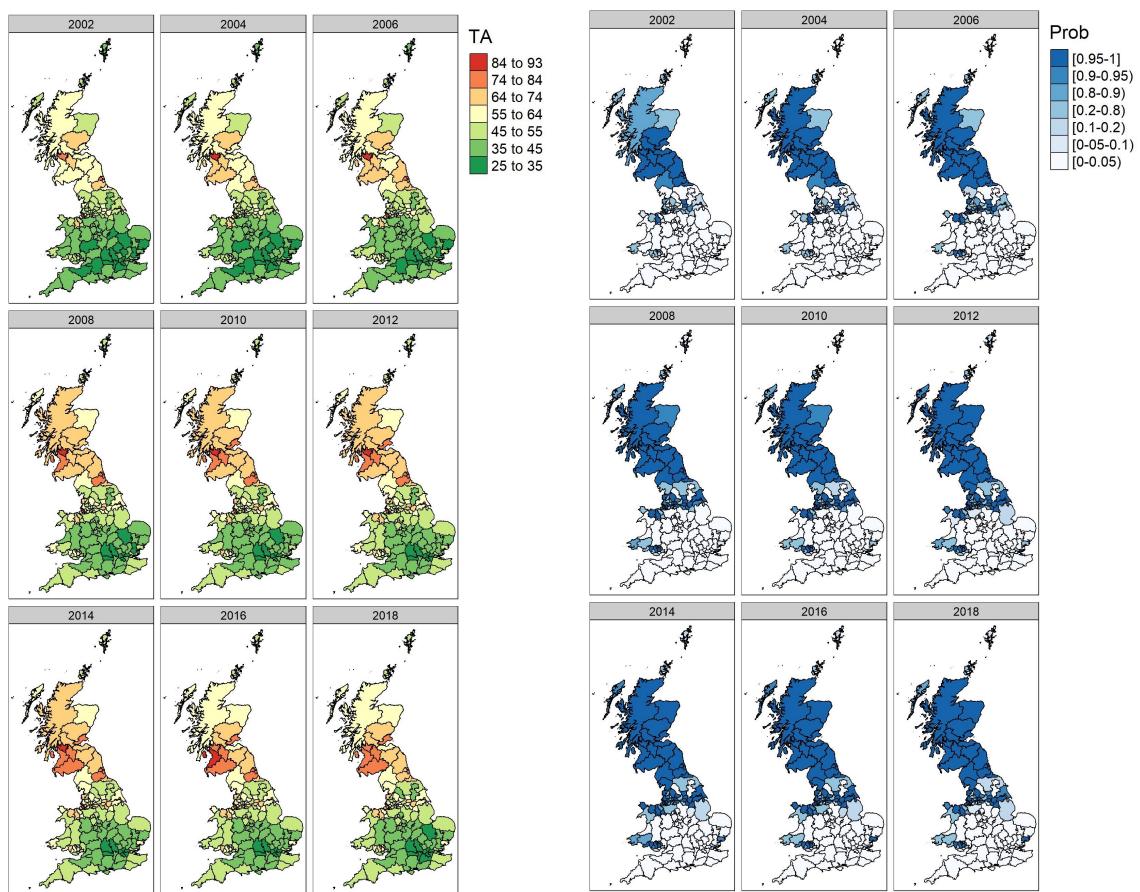


Figura C.4: Tasas ajustadas (izquierda) y probabilidades de exceso (derecha) obtenidas con los datos de mortalidad en mujeres.