Máster Universitario en Modelización e Investigación Matemática, Estadística y Computación 2023/2024

Trabajo Fin de Máster

Validación de modelos predictivos espacio-temporales de la incidencia y mortalidad por cáncer

---

Mikel Bergara Martinez

Tutores

Jaione Etxeberria Andueza

Aritz Adin Urtasun

Leioa, febrero de 2024

Resumen

Este Trabajo Fin de Máster se enmarca en el ámbito de la representación cartográfica de enfermedades o *disease mapping* en inglés, el cual tiene como propósito proporcionar herramientas de modelización estadística para analizar la distribución geográfica de una o varias enfermedades, y su evolución en el tiempo. Este trabajo tiene dos objetivos principales. En primer lugar, exponer teóricamente los fundamentos de la estadística Bayesiana mostrando las ventajas de utilizarla ante la probabilidad clásica, para después introducir los conceptos básicos de la modelización espacio-temporal de enfermedades crónicas como el cáncer, caracterizadas por sus medidas clásicas de estimación como las tasas crudas. Estos modelos estarán basados en la técnica INLA (*Integrated Nested Laplace Approximation*), un algoritmo determinista utilizado para realizar inferencia Bayesiana aproximada, produciendo resultados precisos y con menor coste computacional que los métodos Bayesianos de simulación MCMC (*Markov Chain Monte Carlo*). Posteriormente, se procederá a estudiar la aplicabilidad de los modelos espacio-temporales planteados con datos reales de cáncer de estómago en hombres de Inglaterra durante el periodo de tiempo 2001-2017. Para realizar la inferencia y lograr estimaciones de tasas suavizadas mediante este estudio, se utilizará el paquete `R-INLA` del software libre `R`. En segundo lugar, se ha desarrollado un proceso de validación de modelos estandarizado, el cual permitirá validar, comparar y seleccionar los mejores modelos espacio-temporales ajustados en base a medidas de validación como el MAE (*Mean absolute error*), RMSE (*Root Mean Squared Error*) o IS (*Interval Score*). Una vez escogido el mejor modelo, se realizarán predicciones a tres años para las tasas de incidencia de cáncer de estómago en hombres ingleses.

Abstract

This Master's Final Project is framed within the field of cartographic representation of diseases, commonly known as *disease mapping*, which has the purpose of providing statistical modeling tools to analyze the geographic distribution of a particular disease, and its evolution in time. This project has two main goals. First, to theoretically present the basics of the Bayesian statistics, giving evidence of its advantages compared to classical probability, and subsequently showing the foundation of the spatio-temporal modelization of chronic diseases such as cancer, which will be characterized by classical measures of estimates, for instance, crude rates. Those models will be based on a technique called INLA (*Integrated Nested Laplace Approximation*), a deterministic algorithm used for carrying out approximate Bayesian inference, giving precise results and better computational costs compared to the well-known MCMC (*Markov Chain Monte Carlo*), simulation-based Bayesian methods. Afterwards, the applicability of the proposed spatio-temporal models will be studied by carrying out a research with real stomach cancer data in men from England, within the period 2001-2017. To perform the necessary inference and obtain the posterior smoothed rate estimates through this study, the package `R-INLA` of the free software `R` will be used. Secondly, a standarized model validation process will be developed, which will allow validation, comparison and selection of the best spatio-temporal models fitted, considering some validation measures such as MAE (*Mean Absolute Error*), RMSE (*Root Mean Squared Error*) or even IS (*Interval Score*). Once the best possible model according to those measures is chosen, the final aim will be to carry out three years ahead predictions for stomach cancer incidence rates on males in England.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

El avance tecnológico de los últimos tiempos ha supuesto un incremento en la disponibilidad de datos espaciales y espacio-temporales utilizados en diversos campos de la investigación científica. Este Trabajo Fin de Máster se enmarca dentro del ámbito de la representación cartográfica de enfermedades, conocida en inglés como *disease mapping*, cuyo objetivo principal es proporcionar herramientas para el análisis de datos a nivel de área, con el fin de estimar patrones espaciales de mortalidad o incidencia por cáncer u otras enfermedades crónicas, e identificar así regiones que presenten un exceso de riesgo. El objetivo del *disease mapping* es proporcionar una representación de la distribución geográfica de una o varias enfermedades, permitiendo de este modo buscar posibles factores de riesgo subyacentes.

En este contexto, se desarrollan técnicas de modelización estadística que permiten estimar el riesgo o las tasas de enfermedades raras (con pocos casos) o en áreas poco pobladas, donde las medidas clásicas como la razón de mortalidad estandarizada o tasas crudas son extremadamente variables. De esta manera, las instituciones sanitarias pueden diseñar y evaluar estrategias de prevención de las enfermedades. No obstante, debido a la complejidad de la recolección, procesamiento y validación de las fuentes de información, los datos oficiales de incidencia o mortalidad están disponibles con dos o tres años de retraso respecto al año natural. Por esta razón, los procedimientos que permiten la predicción de las tasas o riesgos a corto plazo resultan de gran utilidad. En este trabajo, se desarrollarán métodos que permitan validar, comparar y seleccionar modelos adecuados para realizar predicciones a corto plazo (ver, por ejemplo, los artículos [1] o [2]).

Los modelos más utilizados en *disease mapping* son los modelos lineales generalizados mixtos (GLMM), los cuales se formulan tanto desde el punto de vista frecuentista como desde el Bayesiano, y generalmente permiten el suavizado de riesgos o tasas. Durante las últimas décadas, se han desarrollado numerosas propuestas de modelización y herramientas computacionales tanto desde el enfoque frecuentista como Bayesiano. En concreto, los modelos espaciales y espacio-temporales que se presentarán en este Trabajo Fin de Máster han sido desarrollados desde una perspectiva Bayesiana, la cual cuenta con una serie de ventajas respecto a la modelización clásica, permitiendo por ejemplo manejar estructuras complejas con datos faltantes, además de incorporar información previa o conocimientos

expertos en forma de distribuciones de probabilidad.

Tradicionalmente, la inferencia Bayesiana (cuyo fundamento se basa en el conocido Teorema de Bayes) se ha realizado utilizando diferentes métodos numéricos con el objetivo de aproximar las distribuciones de probabilidad. Uno de los algoritmos más conocidos y utilizados en este ámbito son los métodos de Monte Carlo basados en cadenas de Markov (técnicas denominadas por las siglas MCMC). No obstante, estos métodos basados en la simulación suelen resultar computacionalmente muy costosos, particularmente cuando se trata con modelos espacio-temporales complejos. Por ello, en este trabajo se utilizará una técnica alternativa conocida como INLA (*integrated nested Laplace approximation*), propuesta por Rue et al. (2009) [3]. Se trata de un método determinista para realizar inferencia Bayesiana aproximada ampliamente utilizado en el campo de la estadística espacial (ver, por ejemplo, los libros [4] y [5], o los artículos [6] y [7]). Este método reduce significativamente el tiempo computacional en modelos espaciales y espacio-temporales en comparación con las técnicas de MCMC y puede ejecutarse en el software libre `R` a través de la librería `R-INLA` [8].

Este Trabajo Fin de Máster tiene dos objetivos principales. En primer lugar, describir la técnica de inferencia Bayesiana aproximada INLA y su aplicabilidad para la estimación de tasas mediante el uso de modelos estadísticos espacio-temporales comúnmente utilizados en *disease mapping*. En segundo lugar, desarrollar un procedimiento estandarizado para validar, comparar y seleccionar modelos de predicción espacio-temporales en el ámbito de la representación cartográfica de enfermedades. Concretamente, se ha implementado una metodología que, una vez seleccionado el modelo óptimo, se realicen predicciones a un futuro a corto plazo, donde los datos oficiales de tasas de incidencia o mortalidad de las enfermedades puede que no estén disponibles. Para ilustrar los resultados, se emplean datos reales de incidencia por cáncer en las distintas regiones de Inglaterra durante el periodo 2001-2017. Además, mediante el desarrollo de este trabajo se pretende poner a prueba e interiorizar los conocimientos adquiridos en el máster gracias a asignaturas como Modelización Estadística, Minería de Datos, o, en general, cualquier conocimiento extra conseguido en lo que respecta a la modelización matemática y el uso del software libre `R`.

El trabajo se organiza de la siguiente manera. El Capítulo 2 presenta el marco teórico necesario para contextualizar el ejemplo práctico de capítulos posteriores. Se realiza una breve introducción a la estadística Bayesiana, comparándola con la probabilidad clásica, para después introducir el concepto de inferencia Bayesiana. A continuación, se presentan los fundamentos de los modelos jerárquicos, y también se analiza la teoría de la técnica INLA junto a las aproximaciones anidadas de Laplace en las cuales se basa. Una vez explicados estos conceptos, se introducen los modelos Gaussianos latentes y se muestra como se desarrolla la inferencia Bayesiana utilizando INLA. Por último, se explican las nociones necesarias de la modelización espacio-temporal para datos de área, y se proporcionan varios criterios de selección de dichos modelos que se emplearán para evaluar los resultados numéricos obtenidos.

En el Capítulo 3, se estudia la aplicabilidad de los modelos espacio-temporales planteados anteriormente, presentando datos reales de cáncer. Primeramente, se construye una base de datos partiendo de los datos publicados en abierto por el

Registro de Cáncer de Inglaterra, [9], y el *Office for National Statistics*, [10], y se lleva a cabo un problema real basado en datos de incidencia y mortalidad de diferentes tipos de cáncer en regiones de Inglaterra durante diferentes periodos de tiempo, realizando un análisis descriptivo de los datos para observar la evolución espacial, temporal y espacio-temporal de las tasas. Después, se ajustan diferentes modelos espacio-temporales de estimación de tasas, utilizando la técnica INLA mediante el paquete `R-INLA` del software `R`, y con ayuda de los diferentes criterios de selección, se interpretan los resultados numéricos obtenidos. Finalmente, se representan las estimaciones obtenidas bajo el mejor modelo seleccionado.

En el Capítulo 4, se implementa un procedimiento estandarizado de validación de los mejores modelos del capítulo anterior, el cual se basa en dividir los datos en subconjuntos de diferentes periodos, y ajustar los modelos a cada uno de estos subconjuntos. Partiendo de los modelos ajustados para cada una de las particiones de datos o configuraciones, se escoge el mejor modelo gracias a diferentes medidas de validación, y finalmente se realizan predicciones de tasas de incidencia a un futuro cercano (uno, dos y tres años) con este último modelo, interpretando los resultados.

En el Capítulo 5, el trabajo termina con las conclusiones finales y agradecimientos.

Capítulo 2

Formulación del problema espacio-temporal en el marco de la estadística Bayesiana e INLA

2.1. Introducción a la estadística Bayesiana

En esta primera sección, se busca introducir brevemente al lector los fundamentos de los métodos Bayesianos. Utilizados hoy en día en diversas áreas de investigación, la teoría Bayesiana se basa en razonar que la probabilidad es subjetiva, es decir, que cada individuo decide el grado de confianza acerca de un posible evento. Es por esto que, una ventaja del pensamiento subjetivo comparado con la probabilidad clásica, es que la premisa de que un evento tiene que ser repetible para poder calcular su probabilidad, no es necesaria.

Comencemos retomando la noción básica de probabilidad condicional. Dados dos eventos A y B, se define el evento *condicional* como el evento A bajo la condición del evento B, el cual suponemos que ya ha ocurrido (se denota $A|B$). De este modo, la *probabilidad condicional* se puede definir así:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

Es importante mencionar el concepto de la condicionalidad entre eventos, puesto que el teorema de Bayes que será de interés durante todo el trabajo se deduce directamente de la definición anterior:

$$P(B|A) = \frac{P(A|B) \times P(B)}{P(A)} \tag{2.2}$$

Este teorema se podría reescribir en caso de cumplirse condiciones de eventos disjuntos y utilizando el teorema de la probabilidad total. No obstante, profundizar en teoría sobre probabilidad clásica no es el objetivo del trabajo, por lo que si el lector desea consultar más detalles puede hacerlo, por ejemplo, en el libro [4]. De este modo, se supondrá para el resto del trabajo un conocimiento básico acerca de la Teoría de la Probabilidad.

2.1.1. Inferencia Bayesiana

El teorema 2.2 es el teorema de Bayes enunciado para eventos observables, sin embargo, también se estudia su uso para análisis estadísticos más generales, donde existen parámetros que son valores desconocidos, y se requiere la especificación de una distribución *a priori* de ellos, para obtener así la correspondiente distribución *a posteriori* (en el teorema 2.2 se especificaría la distribución *a priori* $P(B)$ para obtener $P(B|A)$ *a posteriori*). Al proceso que se detalla a continuación se le denomina *inferencia Bayesiana*, en el cual es fundamental diferenciar cantidades observables y desconocidas.

Sea X una variable aleatoria y θ un parámetro de interés. La incertidumbre de la variable X se mide utilizando una función de probabilidad o densidad (dependiendo si la variable aleatoria es contínua o discreta). Sea $V(\theta) = p(X = x|\theta) = p(x|\theta)$ la llamada función de *verosimilitud*, la cual especifica la distribución de los datos observados x bajo el parámetro θ (notar que estamos usando una p minúscula que denota una distribución de probabilidad; no confundirlo con la anterior P mayúscula que denotaba una probabilidad numérica de eventos).

Por otro lado, al haber adoptado un punto de vista Bayesiano, tenemos que el parámetro de interés θ es una cantidad desconocida, que se modelizará mediante una distribución de probabilidad *a priori* $p(\theta)$, antes de observar valores x de la variable aleatoria X. Si nos fijamos, la distribución $p(\theta)$ se puede definir como el grado de confianza previo a las observaciones que depositamos sobre el parámetro θ, es decir, nuestra hipótesis inicial acerca del parámetro. Es importante mencionar que más adelante, cuando se introduzca una estructura jerárquica de los datos, y una relación espacial o temporal entre parámetros, veremos que el conocimiento acerca del parámetro θ se modelizará condicionado a una serie de hiperparámetros que llamaremos ω, de tal manera que la distribución *a priori* se denotará como $p(\theta|\omega)$.

De este modo, se han introducido todos los conceptos necesarios para resolver el problema de inferencia Bayesiana que se plantea: teniendo la función de verosimilitud y la distribución *a priori* del parámetro, se obtiene la distribución *a posteriori*, utilizando el teorema de Bayes 2.2:

$$p(\theta|x) = \frac{p(x|\theta) \times p(\theta)}{p(x)} \tag{2.3}$$

Finalmente, teniendo en cuenta que $p(x)$ es la distribución que siguen los datos y funciona como constante de normalización, $p(\theta|x)$ (es decir, la distribución del parámetro de interés después de observar los datos) se puede expresar como proporcionalidad:

$$p(\theta|x) \propto p(x|\theta) \times p(\theta) \tag{2.4}$$

En este Trabajo Fin de Máster, se va a tratar con datos reales de incidencia de cáncer, por lo que la variable aleatoria X representará en este caso el número total de enfermos y se modelizará mediante una distribución de *Poisson*. Por otro lado, uno de los parámetros de interés al cual denominaremos una vez más θ, será la *tasa cruda*, una de las medidas de estimación clásicas más utilizadas, que será definida más adelante.

Una de las ventajas de trabajar con estructuras Bayesianas, es que se obtiene como resultado una distribución $p(\theta|x)$ *a posteriori* para los parámetros de interés.

De este resultado se pueden obtener indicadores estadísticos como la *media a posteriori*, la cual también será utilizada en los siguientes capítulos, y se calcula de esta manera (para un parámetro contínuo, si fuera discreto tendríamos una suma en vez de la integral):

$$E(\theta|x) = \int_{\theta} \theta p(\theta|x) d\theta, \quad \forall\theta \tag{2.5}$$

De un modo similar, se podría calcular la *mediana a posteriori* $\theta_{0,5}$, la cual se define como el valor que divide la distribución de probabilidad en dos mitades iguales:

$$p(\theta \leq \theta_{0,5}|x) = 0{,}5 \;\; y \;\; p(\theta \geq \theta_{0,5}|x) = 0{,}5 \tag{2.6}$$

Por último, también se pueden construir los llamados *intervalos de credibilidad*, que son similares a los intervalos de confianza para un enfoque frecuentista, por lo que es importante diferenciarlos. Los intervalos de confianza del $(100-\alpha)\,\%$ se interpretan como intervalos de los cuales el parámetro θ estará excluído un $\alpha\,\%$ de veces si se repitiera el mismo experimento varias veces. En cambio, el intervalo de credibilidad o IC del $(100-\alpha)\,\%$ solamente muestra la probabilidad *a posteriori* de que el parámetro de interés θ caiga dentro de dicho intervalo, es decir, $p(\theta \in IC|x) = \frac{100-\alpha}{100}$, para cierto valor de α. El cálculo de esta probabilidad es posible puesto que el parámetro de interés θ está sujeto a una distribución de probabilidad, de modo que es posible exponer hechos probabilísticos.

2.2. Modelos jerárquicos

A menudo, si se quiere construir un modelo estadístico se deberá caracterizar con una serie de parámetros θ. Por ejemplo, en el marco del *disease mapping*, si el objetivo es llevar a cabo un estudio del riesgo de una enfermedad en particular, se considerarán parámetros específicos para medir el efecto espacial o temporal de dicha enfermedad. En general, cuando se dispone de datos estructurados en diferentes grupos, como pudieran ser regiones de un área, periodos de tiempo, grupos de edad o sexo, se busca caracterizar cada uno de los grupos mediante un parámetro.

Supongamos que disponemos de unos datos u observaciones clasificados en J grupos numerados por $j \in \{1, \ldots, J\}$, y para los cuales n_j indica el número de elementos de cada grupo. Llamaremos, a su vez, y_{ij} a la observación $i \in \{1, \ldots, n_j\}$ del grupo j. En este caso, diremos que hay un *primer nivel* que se identifica con las observaciones y_{ij} , las cuales siguen una distribución de probabilidad adecuada (en el caso de incidencia de enfermedades, tendremos que los datos siguen una distribución de Poisson). Del mismo modo, llamaremos unidades del *segundo nivel* a cada uno de los J grupos en los que se clasifican los datos.

Volviendo a la noción de asignación de parámetros a cada uno de los J grupos, la modelización de datos se puede dar de diferentes formas. En un extremo está el caso de caracterizar todos los grupos del segundo nivel con el mismo y único parámetro θ. Así, se ignora a las unidades del segundo nivel, ya que al definir solamente un parámetro para todos los grupos, se ignora el hecho de que cada grupo pueda tener características diferentes al resto. A este tipo de modelo se le denomina en inglés *pooling model*.

Por el contrario, se puede modelizar también asignando un parámetro diferente por cada grupo, es decir, caracterizando cada grupo j con un parámetro θ_j independiente al resto, lo cual se conoce como *no-pooling model*. Está claro que este segundo caso va a presentar mayor flexibilidad que el primero, puesto que tiene en cuenta las diferentes características de cada grupo. No obstante, la desventaja de este tipo de modelos reside en que si alguno de los grupos no contiene información suficiente (la muestra no es considerable), las estimaciones dadas por la inferencia pueden ser extremadamente variables. Además, al haber considerado independencia total entre los parámetros θ_j, es imposible intercambiar información entre ellos.

Para tener en cuenta completamente la estructura jerárquica de los datos, los modelos jerárquicos (de ahí su nombre), buscan una solución a los modelos anteriores introduciendo un parámetro o parámetros extra de control llamados *hiperparámetros*, y que denotaremos como ω (como normal general este hiperparámetro será una variable aleatoria para la cual definiremos más adelante una distribución *a priori*). Suponiendo que tenemos el vector aleatorio de parámetros $\theta = (\theta_1, \ldots, \theta_J)^T$ para J grupos de datos, se obliga que la precisión (es decir, la inversa de la varianza) de θ dependa del hiperparámetro ω. Esta modificación nos permite poder intercambiar información entre parámetros del vector θ, ya que cada parámetro además de guardar información sobre su grupo, tendrá que compartir información con el resto, para ajustarse a la distribución conjunta controlada por el hiperparámetro ω. En los modelos espacio-temporales que se analizarán en este trabajo, se considerará una distribución normal multivariante de media 0 para el vector aleatorio de parámetros (también se verá el porqué), y en el caso más simple si consideramos que los parámetros son variables independientes idénticamente distribuidas, obtendríamos lo siguiente:

$$\theta = (\theta_1, \ldots, \theta_J)^T \sim \mathcal{N}(0, [\omega I_J]^-) \tag{2.7}$$

donde I_J es la matriz identidad de dimensión $J \times J$, y se ve que la distribución depende del parámetro ω, al cual también se denomina directamente parámetro de precisión.

Recapitulando todo lo mencionado, tendríamos: una primera etapa en la que se encuentra el vector de datos u observaciones $\vec{y} = (y_{11}, \ldots, y_{n_1 1}, \ldots, y_{n_J J})$ para el cual se especificará una distribución adecuada; una segunda etapa donde se encuentra el vector de parámetros θ que caracteriza a cada grupo de observaciones, y por último, una tercera etapa donde aparecen los hiperparámetros ω de cuyas distribuciones θ depende conjuntamente.

Es importante recalcar el haber introducido este enfoque de modelización jerárquica de datos, ya que ayudará a comprender los conceptos basados en la técnica INLA que vienen a continuación, puesto que los modelos espacio-temporales que se detallarán más tarde se construirán como modelos jerárquicos Bayesianos de tres etapas.

2.3. Integrated Nested Laplace Approximation (INLA)

Una vez ha sido introducido el concepto de inferencia Bayesiana, el siguiente paso es introducir la técnica INLA, la cual es utilizada para hacer dicha inferencia Bayesiana de un modo determinista, y del cual se esperan resultados precisos y con tiempo computacional razonable. INLA, como su nombre indica, se basa en aproximaciones de Laplace anidadas integradas, por lo que primero es importante conocer las nociones básicas de esta aproximación.

2.3.1. Aproximación de Laplace

Supongamos que nuestro interés reside en calcular la siguiente integral:

$$\int f(x)dx$$

donde $f(x)$ es simplemente la función de densidad de una variable aletaoria X. Se puede reescribir la integral de este modo:

$$\int f(x)dx = \int exp(logf(x))dx$$

Ahora se toma el logaritmo de la función de densidad, y se desarrolla como serie de Taylor evaluada en el punto $x = x_0$:

$$logf(x) \approx logf(x_0) + (x - x_0)\frac{\partial logf(x)}{\partial x}\bigg|_{x=x_0} + \frac{(x - x_0)^2}{2}\frac{\partial^2 logf(x)}{\partial x^2}\bigg|_{x=x_0}$$

Si se fija que, $x_0 = x^* = argmax_x logf(x)$, entonces el termino correspondiente a la primera derivada parcial se anulará, obteniendo:

$$logf(x) \approx logf(x^*) + \frac{(x - x^*)^2}{2}\frac{\partial^2 logf(x)}{\partial x^2}\bigg|_{x=x^*}$$

Sustituyendo esta expresión en la integral inicial,

$$\int f(x)dx \approx \int exp\left(logf(x^*) + \frac{(x - x^*)^2}{2}\frac{\partial^2 logf(x)}{\partial x^2}\bigg|_{x=x^*}\right)dx$$

Simplificando y fijando el valor $\sigma^2 = -1/\frac{\partial^2 logf(x)}{\partial x^2}\Big|_{x=x^*}$, se tiene que,

$$\int f(x)dx \approx f(x^*)\int exp\left(-\frac{(x - x^*)^2}{2\sigma^2}\right)dx$$

Y si lo pensamos, se ha conseguido llegar a una integral donde tenemos un nucleo de Gauss para una varianza σ^2 y una media x^*. Por tanto, evaluando la integral en el intervalo (a, b), la aproximación quedaría así:

$$\int_a^b f(x)dx \approx f(x^*)\sqrt{2\pi\sigma^2}(\Phi(b) - \Phi(a)) \tag{2.8}$$

siendo Φ la función de densidad cumulativa de la distribución $Normal(x^*, \sigma^2)$. Dicho de otro modo, la aproximación de Laplace de la variable aleatoria inicial es la siguiente: $X \approx Normal(x^*, \sigma^2)$.

2.3.2. Modelos Gaussianos latentes

El método INLA fue creado para trabajar con una clase específica de modelos llamados *modelos Gaussianos latentes*, por lo que es importante precisar la estructura de estos modelos.

El primer paso para definir los modelos Gaussianos latentes desde una perspectiva Bayesiana, y teniendo en mente la estructura de datos jerárquica explicada en la sección 2.2, es identificar una distribución para los datos observados a los que llamaremos $\vec{y} = (y_1, \ldots, y_n)$. Comúnmente, se caracteriza la distribución de cada observación y_i con un parámetro ϕ_i, que en su defecto será la media de la observación denotada por $E(y_i)$. Por ejemplo, en el ámbito del *disease mapping* que nos concierne, para modelizar los datos de incidencia (número de casos) de una enfermedad crónica como el cáncer, estos datos seguirán una distribución de Poisson caracterizada por su media (se verá en la ecuación 2.25).

A su vez, también se debe caracterizar esta media $E(y_i)$ en función de una serie de parámetros. Para ello, se define la misma en función de un predictor aditivo estructurado, utilizando una función de enlace llamada $g(\cdot)$:

$$g(E(y_i)) = \eta_i = \beta_0 + \sum_{k=1}^{K} \beta_k x_{ki} + \sum_{l=1}^{L} f_l(z_{li}); \quad \forall i = 1, \ldots, n \tag{2.9}$$

donde β_0 representa un intercepto global, $\beta = \{\beta_1, \ldots, \beta_K\}$ el efecto lineal de algunas covariables $x = (x_1, \ldots, x_K)$, y finalmente, $f = \{f_1(\cdot), \ldots, f_L(\cdot)\}$ son funciones de las covariables $z = (z_1, \ldots, z_L)$. Estas funciones desconocidas, pueden ser de varios tipos, como por ejemplo efectos espaciales o temporales, como veremos a posteriori al analizar el ejemplo práctico. Es por esto que estos modelos Gaussianos latentes pueden ajustar una gran cantidad de efectos, entre ellos los espacio-temporales, los cuales son el interés principal del trabajo.

Se dice que el predictor aditivo definido en 2.9 es estructurado, y los modelos llamados *Gaussianos latentes*, por una simple razón. Primeramente, porque se recogen todas las componentes no observables (latentes) en un conjunto θ de parámetros llamado *conjunto latente*, $\theta = \{\beta_0, \beta, f\}$. Después, porque al denotar con $\omega = \{\omega_1, \ldots, \omega_R\}$ al conjunto de R hiperparámetros, se asume una distribución *a priori* Normal (Gaussiana) Multivariante de media 0 y matriz de precisión $\mathcal{Q}(\omega)$ (que es la matriz inversa a la varianza) para θ, es decir, $\theta \sim \mathcal{N}(0, \mathcal{Q}(\omega))$. Por último, se verá en la sección 2.4.4 que en el caso de este estudio, cada parámetro dentro del conjunto latente estará modelizado mediante una distribución Normal Multivariante, y que la matriz de precisión no será más que el producto entre el hiperparámetro ω y una matriz de estructura que se especificará en cada caso.

2.3.3. Inferencia Bayesiana con INLA

En esta subsección finalmente se juntan todos los conceptos definidos anteriormente para analizar como utilizar la técnica INLA a la hora de hacer inferencia Bayesiana.

Volviendo a la ecuación 2.4, se debe definir primero la función de *verosimilitud* o incertidumbre de la variable observada $\vec{y} = (y_1, \ldots, y_n)$ condicionada al conjunto latente de parámetros e hiperparámetros. Para ello, se asume independencia

condicional entre las n observaciones, y se define la verosimilitud de la siguiente manera:

$$p(y|\theta,\omega) = \prod_{i=1}^{n} p(y_i|\theta_i,\omega) \tag{2.10}$$

Por otro lado, es necesario especificar también una distribución *a priori* para el conjunto de parámetros. Como ya se ha mencionado antes, INLA trabaja con modelos Gaussianos latentes, por lo que del mismo modo que en la sección anterior, se asume una distribución Normal Multivariante de media 0 y matriz de precisión $\mathcal{Q}(\omega)$, tal que se obtiene la siguiente función de densidad:

$$p(\theta|\omega) = (2\pi)^{-n/2}|\mathcal{Q}(\omega)|^{1/2}\exp\left(-\frac{1}{2}\theta'\mathcal{Q}(\omega)\theta\right) \tag{2.11}$$

Para más detalles, se ruega al lector consultar capítulo 4.7.1 del libro [4], puesto que no es la prioridad de este trabajo entrar en la más profunda teoría. Solamente cabe destacar que a esta especificación se le denomina GMRF (*Gaussian Markov Random Field*), y resulta computacionalmente beneficioso a la hora de hacer inferencia. Dicho esto, se hace inferencia aplicando el teorema 2.2 para lograr la distribución *a posteriori* de los parámetros e hiperparámetros de interés:

$$p(\theta,\omega|y) \propto p(\omega)p(\theta|\omega)p(y|\theta,\omega) \tag{2.12}$$

Como se puede observar, faltaría especificar una adecuada distribución *a priori* para los hiperparámetros de precisión ω, lo cual veremos como hacer al final de este capítulo, una vez introducida la teoriá de modelos espacio-temporales. Si nos fijamos, la modelización de 2.9 se puede plantear como una estructura jerárquica. En el primer nivel, se tiene la distribución de las observaciones condicionada a los parámetros e hiperparámetros $p(y|\theta,\omega) = \prod_{i=1}^{n} p(y_i|\theta_i,\omega)$. En un segundo nivel, tenemos el conjunto latente de parámetros el cual estaría caracterizado por una función de densidad dada por una distribución Normal Multivariante a la cual hemos llamado $p(\theta|\omega)$. Por último, en el tercer nivel se tendría una distribución *a priori* para el conjunto de hiperparámetros, $p(\omega)$. Por tanto, una estructura jerárquica de tres etapas.

Una vez definida la distribución $p(\omega)$, se obtendrían las distribuciones conjuntas de los parámetros e hiperparámetros, no obstante, el objetivo final de la inferencia Bayesiana es marginalizar estas distribuciones, es decir, obtener las distribuciones marginales *a posteriori* para cada uno de los parámetros e hiperparámetros:

$$p(\theta_i|y) = \int p(\theta_i,\omega|y)d\omega = \int p(\theta_i|\omega,y)p(\omega|y)d\omega \tag{2.13}$$

$$p(\omega_r|y) = \int p(\omega|y)d\omega_{-r} \tag{2.14}$$

Notar que ω_{-r} denota todos los hiperparámetros que no sean ω_r. Entonces, está claro que para conseguir las distribuciones marginales *a posteriori*, se necesita calcular los términos $p(\omega|y)$ y $p(\theta_i|\omega,y)$. Es en este punto donde utilizaremos las aproximaciones dadas por INLA. Empecemos por calcular $p(\omega|y)$:

$$p(\omega|y) = \frac{p(\theta,\omega|y)}{p(\theta|\omega,y)}$$

El numerador ya lo tenemos de la ecuación 2.12, y el denominador se aproxima mediante Laplace obteniendo la siguiente expresión:

$$p(\omega|y) = \frac{p(\theta,\omega|y)}{p(\theta|\omega,y)} \propto \frac{p(\omega)p(\theta|\omega)p(y|\theta,\omega)}{p(\theta|\omega,y)} \approx \frac{p(\omega)p(\theta|\omega)p(y|\theta,\omega)}{\tilde{p}(\theta|\omega,y)} =: \tilde{p}(\omega|y) \tag{2.15}$$

siendo $\tilde{p}(\theta|\omega,y)$ la aproximación dada por el método de Laplace. Por otro lado, queda calcular $p(\theta_i|\omega,y)$, lo cual puede ser muy costoso computacionalmente. Sin entrar en demasiados detalles (una vez más, si el lector desea consultarlo, puede hacerlo mediante el libro [4], capítulo 4.7.2), hay tres enfoques generales para aproximar dichas distribuciones marginales: aproximación plena de Laplace (*full Laplace approximation*), aproximación simplificada de Laplace (*simplified Laplace approximation*) y aproximación Gaussiana (*Gaussian aproximation*). Para el problema práctico de los capitulos siguientes, se ha decidido adoptar una aproximación de Laplace simplificada, ya que esta nos permitirá reducir costes computacionales obteniendo resultados razonablemente buenos.

2.4. Modelos espacio-temporales

Volviendo al ámbito epidemiológico y del *disease mapping*, se busca en esta sección introducir un tipo de modelo espacio-temporal de regresión ajustado el marco de trabajo Bayesiano. Este modelo, evidentemente, cumplirá con las condiciones y estructura de los ya mencionados *modelos Gaussianos latentes* (ver sección 2.3.2), y definirá una estructura para los efectos o parámetros espaciales, temporales e incluso espacio-temporales que caractericen a los datos de incidencia de cáncer reales observados.

2.4.1. Medidas clásicas de estimación de riesgo

Como ya se dijo en la introducción, es esencial desarrollar técnicas de modelización que permitan suavizar las medidas clásicas de estimación de riesgo, las cuales pueden resultar extremadamente variables en áreas o zonas poco pobladas, o incluso con enfermedades que presentan poca incidencia.

No obstante, estas técnicas de estimaciones de riesgo pueden utilizarse con asiduidad en análisis descriptivos previos a la modelización, y tienen como propósito medir el efecto producido por las enfermedades, que en el caso de este trabajo serán diferentes tipos de cáncer, entre las diferentes regiones del área de estudio. En concreto, el indicador de riesgo que se utilizará durante la aplicación práctica son las tasas crudas (en inglés, *crude rates*), aunque otras medidas como la razón de mortalidad estandarizada (en inglés, *standarized mortality ratio* o SMR) también sean muy utilizadas.

Supongamos que disponemos de los datos de incidencia y mortalidad para las n regiones dentro de nuestro área de estudio, donde $i \in \{1,\dots,n\}$, y que asimismo contamos con los datos sobre las regiones durante un periodo de tiempo de T instantes donde $t \in \{1,\dots,T\}$. Se define como y_{it} al número de casos observados en la región i en el instante t, y P_{it} a la población total en riesgo.

Tasas crudas

Se define como *tasa cruda* de la región i en el instante t a la tasa por cada 100.000 habitantes, es decir:

$$TC_{it} = \frac{y_{it}}{P_{it}} \times 100000 \tag{2.16}$$

Estas tasas simplemente indican la proporción de los casos observados de la enfermedad por cada 100.000 habitantes. Por lo tanto, pueden no ser un indicador muy eficaz ya que a la hora de analizar los factores que influyen en el riesgo de la enfermedad, como por ejemplo, la edad o el sexo, no se tienen en cuenta, puesto que se toma la población a estudiar sin ser desagregada por grupos.

SMR

Con la motivación de definir un indicador de riesgos más eficaz y que tenga en cuenta grupos de edad (por ejemplo, no tendrá la misma incidencia un cáncer de pulmón en niños que en ancianos), se definen los denominados métodos de estandarización, de los cuales se pueden destacar el método directo y el indirecto. Como el método de estandarización directo se centra en una población de referencia (por ejemplo, datos europeos), y por el contrario el método indirecto trabaja con datos observados en nuestro área de estudio, nos centraremos en definir solamente este segundo, puesto que es más acorde al problema práctico que se verá en el Capítulo 3. Concretamente, se define la *razón de mortalidad estandarizada*, que utiliza solamente los datos de mortalidad (y no los de incidencia), y para la cual necesitamos primero introducir el número de casos esperados:

$$e_{it} = \sum_{j=1}^{J} P_{itj} \frac{y_j}{P_j} \tag{2.17}$$

donde es preciso mencionar que se necesita hacer una partición de los datos por grupos de edad, es decir, P_{ijt} indica la población en riesgo de la región i en el instante t y grupo de edad j, para cada grupo de edad $j \in \{1, \ldots, J\}$. Además, $y_j = \sum_{i=1}^{n} \sum_{t=1}^{T} y_{itj}$ y $P_j = \sum_{i=1}^{n} \sum_{t=1}^{T} P_{itj}$ son el número de casos y población en riesgo para el grupo de edad j, respectivamente. Por último, la razón de mortalidad estandarizada o SMR para cada región i en el instante j se calcula dividiendo las observaciones de mortalidad entre el número de casos esperados:

$$SMR_{it} = \frac{y_{it}}{e_{it}} \tag{2.18}$$

Interpretar los ratios de mortalidad es algo más complicado que las tasas. Estos ratios describen si cada una de las regiones del estudio tiene más o menos riesgo de la enfermedad que el total del territorio bajo estudio en el periodo de tiempo completo. Dicho de otra manera, si los valores de los SMR son menores que la unidad, significa que se han observado menos muertes de las esperadas en la población total con la que hemos trabajado. En cambio si los valores superan la unidad, nuestra población habrá sufrido más muertes de las esperadas.

2.4.2. Modelización espacial para datos de área

A la hora de introducir efectos o parámetros espaciales al modelo general definido en 2.9, se debe mencionar que este trabajo solamente tratará con datos de área, que según el libro [4], se definen como datos u observaciones de tipo $y(s)$ donde s será una de las regiones del dominio o área D que será irregular y estará delimitado por un número contable de fronteras administrativas bien definidas. En nuestro caso práctico, el dominio D estará marcado por Inglaterra, mientras que cada región la delimitarán los diferentes CCGs (en inglés, Clinical Comissioning Groups), como se verá más adelante.

Esta claro que al tener regiones delimitadas por fronteras, va a ser posible definir una *matriz de estructura* $\mathbf{R}$ que tenga en cuenta la cercanía entre regiones en el mismo área. Para ello, es necesario primero introducir el concepto de vecindad entre dos regiones; se dice que las regiones i y j del dominio D están relacionadas, o en su defecto, que son vecinas, si comparten frontera, y denotaremos a la relación como $i \sim j$. Además, se denotará como $\mathcal{V}(i)$ al número de vecinos totales de la región i.

Por último, se define la matriz de estructura espacial $\mathbf{R}$ así:

$$R_{ij} = \begin{cases} \mathcal{V}(i) & si\ i = j \\ -1 & si\ i \sim j \\ 0 & otro\ caso \end{cases} \tag{2.19}$$

Es esencial tener en cuenta esta matriz de estructura espacial, puesto que de acuerdo con la teoría vista en la sección 2.3.2, el vector de parámetros espaciales de los modelos que se construirán (el cual denominaremos vector de efectos espaciales aleatorios), estará modelizado mediante una distribución *a priori* Normal Multivariante, para la cual se definirá una matriz de precisión. A su vez, esta matriz de precisión se construirá a partir de la matriz de estructura espacial $\mathbf{R}$, consiguiendo que la dependencia entre los efectos espaciales dependa de esta misma matriz.

Volviendo a la idea principal del suavizado del riesgo o las tasas de enfermedades crónicas, y teniendo en cuenta la estructura general de modelos Gaussianos latentes de la ecuación 2.9, denominamos como r_i a los ratios o tasas definidos para cada región de estudio i. En este tipo de estudios epidemiológicos, se considera habitualmente que los datos u observaciones que en este caso son conteos de incidencia o mortalidad observados para cada región, definidos en la sección 2.4.1 como y_i, siguen una distribución de Poisson condicionados a los riesgos o tasas r_i, es decir: $y_i|r_i \sim Poisson(\lambda_i)$. Dependiendo de si r_i indica riesgos o tasas, la media λ_i se podrá reescribir como $\lambda_i = P_i r_i$ (en caso de tasas, ya vimos que P_i era la población total de la región i) o $\lambda_i = e_i r_i$ (en caso de riesgos, vimos que e_i era el número de casos esperados de la región i).

Con esta idea en mente, se utiliza la función de enlace *log* para estimar tasas o riesgos como función de un predictor aditivo estructurado, definiendo el modelo espacial genérico del siguiente modo:

$$log(r_i) = \beta_0 + \xi_i \quad \forall i = 1, \ldots, n \tag{2.20}$$

donde β_0 indica el intercepto global y el vector de parámetros $\xi=(\xi_1,\ldots,\xi_n)$ muestra el efecto espacial aleatorio de cada región (se están suponiendo n regiones dentro del área de estudio). El siguiente paso consiste en definir una adecuada distribución *a priori* para el vector de efectos espaciales de cada región, lo cual se puede realizar mediante los diferentes tipos de distribuciones *CAR (conditional autoregressive)*. A continuación, se presentan dos de las distribuciones más utilizadas hoy en día, las cuales serán implementadas también en esta disertación.

iCAR

Por un lado, tenemos la distribución *a priori* conocida como *iCAR (intrinsec conditional autoregressive)*, la cual modeliza al vector de efectos espaciales aleatorios mediante una distribución Normal Multivariante definida tal que,

$$\xi \sim \mathcal{N}(0, [\omega_\xi R_s]^-)$$

donde ω_ξ es el hiperparámetro de precisión y R_s es la matriz mencionada anteriormente como matriz de estructura espacial. Por último, el símbolo $^-$ denota la inversa generalizada de Moore-Penrose.

BYM

Un modelo alternativo para el iCAR que permite tener en cuenta la variabilidad espacial no estructurada de los datos, es la modificación conocida como modelo *BYM*. En este caso, se incluyen dos efectos espaciales aleatorios (uno estructurado y el otro no). Uno de ellos, asumirá como antes una distribución *a priori* iCAR, mientras que el segundo, otra distribución Gaussiana para modelizar el efecto espacial no estructurado:

$$\xi = \xi_1 + \xi_2$$

donde

$$\xi_1 \sim \mathcal{N}(0, [\omega_{\xi_1} R_s]^-) \tag{2.21}$$

$$\xi_2 \sim \mathcal{N}(0, \omega_{\xi_2}^{-1} I_n) \tag{2.22}$$

Igual que antes, ω_{ξ_1} y ω_{ξ_2} son los hiperparámetros de precisión para los efectos espaciales, R_s es la misma matriz de estructura espacial, y por último, I_n es la matriz identidad de dimensión $n \times n$, al tener disponible un área con n regiones.

Cabe mencionar que además del modelo iCAR y BYM, también son utilizados frecuentemente otros modelos espaciales como el llamado LCAR (modelo de Leroux) o variaciones del modelo BYM.

2.4.3. Modelización temporal

Cuando se dispone de datos que además de estar desagregados espacialmente, están recogidos en diferentes periodos de tiempo, es de mayor interés introducir modelos con estructuras temporales que tengan en cuenta un posible comportamiento similar entre datos en instantes de tiempo cercanos.

Del mismo modo que se hizo para los datos de área, se puede construir una matriz de estructura temporal, definiendo antes el concepto de vecindad para los

diferentes instantes de tiempo. Supongamos que tenemos ordenados los intervalos de tiempo (en el caso de este trabajo, será un vector de años ordenado) sobre los que tenemos información en un vector llamado $T = (z_1, \ldots, z_T)$. Denominamos como orden k al factor que limitará el vecindario, es decir, dos intervalos de tiempo z_i y z_j serán vecinos o estarán relacionados (y, por tanto, escribiremos $z_i \sim z_j$), si están separados en el vector de intervalos de tiempo ordenado en un máximo de k posiciones. Asímismo, se denota como $\mathcal{V}_k(z_i)$ al número de vecinos del intervalo de tiempo z_i habiendo considerado un orden k. Por tanto, es natural construir la matriz de estructura temporal de la siguiente manera:

$$R_{kz_iz_j} = \begin{cases} \mathcal{V}_k(z_i) & si\ z_i = z_j \\ -1 & si\ z_i \sim z_j \\ 0 & otro\ caso \end{cases} \tag{2.23}$$

Generalmente, y durante este trabajo así será, se suele escoger $k = 1$ (aunque $k = 2$ es frecuente también), de tal manera que cada instante de tiempo estará relacionado con el instante anterior y posterior. A estos modelos se les denomina *Random Walk* (RW) de orden k, por lo que este último caso sería un Random Walk de primer orden.

Por último, llamando $\gamma = (\gamma_1, \ldots, \gamma_T)$ al vector de parámetros o efectos temporales aleatorios, se define de la misma manera que para los efectos espaciales una distribución *a priori* Normal Multivariante de media 0:

$$\gamma \sim \mathcal{N}(0, [\omega_\gamma R_t]^-)$$

donde ω_γ es el hiperparámetro de precisión y R_t la matriz de estructura temporal. Para el caso de RW de orden $k = 1$, la matriz R_t viene definida por:

$$R_t = R_{1z_iz_j} = \begin{cases} \mathcal{V}_1(z_i) & si\ z_i = z_j \\ -1 & si\ z_i \sim z_j \\ 0 & otro\ caso \end{cases} \tag{2.24}$$

2.4.4. Modelización espacio-temporal

Una vez han sido introducidos los modelos espaciales y temporales por separado, a continuación se describirá como se construyen los modelos espacio-temporales.

Primero de todo, es fundamental extender la teoría de la sección 2.4.2 en la cual se asume que, condicionado a los riesgos o tasas, cada observación de incidencia o mortalidad para cada región sigue una distribución de Poisson $(y_i|r_i \sim Poisson(\lambda_i))$. Si además disponemos de datos para T periodos de tiempo donde $t \in \{1, \ldots, T\}$, podemos asumir que,

$$y_{it}|r_{it} \sim Poisson(\lambda_{it}) \quad \forall i = 1, \ldots, n \ \ \forall t = 1, \ldots, T \tag{2.25}$$

donde λ_{it} sigue siendo la media de la variable aleatoria, la cual si trabajamos con tasas se reescribiría como $\lambda_{it} = P_{it}r_{it}$ (ahora P_{it} indicaría la población en la región i en el instante t), y $\lambda_{it} = e_{it}r_{it}$ (ahora e_{it} indicaría el número de casos esperado en la región i en el instante t) en caso de riesgos.

Nuestros indicadores de interés r_{it} se modelizan de nuevo usando la función *log* como link, obteniendo la siguiente expresión:

$$log(r_{it}) = \beta_0 + \xi_i + \gamma_t + \rho_t + \delta_{it} \tag{2.26}$$

Comparando la expresión 2.26 con la ecuación 2.20, se puede observar la aparición de los terminos relacionados a la componente temporal. β_0 muestra al igual que antes, el intercepto global, mientras que ξ_i es la componente espacial, para cuyo vector de efectos aleatorios se vieron diferentes tipos de distribuciones *a priori*. Después, para la componente temporal estructurada γ_t se asume una distribución RW de primer orden ($\gamma \sim \mathcal{N}(0, [\omega_\gamma R_t]^-)$) y también se puede añadir a este modelo una componente temporal no estructurada, a la cual llamaremos ρ_t, y que modelizaremos con la siguiente distribución *a priori*:

$$\rho_t \sim \mathcal{N}(0, \omega_\rho^{-1} I_T) \tag{2.27}$$

siendo ω_ρ el hiperparámetro de precisión y I_T la matriz identidad $T \times T$. Por último, queda mencionar el término δ_{it} que corresponde al efecto aleatorio dado por la interacción espacio-temporal. Esta interacción puede darse de cuatro formas diferentes (ver artículo [11]), dependiendo si las matrices de precisión de los efectos aleatorios incluyen o no efectos espaciales/temporales estructurados. La tabla que se muestra a continuación resume las cuatro posibles situaciones:

Interacción	Parámetros interaccionando	Matriz resultante R_δ
I	No estructurados	$I_n \otimes I_T$
II	Estructurado temporal	$I_n \otimes R_t$
III	Estructurado espacial	$R_s \otimes I_T$
IV	Estructurado espacial y temporal	$R_s \otimes R_t$

Tabla 2.1: Diferentes tipos de interacción espacio-temporales dependiendo del tipo de correlación espacial/temporal que introducen en los datos, y sus respectivas matrices de estructura.

Utilizando la matriz de estructura resultante R_δ, la distribución *a priori* para el vector de efectos espacio-temporales de interacción aleatorio se especifica de la siguiente manera:

$$\delta \sim \mathcal{N}(0, [\omega_\delta R_\delta]^-) \tag{2.28}$$

Una vez más, ω_δ es el hiperparámetro de precisión para la interacción, y R_δ es la matriz de dimensiones $nT \times nT$ obtenida interaccionando los diferentes casos de los parámetros (ver Tabla 2.1).

Cabe mencionar que la matriz resultante en cada caso es calculada mediante el producto de *Kronecker* entre la matriz espacial y la matriz temporal. Supongamos que tenemos dos matrices $A = (a_{ij})$ y $B = (b_{ij})$ para las cuales sus dimensiones son $m \times n$ y $p \times q$, respectivamente. Entonces el producto de Kronecker denotado $A \otimes B$ es una matriz de dimensiones $mp \times nq$ definida así:

$$\begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix} \tag{2.29}$$

2.4.5. Distribuciones *a priori* para los parámetros de precisión

Tal y como se ha descrito en la introducción de los modelos jerárquicos en la sección 2.2, para poder llevar a cabo la inferencia Bayesiana y tener completa la estructura jerárquica de tres etapas, además de una distribución para los datos observados condicionados a los riesgos (hemos visto que para los datos de incidencias Poisson va bien), y distribuciones *a priori* para el vector de parámetros aleatorios (también se han visto distribuciones para las componentes espacial, temporal y espacio-temporal), es indispensable especificar unas distribuciones *a priori* para todos los hiperparámetros de precisión presentes en las matrices de precisión de las distribuciones de los efectos aleatorios. Volviendo a la notación de la sección 2.2, se ha obtenido el siguiente conjunto latente de vectores de parámetros: $\theta = \{\beta_0, \xi, \gamma, \rho, \delta\}$, a la vez que el conjunto de hiperparámetros $\omega = \{\omega_{\xi_1}, \omega_{\xi_2}, \omega_\gamma, \omega_\rho, \omega_\delta\}$. Para estos hiperparámetros de precisión, se han considerado las siguientes distribuciones uniformes *a priori*:

$$\sigma = \frac{1}{\sqrt{\omega}} \sim \mathcal{U}(0, \infty) \tag{2.30}$$

2.4.6. Criterios de selección de modelos

La descripción teórica de esta sección va a ser muy relevante cuando se planteen diferentes modelos espacio-temporales para el caso práctico en el siguiente capítulo. Es importante tener alguna medida para saber cual o cuales de los modelos construídos se ajustan mejor a los datos observados, y a la hora de comparar los modelos respecto a su desempeño, los criterios de información basados en la *deviance* son muy utilizados.

Supongamos que tenemos los datos $\vec{y} = (y_1, \ldots, y_n)$ y su función de verosimilitud $p(\vec{y}|\theta)$ para los parámetros θ. Entonces la *deviance* del modelo se define tal que

$$D(\theta) = -2log(p(\vec{y}|\theta))$$

Dicha *deviance* mide la variabilidad relacionada a la verosimilitud, la cual es la función utilizada para las observaciones, dados los parámetros.

Deviance Information Criterion (DIC)

El criterio DIC o *Deviance Information Criterion* (ver artículo [12]), es el criterio basado en la *deviance* más utilizado para valorar el ajuste de un modelo creado a partir de inferencia Bayesiana. En esencia, se define como suma de dos factores:

$$DIC = \bar{D} + p_D \tag{2.31}$$

donde $\bar{D} = E_{\theta|\vec{y}}(D(\theta))$ es la media de la *deviance a posteriori*, y $p_D = \bar{D} - D(\bar{\theta})$.

En este caso, $\bar{D}$ es el factor que indica el ajuste del modelo. Naturalmente, una buena medida de selección deberá considerar el *tradeoff* entre la calidad del modelo y su coste computacional, y es por eso que se introduce al criterio el factor p_D (número de parámetros efectivos), el cual mide la complejidad del modelo. En resumen, cuanto más pequeño sea el valor dado por DIC, mejor *tradeoff* entre ajuste del modelo y complejidad se habrá obtenido.

Watanabe-Akaike Information Criterion (WAIC)

Otro criterio de información utilizado recientemente es el criterio de Watanabe-Akaike o WAIC (ver artículo [13]). Este criterio, a diferencia del DIC, es invariante a la parametrización, y se menciona en esta sección puesto que cuando se ajusten los modelos con `R`, el mismo programa nos dará la opción de calcularlo. Una vez más, cuanto menor sea el valor proporcionado por el WAIC, mejores serán los resultados obtenidos.

Por último, cabe mencionar que a parte de los criterios basados en la *deviance*, también existen criterios que tienen en cuenta la distribución predictiva, como podría ser el denominado *logarithmic score* (ver artículo [14]). No obstante, en este trabajo no se tendrán en cuenta y se valorará solamente haciendo uso del DIC y WAIC.

2.5. Predicción a corto plazo mediante modelos espacio-temporales Bayesianos con `R-INLA`

Por último, en relación al estudio de validación y predicción de modelos a un futuro cercano que se llevará a cabo en el Capítulo 4, se detalla el proceso utilizado por `R-INLA` para realizar y evaluar dichas predicciones (para más detalles, se muestran los artículos [15] o [16]).

2.5.1. Predicción con datos faltantes: `NAs`

Al ejecutar un modelo espacio-temporal planteado en `R` mediante la llamada de la función general `inla()`, es posible obtener estimaciones marginales *a posteriori* para nuestros indicadores de interés r_{it} (recordemos que en nuestro caso trabajaríamos con tasas crudas), con los cuales trabajaremos durante el Capítulo 3. No obstante, la función `inla()` también permite trabajar con datos faltantes en la variable de observaciones a la cual llamaremos ahora y_{it^*}, de modo que si disponemos de los datos correspondientes a la población en riesgo P_{it^*} para un punto de tiempo en un futuro cercano t^*, podemos obtener estimaciones predichas de las tasas r_{it^*}. Es decir, realizar predicciones en `R-INLA` es computacionalmente muy sencillo ya que solamente se requiere asignar como "`NA`" a los valores faltantes de la variable de observaciones, para después ejecutar una vez más el modelo mediante `inla()` y obtener predicciones de la variable de interés.

2.5.2. Distribución predictiva *a posteriori* de las observaciones

Durante la aplicación práctica del Capítulo 4, se predecirán tasas un año, dos y tres a futuro, por lo que una de las grandes incógnitas residirá en formular diferentes medidas que permitan seleccionar los mejores modelos respecto a su capacidad predictiva.

En estos estudios de predicciones a corto plazo, es común utilizar medidas de validación que se basan en el cálculo de la llamada distribución predictiva *a*

posteriori. Puesto que `R-INLA` solamente proporciona las distribuciones marginales de cada una de las predicciones o predictores lineales estimados por el modelo, no se obtiene información acerca de la distribución predictiva de las observaciones. Es decir, para cada predicción estimada r_{it^*}, `R-INLA` solamente nos facilita su distribución marginal $p(r_{it^*}|\vec{y})$ siendo $\vec{y}$ el vector de observaciones. Ahora bien, el objetivo final es conseguir la función de densidad o distribución predictiva de cada observación y_{it^*}, dada por la siguiente expresión:

$$p(y_{it^*}|y_{-it^*}) = \int p(y_{it^*}|r_{it^*})p(r_{it^*}|y_{-it^*})dr_{it^*} \tag{2.32}$$

donde y_{-it^*} equivale a todo el vector de observaciones excepto la correspondiente a la estimada. Como se puede apreciar en la ecuación 2.32, se integra sobre el parámetro recién predicho r_{it^*}, y resolverlo explicitamente es imposible. Es por esto que se adoptan diferentes estrategias de resolución del problema, como la integración numérica. No obstante, en este trabajo se ha optado por utilizar la técnica del muestreo.

Este muestreo trata de dos pasos principales. Primero, se generan muestras a las cuales llamaremos $r_{it^*}^m$ (en nuestro caso se emplearán $m = 1, \ldots, 3000$) de la distribución marginal $p(r_{it^*}|\vec{y})$, utilizando la función `inla.rmarginal()`. Segundo, se generan valores $y_{it^*}^m$ para las observaciones mediante una distribución de Poisson con parámetro o media $P_{it^*}r_{it^*}^m$. Todos estos valores (en nuestro caso son 3000) conformarán conjuntamente la distribución predictiva que estamos buscando. Además, cabe mencionar que de esta distribución es fácil obtener los cuantiles *a posteriori* o el valor esperado de las observaciones predictivas, los cuales necesitaremos también a la hora de calcular las medidas de validación en el Capítulo 4.

Capítulo 3

Análisis espacio-temporal de la incidencia y mortalidad por cáncer

En el capítulo anterior, se han propuesto varios modelos espacio-temporales jerárquicos utilizados en el ámbito del *disease mapping* y dentro del marco Bayesiano. El objetivo principal de estos residía en estimar las medidas clásicas de riesgo como las tasas crudas o razón de mortalidad estandarizada (SMR), permitiendo su suavizado, puesto que estas podían resultar de gran variabilidad en regiones con enfermedades de poca incidencia, o poco pobladas.

Para ilustrar la aplicabilidad de este tipo de modelos con conjuntos de datos reales, se utilizan datos de incidencia y mortalidad de diferentes tipos de cáncer en la isla de Gran Bretaña, particularmente en el territorio de Inglaterra, durante el periodo de años 2001-2017. Concretamente, se utilizan tanto los datos de incidencia como los de mortalidad extraídos para llevar a cabo un análisis descriptivo de tasas previo a la modelización, mientras que el ajuste de modelos espacio-temporales se realizará unicamente con los datos de incidencia.

3.1. Extracción de datos

Se han creado dos bases de datos principales partiendo de los datos proporcionados en abierto por *Office for National Statistics* (ONS, [10]) y el Registro de Cáncer de Inglaterra (NHS, [9]), una con datos de incidencia y otra con datos de mortalidad. Para ello, se han extraído primeramente de [9] datos que recogen el número de casos observados y el número de muertos correspondiente a cada región del territorio inglés, desagregados por año, sexo, grupo de edad, índice del tipo de cáncer (según la clasificación internacional de enfermedades, CIE) y nombre del propio cáncer. Después, utilizando la fuente [10], se ha agregado a cada una de las bases de datos construída la población total de cada región y desagregada otra vez por año, sexo, grupo de edad y tipo de cáncer.

Es conveniente mencionar, que la partición de regiones del territorio inglés se ha realizado de acuerdo con el sistema sanitario local, el cual divide el área total en los llamados *Clinical Commissioning Groups* (CCGs) (ver Figura 3.1).

Figura 3.1: División de Inglaterra según sus 105 Clinical Commissioning Groups.

Esta división geográfica ha sufrido modificaciones a lo largo del periodo analizado, por lo que se ha llevado a cabo un proceso de homogeneización de las áreas que han sido objeto de estudio. Así, el análisis descriptivo que viene en la siguiente sección, se ha realizado escogiendo los siguientes casos: cáncer de mama (C50) en mujeres, cáncer de estómago (C16) en hombres, cáncer de pulmón (C33-C34) en mujeres y cáncer de pulmón (C33-C34) en hombres.

A continuación, se muestra una captura de pantalla del inicio de la base de datos de incidencia construida en R (para mortalidad es análoga), habiéndola reducido al caso de cáncer de estómago para hombres.

	Age.group	Year	Region	Gender	ICD10_code	Cancer_site	Count	Pop
1	Under 25	2001	E38000006	Male	C16	Malignant neoplasm of stomach	0	33346
2	Under 25	2001	E38000007	Male	C16	Malignant neoplasm of stomach	0	36643
3	Under 25	2001	E38000008	Male	C16	Malignant neoplasm of stomach	0	16336
4	Under 25	2001	E38000014	Male	C16	Malignant neoplasm of stomach	0	25926
5	Under 25	2001	E38000015	Male	C16	Malignant neoplasm of stomach	0	20254
6	Under 25	2001	E38000016	Male	C16	Malignant neoplasm of stomach	0	43000
7	Under 25	2001	E38000021	Male	C16	Malignant neoplasm of stomach	0	35923
8	Under 25	2001	E38000024	Male	C16	Malignant neoplasm of stomach	0	28979
9	Under 25	2001	E38000025	Male	C16	Malignant neoplasm of stomach	0	30066
10	Under 25	2001	E38000026	Male	C16	Malignant neoplasm of stomach	0	120023

Figura 3.2: Base de incidencia para cáncer de estómago en hombres.

3.2. Análisis descriptivo

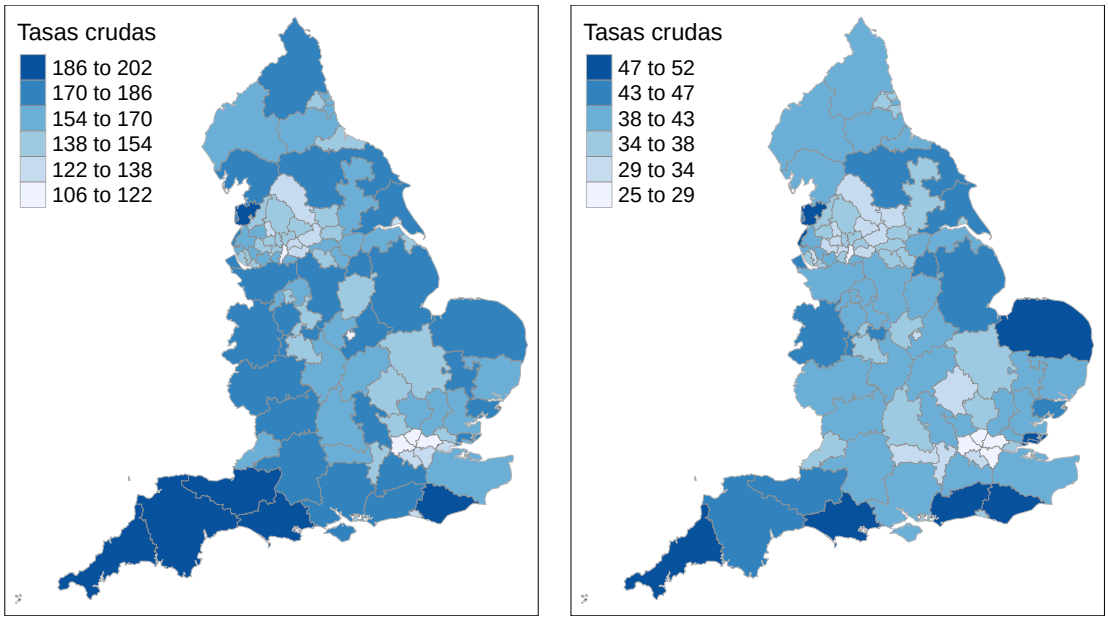
Antes de iniciar el proceso de modelización, se ha llevado a cabo un análisis descriptivo mediante el cual se han ilustrado las tasas crudas por cien mil habitantes correspondientes a los cuatro casos de cáncer anteriormente mencionados. Como ya se dijo en la introducción teórica, es mucho más preciso

representar cada enfermedad con medidas clásicas de riesgo que con el número de casos de incidencia o mortalidad, ya que estas toman en cuenta la población total de cada una de las regiones.

Esta representación geográfica, se ha realizado tanto para los datos de incidencia como los de mortalidad extraídos anteriormente. Además, con el objetivo de poder visualizar el impacto de la componente espacial y temporal de los datos, se ha procedido a dividir este análisis en tres secciones. Por un lado, se dibujan mapas con la distribución geográfica de las tasas para todo el periodo (años 2001-2017). Por otro lado, se grafica una sencilla linea indicativa de la evolución temporal de las tasas para toda la geografía inglesa en conjunto y, por último, se tienen en cuenta tanto la componente espacial como la temporal para dibujar las cartografías correspondientes a dichas tasas en el tiempo.

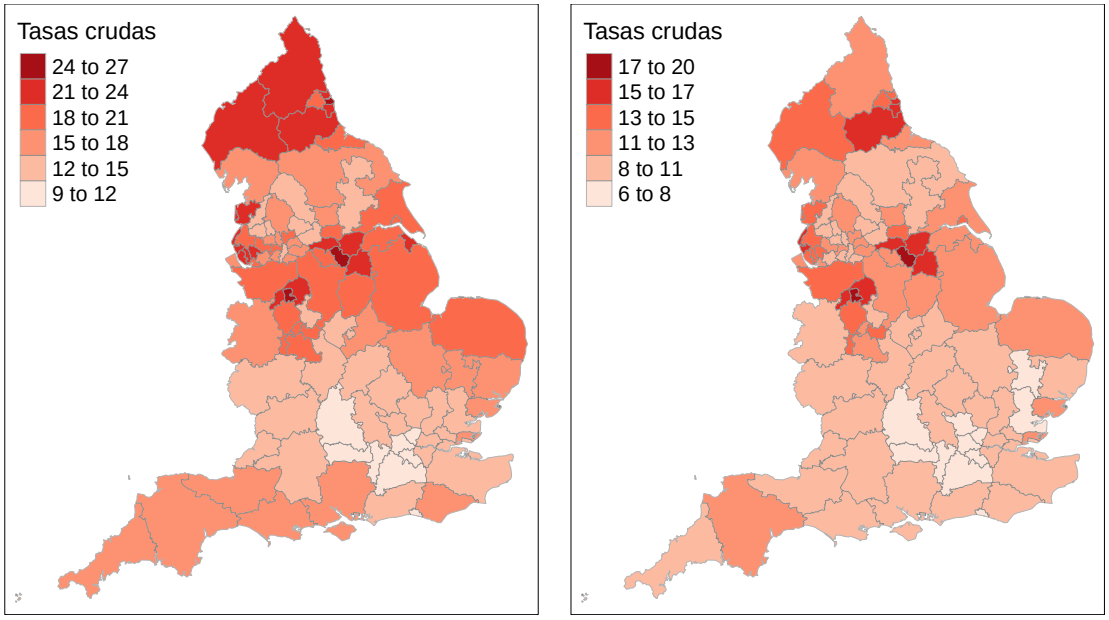
3.2.1. Patrón espacial

En esta sección, se muestra la distribución espacial de las tasas crudas (para incidencia y mortalidad) para el periodo 2001-2017, procedentes del cáncer de mama en mujeres, cáncer de estómago en hombres y cáncer de pulmón tanto en hombres como mujeres:



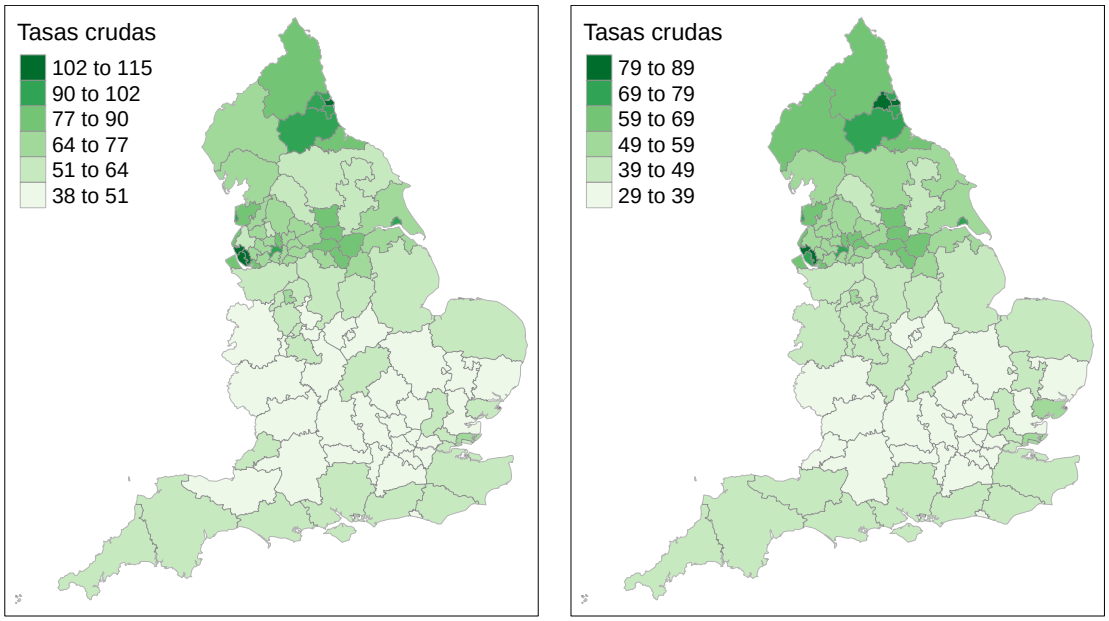
(a) Tasas incidencia (b) Tasas mortalidad

Figura 3.3: Tasas crudas (por 100.000 habitantes) cáncer de mama en mujeres.



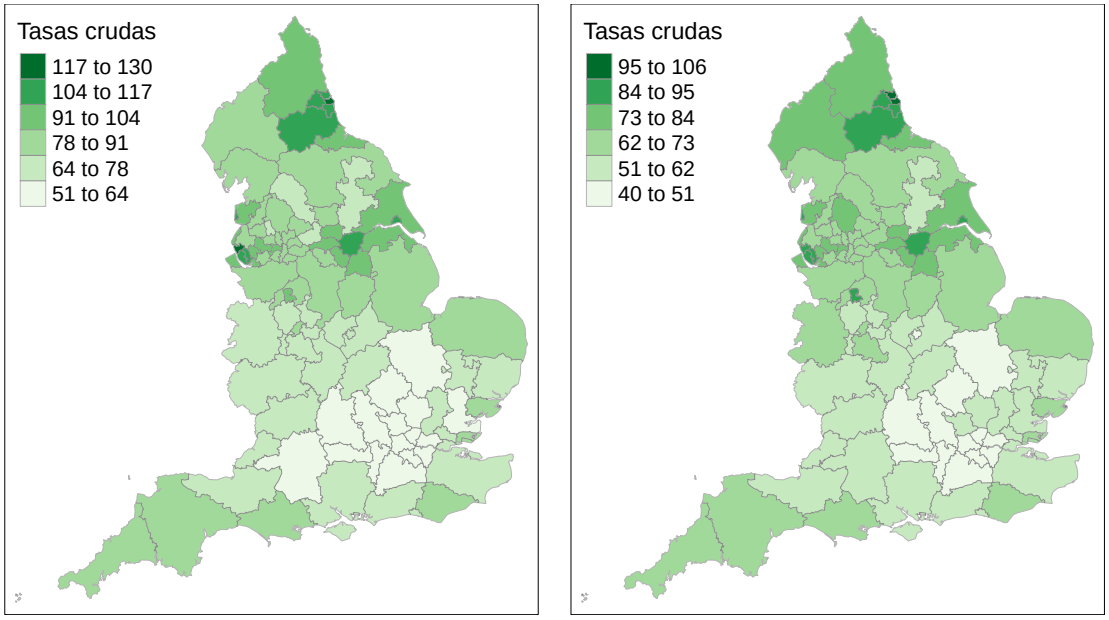
(a) Tasas incidencia (b) Tasas mortalidad

Figura 3.4: Tasas crudas (por 100.000 habitantes) cáncer de estómago en hombres.



(a) Tasas incidencia (b) Tasas mortalidad

Figura 3.5: Tasas crudas (por 100.000 habitantes) cáncer de pulmón en mujeres.



(a) Tasas incidencia

(b) Tasas mortalidad

Figura 3.6: Tasas crudas (por 100.000 habitantes) cáncer de pulmón en hombres.

Observando los mapas anteriores, y teniendo en cuenta que regiones más oscuras en las cartografías indican una tasa mayor de afectados, se puede deducir que es probable que la geografía del área total influya en el desarrollo de la enfermedad. Por ejemplo, considerando la Figura 3.3 correspondiente al cáncer de mama, se observa que las regiones del suroeste de Inglaterra presentan, en general, una mayor tasa de incidencia, mientras que observando las Figuras 3.4, 3.5 o 3.6, se ve que los territorios del norte presentan tasas mayores para el cáncer de estómago o cáncer de pulmón.

3.2.2. Patrón temporal

Se realiza un análisis similar al anterior para analizar la evolución temporal global (es decir, todas las regiones de forma conjunta) de las tasas crudas de incidencia y mortalidad. Para ello, se dibuja para cada una las casuísticas analizadas anteriormente, una línea azul que indica la tasa cruda por cien mil habitantes para cada año del periodo de estudio. Una vez más, para cada cáncer se dibujan dos gráficos, uno para los datos de incidencia, y otro para mortalidad.

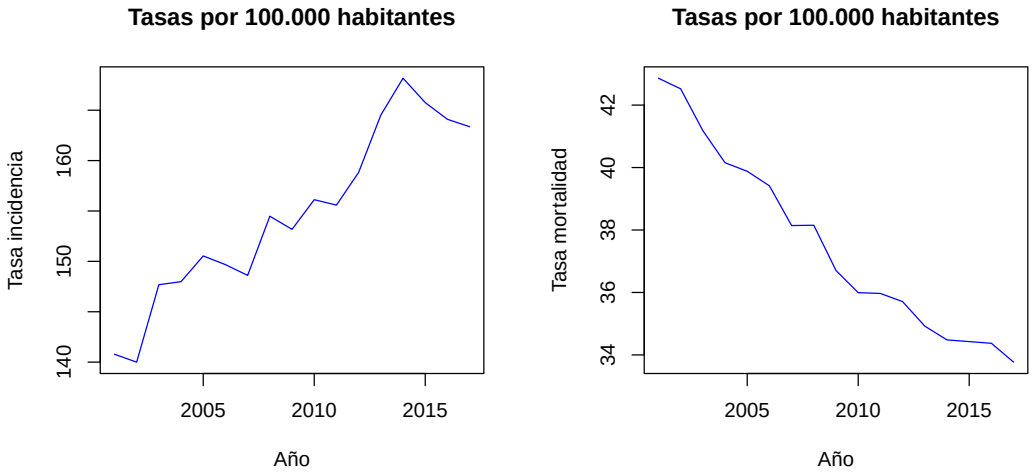


Figura 3.7: Evolución temporal cáncer de mama en mujeres (izquierda incidencia, derecha mortalidad).

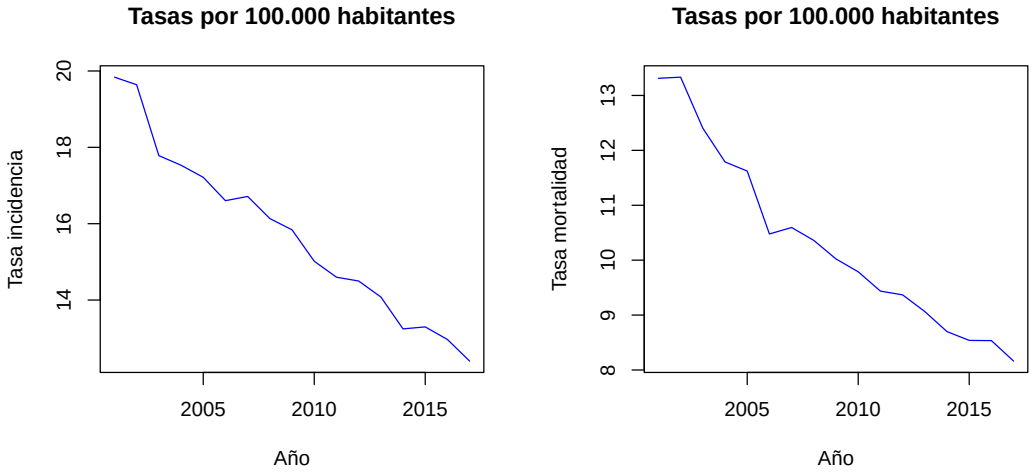


Figura 3.8: Evolución temporal cáncer de estómago en hombres (izquierda incidencia, derecha mortalidad).

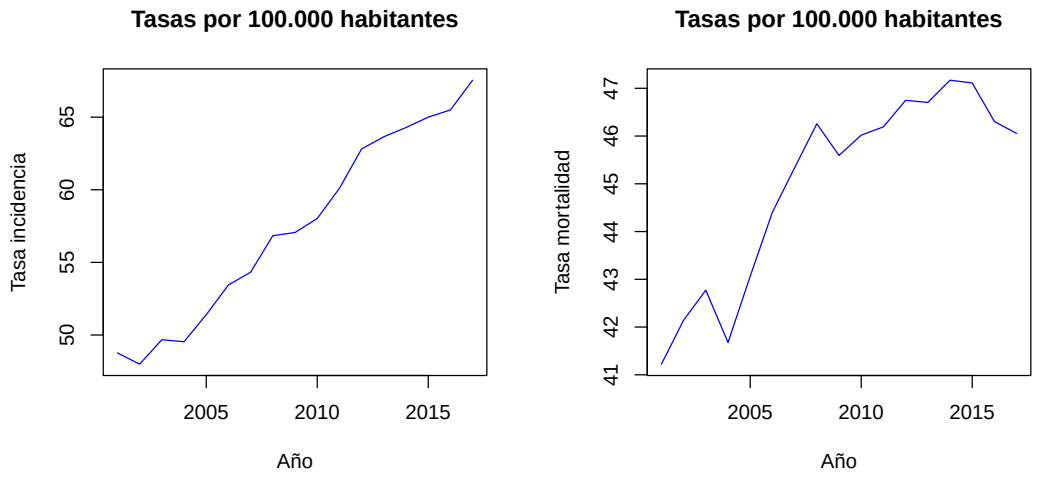


Figura 3.9: Evolución temporal cáncer de pulmón en mujeres (izquierda incidencia, derecha mortalidad).

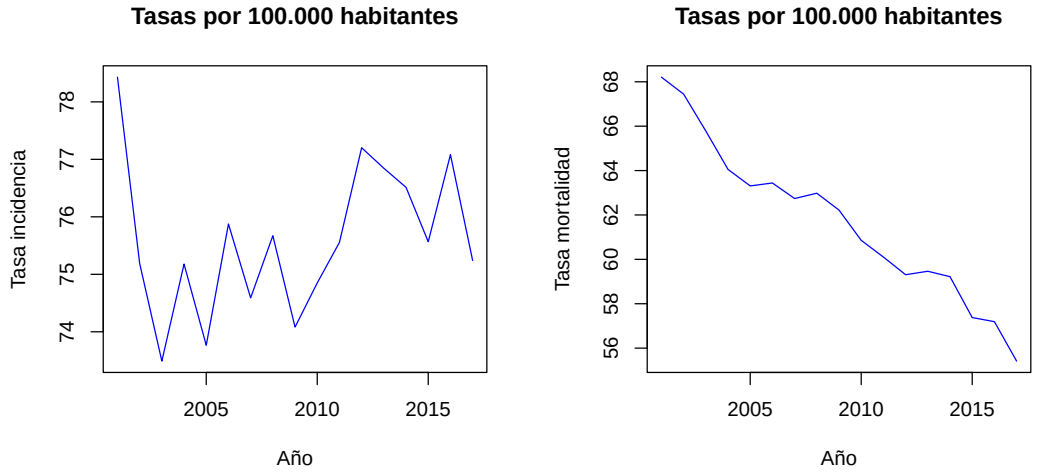


Figura 3.10: Evolución temporal cáncer de pulmón en hombres (izquierda incidencia, derecha mortalidad).

En este caso, la mayoría de casos analizados muestra un crecimiento o decrecimiento de la línea temporal de tasas, lo cual indicaría que efectivamente el patrón temporal es una variable fundamental para explicar la evolución de la incidencia o mortalidad de este tipo de enfermedades crónicas. En casos como el cáncer de estómago o cáncer de pulmón para mujeres, la tendencia de la incidencia y mortalidad es la misma, es decir, creciente o decreciente. Llama la atención, por ejemplo, la tendencia creciente de la incidencia en el cáncer de mama, puesto que la tasa de mortalidad decrece de un modo considerable.

3.2.3. Patrón espacio-temporal

Los análisis efectuados en las secciones anteriores se han realizado para tipos de cáncer y sexo concretos. Además, para todos ellos, se ha calculado como medida de estimación de riesgos la tasa cruda. No obstante, los resultados obtenidos se podrían reproducir para cualquier tipo de cáncer y sexo en general, ya que las bases de datos construídas inicialmente lo permiten. Del mismo modo, se pueden considerar los SMRs como medidas de riesgo, en caso de que se quisiera corregir los sesgos por edad (por ejemplo, no tendrá la misma incidencia un cáncer de mama para una joven que para una mujer mayor).

En adelante, este trabajo se centrará solamente en las tasas crudas de incidencia por cáncer de estómago para el sexo masculino, ya que el objetivo principal es ilustrar la aplicabilidad de los modelos que se propondrán más tarde, y se ha considerado que este caso es un buen ejemplo. El análisis previo de tasas para todo el periodo, y su evolución temporal, corroboran la importancia del patrón espacial y temporal en el cálculo de estas tasas y en general en el ámbito del *disease mapping*, por lo que tan solo faltaría ilustrar la evolución espacio-temporal de las tasas crudas. Para ello, como se acaba de mencionar, se utilizarán los datos de incidencia correspondientes al cáncer de estómago (su identificación según la CIE es C16):

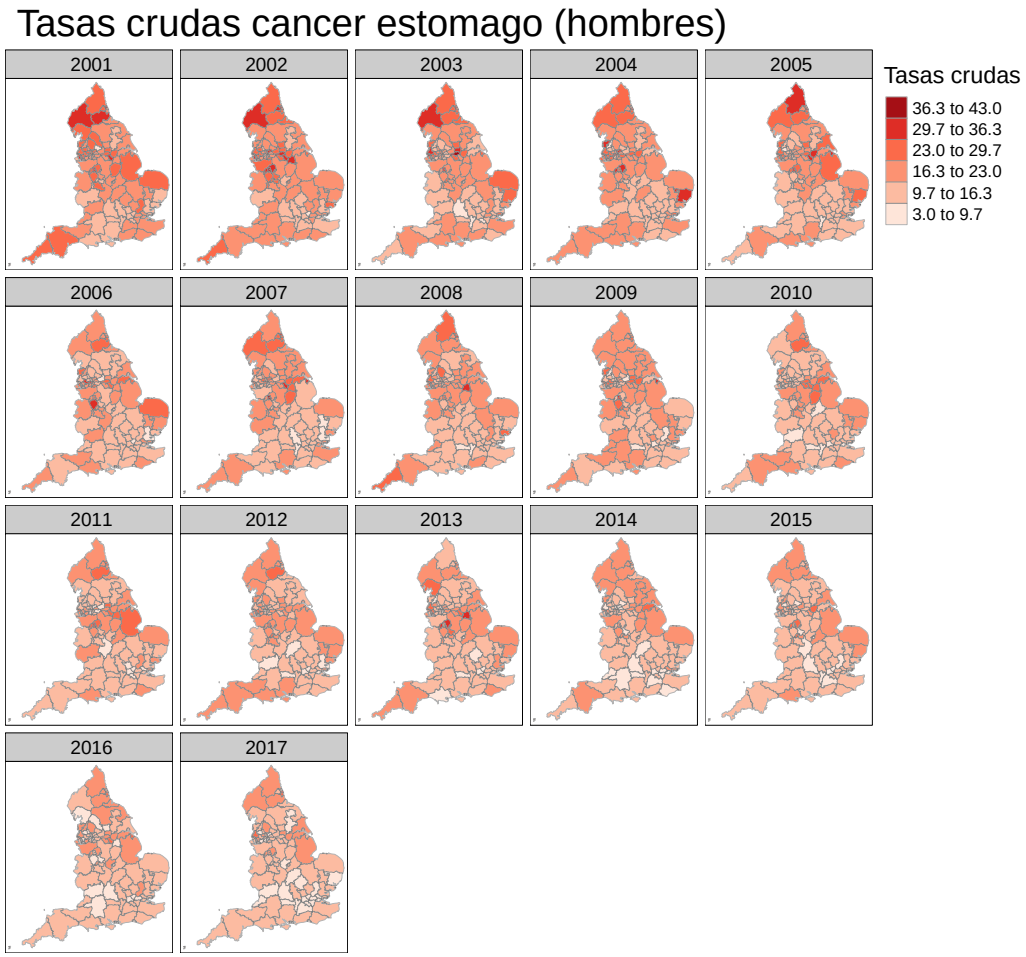


Figura 3.11: Evolución espacio-temporal de tasas de incidencia por cáncer de estómago en hombres.

Si nos fijamos en las Figuras 3.4 y 3.8 se observa claramente la influencia temporal y la espacial en los mapas de la Figura 3.11, donde las tasas se reducen considerablemente con el paso de los años, y zonas cercanas mantienen generalmente parecidos niveles de estas tasas de incidencia.

3.3. Modelización espacio-temporal

En esta sección se presenta el estudio correspondiente a la modelización espacio-temporal de las tasas crudas por incidencia para el cáncer de estómago en hombres del territorio inglés, durante el periodo de tiempo 2001-2017. Hay que recordar, que al estar trabajando en un enfoque Bayesiano, los modelos construidos y ejecutados mediante `R-INLA`, nos permitirán obtener aproximaciones de las distribuciones marginales posteriores de cada una de las tasas, además de las distribuciones marginales para los efectos espaciales, temporales, espacio-temporales e incluso hiperparámetros de precisión.

Por último, también se debe mencionar que los modelos que se ajustarán a continuación, se encuentran dentro de los modelos GLMM (*generalized linear*

mixed models), y estos requieren especificar una serie de restricciones de suma a cero para obtener resultados viables, ya que los efectos aleatorios de los modelos generalmente suelen presentar problemas de identificación. Por esta razón, se han implementado las restricciones necesarias implícitamente en cada uno de los modelos ajustados (ver, por ejemplo, [17], para más información sobre la definición de estas restricciones).

En total, se ilustrarán 8 modelos espacio-temporales diferentes partiendo del modelo genérico introducido en la sección teórica 2.4.4, dado por la ecuación 2.26, por lo que se comienza por ver como construir cada uno de ellos. Primero, se asume que las observaciones y_{it}, es decir, los casos de incidencia de cada región i (para $i \in \{1, \ldots, 105\}$) en el año t (para $t \in \{1, \ldots, 17\}$), siguen una distribución de Poisson condicionados a las tasas las cuales denotamos como r_{it}. Por tanto, se escribe de manera equivalente a la ecuación 2.25 la distribución de los datos:

$$y_{it}|r_{it} \sim Poisson(\lambda_{it}) \quad \forall i = 1, \ldots, 105 \;\; \forall t = 1, \ldots, 17 \tag{3.1}$$

En este caso, como se está condicionando los datos respecto a las tasas, ya vimos que la media de la distribución λ_{it} se podría reescribir como el producto entre la tasa y la población: $\lambda_{it} = r_{it}P_{it}$, donde P_{it} es la población de la región i en el instante de tiempo t. Ahora, esta misma tasa que es una proporción de la media y la cual a su vez caracteriza la distribución de los datos, se enlaza mediante una función link logaritmo a un predictor lineal η_{it} (recordar ecuación 2.9), obteniendo la siguiente expresión o modelo genérico:

$$log(r_{it}) = \eta_{it} = \beta_0 + \xi_i + \gamma_t + \rho_t + \delta_{it} \tag{3.2}$$

Recordemos que β_0 representa un intercepto global que muestra un valor general de la tasa.

Por un lado, se vió en la sección teórica que el vector de parámetros o efectos aleatorios espaciales $\xi = (\xi_1, \ldots, \xi_{105})$ se podía modelizar de formas diferentes. Consideraremos por un lado, el modelo iCAR, para el cual el vector de parámetros se modeliza mediante una distribución Normal Multivariante de media 0, tal que, $\xi^1 = (\xi^1_1, \ldots, \xi^1_{105}) \sim \mathcal{N}(0, [\omega_{\xi^1} R_s]^-)$. Por otro lado, consideraremos el modelo BYM, el cual modelizaba al vector de parámetros de la siguiente manera: $\xi^2 = (\xi^2_1, \ldots, \xi^2_{105}) = \xi_1 + \xi_2$, tal que, $\xi_1 \sim \mathcal{N}(0, [\omega_{\xi_1} R_s]^-)$ y $\xi_2 \sim \mathcal{N}(0, \omega^{-1}_{\xi_2} I_{105})$. Para ambos modelos espaciales, a parte de utilizar la misma distribución *a priori* uniforme no-informativa para los parámetros de precisión ω (consultar la teoría sobre la distribución asignada a los hiperparámetros 2.4.5), era necesario construir la matriz de estructura espacial denominada R_s, lo cual se ha hecho utilizando la ecuación 2.19. A continuación, se muestra el grafo de vecindad utilizado para construir dicha matriz de estructura espacial necesaria en la modelización:

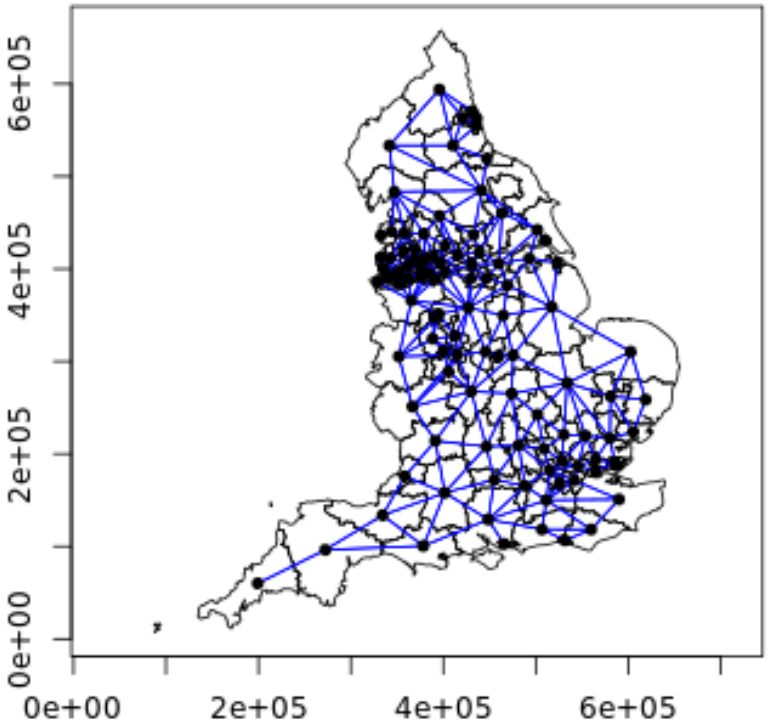


Figura 3.12: Grafo de vecindad del territorio inglés, donde las regiones vecinas se unen con una línea azul.

Por otro lado, tenemos los parámetros correspondientes al patrón temporal. Por un lado, el vector de efectos temporales o componente estructurada $\gamma = (\gamma_1, \ldots, \gamma_{17})$ se modelizará mediante un *Random Walk* de primer orden, de modo que $\gamma \sim \mathcal{N}(0, [\omega_\gamma R_t]^-)$. Una vez más, la distribución *a priori* para el hiperparámetro ω será la misma, y la matriz de estructura temporal R_t se construirá de acuerdo con la ecuación 2.24. Para ello, se ordenarán en un vector todos los años disponibles en el estudio (desde 2001 hasta 2017), y se considerarán vecinos solamente los años contiguos, puesto que se ha escogido un modelo de primer orden. Por otra parte, aunque inicialmente también haya sido considerado en todos los modelos un vector de parámetros temporales no estructurado, $\rho = (\rho_1, \ldots, \rho_{17})$, modelizado tal que $\rho \sim \mathcal{N}(0, \omega_\rho^{-1} I_{17})$, se ha visto que su contribución es mínima y que las estimaciones apenas mejoran, por lo que se ha decidido finalmente no incluirlo.

Por último, todos los modelos contarán con algún tipo de interacción espacio-temporal. Se han implementado todos los tipos, es decir, las interacciones I, II, III y IV, por lo que la modelización del vector de efectos espacio-temporales $\delta = (\delta_1, \ldots, \delta_{105\times17})$ se ha denotado de la siguiente manera:

- Interacción tipo I:

$$\delta^1 = (\delta_1^1, \ldots, \delta_{105\times17}^1) \sim \mathcal{N}(0, [\omega_{\delta^1} R_{\delta^1}]^-)$$

- Interacción tipo II:
  $\delta^2 = (\delta_1^2, \ldots, \delta_{105\times17}^2) \sim \mathcal{N}(0, [\omega_{\delta^2} R_{\delta^2}]^-)$
- Interacción tipo III:
  $\delta^3 = (\delta_1^3, \ldots, \delta_{105\times17}^3) \sim \mathcal{N}(0, [\omega_{\delta^3} R_{\delta^3}]^-)$
- Interacción tipo IV:
  $\delta^4 = (\delta_1^4, \ldots, \delta_{105\times17}^4) \sim \mathcal{N}(0, [\omega_{\delta^4} R_{\delta^4}]^-)$

Para todos estos casos, las matrices de estructura llamadas como R_{δ^1}, R_{δ^2}, R_{δ^3} y R_{δ^4} se definen en base al producto de Kronecker definido en la Tabla 2.1, y el parámetro de precisión ω, es modelizado como todos los anteriores.

Recapitulando, se han ajustado en total 8 modelos espacio-temporales combinando las diferentes distribuciones *a priori* para cada uno de los vectores aleatorios. La componente espacial se modeliza con BYM o iCAR, la temporal estructurada con un RW de primer orden, y además se incluye una interacción de tipo I, II, III o IV. A continuación se muestra el nombre y la estructura aditiva de cada uno de los modelos planteados:

- BYM-I: $\eta_{it} = \beta_0 + \xi_i^1 + \gamma_t + \delta_{it}^1$
- BYM-II: $\eta_{it} = \beta_0 + \xi_i^1 + \gamma_t + \delta_{it}^2$
- BYM-III: $\eta_{it} = \beta_0 + \xi_i^1 + \gamma_t + \delta_{it}^3$
- BYM-IV: $\eta_{it} = \beta_0 + \xi_i^1 + \gamma_t + \delta_{it}^4$
- iCAR-I: $\eta_{it} = \beta_0 + \xi_i^2 + \gamma_t + \delta_{it}^1$
- iCAR-II: $\eta_{it} = \beta_0 + \xi_i^2 + \gamma_t + \delta_{it}^2$
- iCAR-III: $\eta_{it} = \beta_0 + \xi_i^2 + \gamma_t + \delta_{it}^3$
- iCAR-IV: $\eta_{it} = \beta_0 + \xi_i^2 + \gamma_t + \delta_{it}^4$

Por último, cabe destacar que la estrategia de aproximación INLA utilizada para ajustar los modelos ha sido la llamada *"simplified.laplace"* (ver Rue et al., 2009 [3]).

3.3.1. Selección de modelos

En base a los criterios de selección de modelos basados en la *deviance* vistos en la sección 2.4.6, se ha decidido escoger los 2 mejores modelos obtenidos. A continuación, se muestran las tablas ilustrando los resultados logrados para cada uno de ellos:

Modelo	DIC	Mean deviance	Eff.params	WAIC	Eff.params
BYM-I	11509,04	11143,43	364,49	11520,38	320,94
BYM-II	11474,61	11225,94	248,44	11501,60	244,74
BYM-III	11516,85	11248,88	254,83	11551,38	256,10
BYM-IV	**11471,27**	11250,09	225,58	**11501,26**	226,36
iCAR-I	11507,33	11143,33	364,89	11519,89	321,01
iCAR-II	11473,21	11225,65	248,99	11501,12	244,99
iCAR-III	11522,50	11272,58	264,29	11555,97	261,39
iCAR-IV	**11471,56**	11245,05	230,85	**11500,17**	230,20

Tabla 3.1: Valores dados por los diferentes criterios de información para los 8 modelos espacio-temporales propuestos.

Si se observan los resultados obtenidos en la tabla anterior respecto al DIC y WAIC (son las medidas que nos conciernen, aunque `R-INLA` proporciona todas las demás medidas también), se deduce que los modelos que proporcionan mejores estimaciones para este conjunto de datos son los dos que incluyen la interacción espacio-temporal IV. En este caso, los resultados obtenidos para la interacción IV combinada con un modelo espacial BYM o iCAR no cambian mucho entre sí, y se ha decidido escoger ambos modelos para los cuales se analizará a continuación los resultados obtenidos.

3.3.2. Resultados de los modelos

Finalmente, una vez seleccionados los dos modelos que proporcionan mejor ajuste (aparecen en negrita en la tabla que precede), el último objetivo es mostrar las estimaciones y resultados obtenidos.

Primeramente, se ilustran las distribuciones *a posteriori* logradas para los interceptos globales o efectos fijos en cada uno de los dos modelos (los que denominamos como β_0).

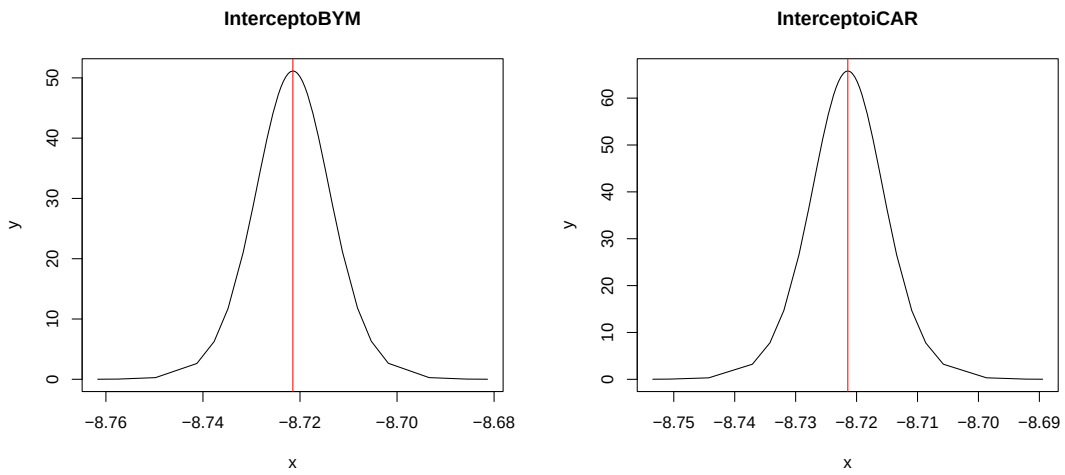


Figura 3.13: Aproximación de las distribuciones *a posteriori* para los interceptos globales de los modelos seleccionados, junto con la mediana (rojo).

Para ambos modelos, la mediana obtenida a partir de las distribuciones *a posteriori* de los interceptos es similar, y se sitúa entorno a -8.72. Teniendo en cuenta que trabajamos con una distribución de Poisson para los datos observados, y que se utilizó una función *link* logaritmo para enlazar las tasas con el predictor lineal del modelo, se necesita exponenciar este valor, es decir, $\exp(-8{,}72) = 1{,}63 \times 10^{-4}$, por tanto, una tasa media global (para todo el periodo y conjunto de regiones) de 16,3 casos por cada cien mil habitantes. Esto simplemente nos indica que las tasas crudas estimadas (las cuales se mostrarán después) por los modelos oscilarán entorno a esos casos, variando hacia arriba o abajo dependiendo de los valores medios estimados por los efectos espaciales, temporales y espacio-temporales.

Por otro lado, se pueden deducir también las aproximaciones de las distribuciones *a posteriori* para cada uno de los hiperparámetros de precisión, los cuales serán en el caso del primer modelo que consideraba el modelo espacial BYM, las precisiones para la componente estructurada (ω_{ξ_1}) y no estructurada espacial (ω_{ξ_2}), componente estructurada (ω_γ) temporal y componente espacio-temporal de interacción de tipo IV (ω_{δ^4}). En cambio, para el caso del modelo espacio-temporal que consideraba el modelo espacial iCAR, tendremos los mismos hiperparámetros de precisión sin la componente espacial no estructurada (en este caso, el parámetro de precisión espacial se denotó ω_{ξ^1}).

Cada uno de estos hiperparámetros explica la variabilidad de su correspondiente parámetro. Por ejemplo, los hiperparámetros espaciales ω_ξ cuanto mayor sean menor será la variabilidad de los parámetros espaciales ξ, y por el contrario valores mas pequeños de ω_ξ aumentarán la variabilidad de los parámetros ξ. Del mismo modo pará el resto de parámetros e hiperparámetros. A continuación se muestran las distribuciones *a posteriori* obtenidas para cada uno de los hiperparámetros para los dos modelos seleccionados.

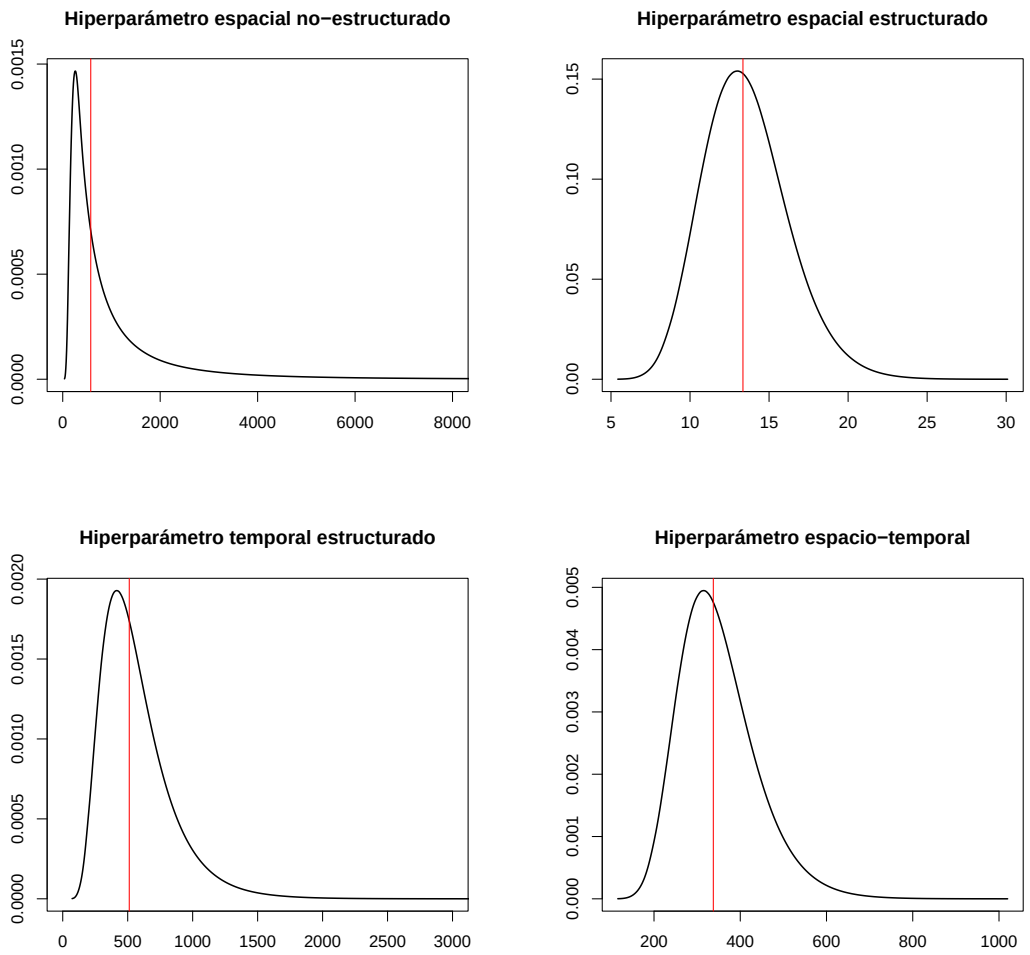


Figura 3.14: Distribuciones *a posteriori* para los hiperparámetros de precisión del modelo BYM-IV y sus respectivas medianas (rojo).

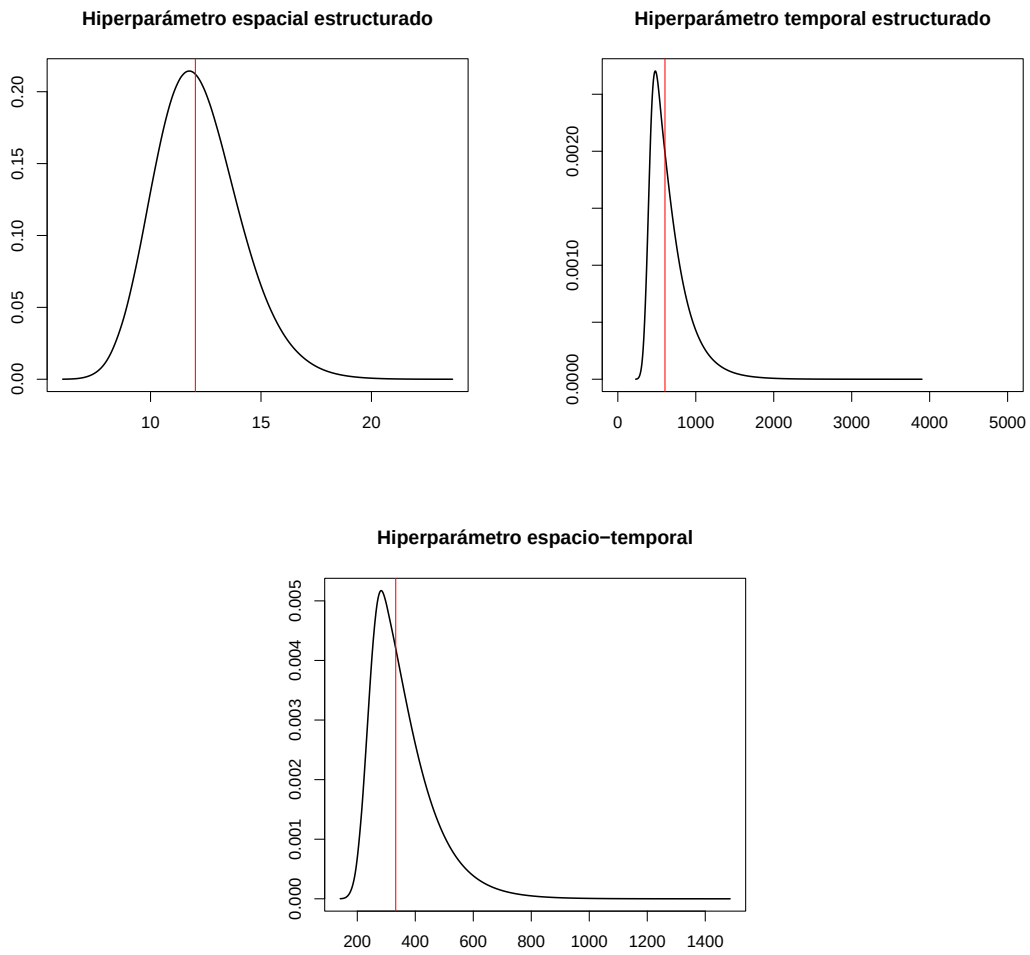


Figura 3.15: Distribuciones *a posteriori* para los hiperparámetros de precisión del modelo iCAR-IV y sus respectivas medianas (rojo).

Por último, y como resultado esencial de este análisis numérico de los modelos propuestos, se muestra la representación mediante mapas de las medias posteriores estimadas para las tasas de cada región y año de estudio dadas por los dos modelos. En este caso, debido a la gran similitud que presentan las estimaciones de los modelos entre sí, se ilustran solamente los resultados para el modelo BYM-IV, siendo practicámente análogos los mapas dados por el modelo iCAR-IV.

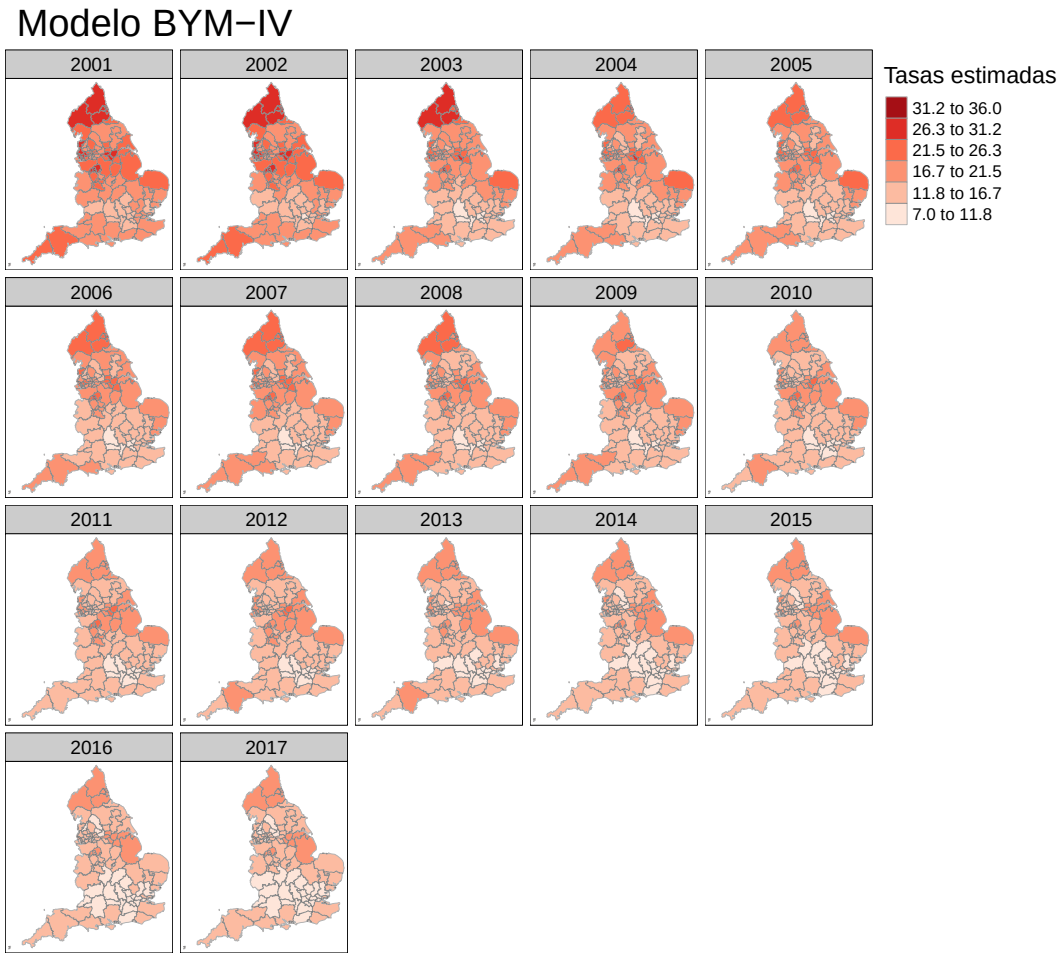


Figura 3.16: Medias posteriores estimadas para cada año y región con el modelo BYM-IV.

Si nos fijamos en los resultados logrados, podemos observar claramente que las tasas ajustadas mediante el modelo BYM-IV (del mismo modo para el modelo iCAR-IV) son bastante similares a las tasas crudas cartografiadas en la Figura 3.11, y que simulan el patrón espacio-temporal de los valores reales de una manera convincente, obteniendo resultados similares a los reales.

Asimismo, la media de 16,3 casos por cada cien mil habitantes que proporcionaban las estimaciones del intercepto, se ven reflejadas en las estimaciones de las tasas, que fluctuan alrededor de este valor, como cabía esperar.
Para acabar, se ve que la tendencia de la incidencia del cáncer con el paso de los años es a la baja, por lo que la estructura temporal introducida a los modelos es vital. Además, se observa que hay un patrón espacial o geográfico, por lo que emplear modelos que incorporen términos espaciales también es apropiado.

De ahora en adelante, para evitar redundancia en los resultados, se trabajará solamente con modelos que utilicen la estructura espacial del modelo BYM, ya que todos los resultados obtenidos mediante el modelo espacial iCAR son prácticamente idénticos.

Capítulo 4

Proceso de validación de los modelos y predicción a futuro

Después de haber ajustado diferentes modelos espacio-temporales con el objetivo de suavizar las tasas, pasamos a este último capítulo. Aquí, se ha diseñado un proceso de validación que nos permitirá determinar qué modelo espacio-temporal es el más adecuado para predecir tasas a corto plazo, ilustrando el procedimiento con los datos de incidencia por cáncer de estómago en hombres para el periodo 2001-2017. Una vez seleccionado el mejor modelo en base a diferentes medidas de validación, se realizarán predicciones de las tasas para los años 2018, 2019 y 2020.

Es vital recalcar la importancia de este capítulo, dado que resulta indispensable que las instituciones sanitarias de cada país desarrollen y evalúen estrategias de prevención para determinadas enfermedades. Habitualmente, los datos de incidencia y mortalidad están disponibles con dos o tres años de retraso respecto al año natural debido a la complejidad en la recolección y procesamiento de los datos. En este sentido, es importante contar con procedimientos estadísticos que permitan realizar las predicciones a corto plazo, además de estudiar la validez de los métodos empleados. Además, si a esto se le añade el hecho de que el año 2020 fue el primero sacudido por la pandemia de la enfermedad SARS-CoV-2 (COVID-19), es posible que los registros sanitarios puedan haber clasificado posibles casos de incidencia o mortalidad por cáncer u otras enfermedades respiratorias como COVID-19, o viceversa. En este contexto, se convierte incluso más importante poder aportar un proceso estadístico de validación y predicción que proporcione predicciones de incidencia fiables a un futuro cercano.

La metodología utilizada en lo referente a los procesos de validación y predicción que se desarrolla durante este capítulo, se ha basado en los trabajos de investigación publicados en los artículos [18] y [16]. Se utiliza un proceso que consiste en comparar mediante diferentes medidas de validación, los valores predichos por los modelos con observaciones de las que se ya se dispone. De esta manera, es posible evaluar la capacidad predictiva de cada uno de los modelos ajustados y seleccionar el mejor.

4.1. Primera etapa del proceso de validación

El proceso de validación desarrollado se centrará en obtener predicciones de tasas en distintos periodos de tiempo para los cuales se dispone de valores observados, de manera que se tratará de comparar lo predicho con lo observado. Esto se realizará de la siguiente forma: en primer lugar se creará una partición adecuada de los datos disponibles, es decir, diferentes subconjuntos partiendo de la base de datos inicial. Después, se ajustarán a cada uno de estos subconjuntos o configuraciones los modelos espacio-temporales descritos en el Capítulo 3.
En este caso, se ha adaptado la metodología implementada en los artículos [18] y [16] a nuestros datos. Como se dispone de datos de incidencia de cáncer para el periodo de años 2001-2017, se ha diseñado en `R` una rutina (una función), la cual realiza la división de datos por periodos de años y que además permite hacer esta partición de diversas maneras, según interese al usuario dependiendo de los datos que disponga. Para el ejemplo práctico con el que se está tratando desde el capítulo anterior, se ha decidio crear los subconjuntos de validación de la siguiente manera:

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
$k=1$																	
$k=2$																	
$k=3$																	
$k=4$																	
$k=5$																	

Tabla 4.1: Tabla de cada uno de los $k \in \{1, \ldots, 5\}$ subconjuntos o configuraciones de validación creados. Las celdas en naranja muestran el periodo de los datos utilizados para ajustar los modelos; las celdas en azul, turquesa y verde son los años para los cuales el modelo realiza predicciones; un año, dos años y tres años a futuro, respectivamente.

Por tanto, se ha creado una partición del conjunto total de datos disponibles en 5 subconjuntos (ver Tabla 4.1), en la cual cada uno de los subconjuntos parte con datos correspondientes a un intervalo de tiempo de 10 años (en la tabla, se muestran de color naranja) que se utilizarán para ajustar los modelos descritos en el Capítulo 3. Una vez los modelos hayan sido ajustados, se realizarán predicciones de casos para los próximos tres años, y se evaluarán los resultados empleando distintas medidas de validación que veremos a continuación.

En resumen, se comienza ajustando los modelos en el periodo k=1 que corresponde a los datos del intervalo 2001-2010, y después se predicen las tasas para los años 2011, 2012 y 2013. Posteriormente, el subconjunto k=2 seleccionará los datos en el intervalo 2002-2011, prediciendo los años 2012, 2013 y 2014, de modo que sucesivamente se llegará al último año disponible en la base de datos, 2017. Ya se ha mencionado anteriormente, que la rutina implementada en `R` permite crear estos subconjuntos de validación de modo que interese más al usuario, y en este caso se ha considerado que esta es una buena partición de datos, ya que en el estudio del cáncer es conveniente contar con periodos de, al menos, 10 años.

4.2. Estrategias para la reducción del coste computacional del proceso de validación

El proceso de validación planteado requiere ajustar múltiples modelos espacio-temporales en cada uno de los subconjuntos de datos o configuraciones de validación. Debemos tener en cuenta que el coste computacional de cada modelo puede llegar a ser realmente elevado, debido principalmente a la gran dimensionalidad de las matrices de precisión de los efectos aleatorios del modelo, así como al número de restricciones de identificación requeridas. No obstante, el mayor coste computacional que requiere a la técnica INLA es la estimación de las distribuciones *a posteriori* de los hiperparámetros del modelo. Teniendo en cuenta que en un modelo de predicción, la estimación de estos parámetros únicamente depende de los datos observados, se ha optado por emplear estrategias que nos permitan reducir el coste computacional de esos modelos.

Para ello, `R-INLA` proporciona una opción dentro de la función general `inla()` llamada `control.mode()`. Este argumento permite recoger los hiperparámetros estimados para un modelo referencia y utilizarlos en otro modelo ajustado a un conjunto de datos más o menos grande. En nuestro caso, utilizaremos como modelo referencia los modelos BYM ajustados al periodo completo 2001-2017, y después se utilizarán sus hiperparámetros estimados como valores iniciales en los modelos ajustados a cada uno de los subperiodos.

La función `control.mode()` toma, entre otros, dos argumentos principales. Por un lado, recoge las modas *a posteriori* de los hiperparámetros de un modelo ya ajustado, y por otro lado, el argumento booleano "`restart`". Este último se puede seleccionar como `TRUE` o `FALSE`, y aquí está la verdadera ventaja de esta función y con la que lograremos reducir el coste computacional.

Esencialmente, se le comunica internamente a `R-INLA` que tome los hiperparámetros del modelo referencia y los utilice para ajustar el nuevo modelo. De este modo, se reduce coste computacional al no ser necesario aproximar nuevas distribuciones *a posteriori* para los hiperparámetros del nuevo modelo. Evidentemente, veremos que esta opción funciona en este caso puesto que se utilizan como referencia los modelos ajustados en el Capítulo 3 para el periodo completo 2001-2017, y ahora solamente se van a querer ajustar los mismos modelos en periodos de tiempo más cortos (en cada uno de las 5 subperiodos vistos en la Tabla 4.1). La diferencia de utilizar `restart = TRUE` o `restart = FALSE` como argumento es simple; si se utiliza el primero, se impone a `R-INLA` que utilice los hiperparámetros del modelo referencia como valores iniciales. Es decir, el modelo nuevo parte de estos parámetros referencia e internamente se siguen optimizando hasta llegar a las aproximaciones óptimas dadas por `inla()`. En cambio, si se utiliza la opción `restart = FALSE`, se fijan los hiperparámetros del modelo nuevo en los del modelo referencia, y no se optimizan más.

Para llevar a cabo el proceso de validación de los cuatro modelos BYM ajustados en el Capítulo 3 en cada uno de los 5 subconjuntos mostrados en la Tabla 4.1, se ha decidido adoptar la estrategia `restart = FALSE`, utilizando como valores fijos los hiperparámetros dados por los modelos BYM ajustados al periodo completo 2001-2017. En el Apéndice A, se comparan los tiempos computacionales realizando el ajuste clásico frente al uso del argumento `control.mode()`. Los resultados

obtenidos muestran reducciones de hasta un 50 % en los tiempos de ejecución.

Además, ha sido fundamental comprobar que la técnica utilizada no altera los resultados dados por los modelos, es decir, que las tasas estimadas son prácticamente idénticas entre el ajuste normal del modelo y la estrategia `restart = FALSE`. Asimismo, es importante destacar que se ha realizado un análisis que determina que las estimaciones de los hiperparámetros apenas cambian entre el modelo global y aquellos ajustados en cada subperiodo, por lo que es viable asumir hiperparámetros globales. Una vez más, se muestran en el Apéndice B algunos resultados adicionales de este estudio donde se observa, efectivamente, que los resultados obtenidos con ambas metodologías son similares. Se ilustran solamente un par de ejemplos, no obstante, los resultados han sido replicados para todos los subperiodos de tiempo y años de estudio.

4.3. Medidas de validación

En esta sección, se introducen varios criterios o medidas de validación mediante los cuales se evaluará a posteriori la capacidad predictiva de los modelos.
En primer lugar, se presentan dos medidas estadísticas basadas en el error de cada predicción respecto al valor de incidencia observado. Estas medidas se calcularán en cada una de las 105 regiones de estudio para las predicciones a $p = 1, 2$ y 3 años.

MAE (*Mean Absolute Error*)

$$MAE_i^{(p)} = \frac{1}{5} \sum_{t=2010+p}^{2017+p} |y_{it} - \hat{y_{it}}| \qquad i = 1, \ldots, 105 \tag{4.1}$$

RMSE (*Root Mean Squared Error*)

$$RMSE_i^{(p)} = \sqrt{\frac{1}{5} \sum_{t=2010+p}^{2017+p} (y_{it} - \hat{y_{it}})^2} \qquad i = 1, \ldots, 105 \tag{4.2}$$

Recordemos que y_{it} es el número de observaciones de la región i en el año t, y en este caso $\hat{y_{it}}$ indica el valor esperado de las predicciones (recalcar la importancia de la sección 2.5.2) en la misma región y año, es decir, la media de la distribución predictiva de los casos (ver apéndice A del artículo [16] para más detalles sobre su cálculo). La fracción $\frac{1}{5}$ se utiliza para calcular el error medio, puesto que contamos con 5 configuraciones o subconjuntos de datos. Además, dependiendo de si se escoge $p = 1, 2$ o 3, se obtendrá una media del error de predicción un año, dos o tres al futuro, es decir, observando la Tabla 4.1, estaríamos calculando el error de predicción de los colores azul ($p = 1$), turquesa ($p = 2$) y verde ($p = 3$).

Otra medida que será utilizada como criterio de validación es el *Interval Score* (IS), una regla de decisión que combina tanto la longitud como la cobertura empírica del intervalo de credibilidad en una medida única de puntuación (artículo [14]).

IS (*Interval Score*)

$$IS_\alpha(y_{it}) = (u_{it} - l_{it}) + \frac{2}{\alpha}(l_{it} - y_{it})I(y_{it} < l_{it}) + \frac{2}{\alpha}(y_{it} - u_{it})I(y_{it} > u_{it}) \qquad (4.3)$$

donde y_{it} denota como antes la observación de la región i en el año t, mientras que l_{it} y u_{it} denotan el límite inferior y superior del intervalo de credibilidad al $(1-\alpha)\cdot 100\,\%$, es decir, el intervalo de credibilidad correspondiente a la distribución predictiva *a posteriori* para cada una de las observaciones predichas (para el análisis posterior, se seleccionará $\alpha = 0{,}05$). Por último, la función $I(\cdot)$ es la función característica que penaliza el tamaño del intervalo de credibilidad, si la observación y_{it} no está contenida en él.

Como se puede ver, para cada uno de los tres criterios presentados, valores más pequeños indicarán un mejor ajuste predictivo del modelo.

4.4. Ajuste de los modelos espacio-temporales a cada configuración o subconjunto de validación

Finalmente se lleva a cabo el proceso de validación expuesto en las secciones anteriores. Para ello, se recuperan del Capítulo 3 los cuatro modelos espacio-temporales ajustados con el modelo espacial BYM. Por tanto, se trabajará con los modelos llamados como BYM-I, BYM-II, BYM-III y BYM-IV. No obstante, estos modelos fueron ajustados para todo el periodo de estudio 2001-2017, y ahora estamos dividiendo este intervalo de tiempo en cinco subconjuntos, por lo que para cada uno de los subconjuntos se deberán reescribir los modelos de tal manera que el vector de efectos temporales estructurado y el vector de efectos espacio-temporales se modelicen de acuerdo con cada subconjunto. Es decir, la distribución *a priori* que modeliza estos vectores aleatorios deberá modificarse, puesto que la construcción de la matriz de estructura temporal definida como R_t cambia al modificarse el periodo de tiempo.

Volviendo al modelo espacio-temporal genérico definido en la ecuación 3.2, está claro que la modelización del vector de parámetros o efectos aleatorios espaciales $\xi = (\xi_1, \ldots, \xi_{105})$ no cambia, ya que que la matriz de estructura espacial R_s se mantiene. En cambio, el vector de efectos temporales modelizado en todos los casos analizados por un RW de primer orden (recordamos que $\gamma \sim \mathcal{N}(0, [\omega_\gamma R_t]^-)$), ahora denotado por $\gamma = (\gamma_1, \ldots, \gamma_{13})$ (cada uno de los subconjuntos dispone ahora de información para 13 años, de los cuales los últimos 3 se predicen), seguirá de nuevo una distribución normal multivariante de media 0, pero a la hora de ajustar los modelos se ha recalculado internamente la matriz de estructura temporal R_t, puesto que de acuerdo con la ecuación 2.24, el vector ordenado de años cambia el vecindario en cada caso. Aunque en teoría también se debería modificar la distribución del vector de efectos temporal sin estructura (denotado como ρ), ya se vio que su relevancia era prácticamente nula y por tanto no sería incluida en los modelos. Por último, debido a que se ha redefinido R_t para cada una de las

5 configuraciones, la distribución del vector de interacciones espacio-temporales también cambiará en cada caso, por lo que ahora $\delta = (\delta_1, \ldots, \delta_{105\times 13})$ se modeliza así: $\delta \sim \mathcal{N}(0, [\omega_\delta R_\delta]^-)$, siendo R_δ el producto de Kronecker dependiendo de la interacción de tipo I, II, III o IV. Resumiendo, debido a que el periodo de tiempo completo ha sufrido una reducción en cada una de las configuraciones, ha sido necesario adaptar las matrices de estructura de los modelos implementados.

Una vez aclaradas estas modificaciones, se ha procedido a ajustar cada uno de los cuatro modelos mencionados previamente, utilizando cada subconjunto de validación ($k \in \{1, \ldots, 5\}$). Primero, se han calculado los valores dados por los criterios de información DIC y WAIC para cada modelo ajustado mediante cada uno de los subconjuntos, y los resultados obtenidos se pueden ver en las tablas del Apéndice C. Como se puede observar, los resultados dados por los modelos en el periodo completo 2001-2017 no cambian mucho analizándolos por subconjuntos, y los modelos con el DIC/WAIC más bajo siguen siendo los modelos que llamamos BYM-II y BYM-IV, mostrando este último valores similares o ligeramente mejores comparados a la interacción tipo II.

Después, se han analizado los resultados de predicción obtenidos en base a las medidas de validación introducidas en la sección 4.3. Concretamente, se han calculado los valores medios del MAE, RMSE e IS (con valor de credibilidad $\alpha = 0{,}05$) sobre las 105 regiones de estudio del territorio inglés, para predicciones un año, dos y tres a futuro, con cada uno de los cuatro modelos ajustados. La tabla que sigue a continuación resume de manera visual las computaciones mencionadas.

	1 año a futuro			2 años a futuro			3 años a futuro		
Modelo	MAE	RMSE	IS	MAE	RMSE	IS	MAE	RMSE	IS
BYM-I	5,38	6,35	34,69	5,55	6,63	30,38	5,92	6,99	34,03
BYM-II	5,33	**6,29**	**30,80**	5,49	**6,59**	30,69	**5,87**	6,98	**32,38**
BYM-III	5,36	6,32	35,51	5,52	**6,59**	**29,30**	**5,87**	**6,93**	32,52
BYM-IV	**5,30**	6,30	32,35	**5,48**	6,60	30,32	5,95	7,05	33,98

Tabla 4.2: Resultados obtenidos para los valores medios de las medidas de validación propuestas para cada modelo con predicciones a un año, dos y tres al futuro.

Como se puede observar, todos los modelos presentan mejores resultados en general respecto a las tres medidas de validación si el número de años predicho es menor; es decir, los valores del MAE, RMSE e IS aumentan si se incrementa el número de años a predecir al futuro, lo cual es lógico.
Además, se aprecia una muy ligera tendencia de mejora del modelo con interacción tipo II respecto al de interacción tipo IV, cuando el número de años a predecir incrementa. Cuando se predice uno o dos años a futuro, parece que las predicciones dadas por el modelo BYM-IV son las mejores; no obstante, las estimaciones a tres años son mejores con el modelo BYM-II, incluso con el BYM-III. Se podría concluir que en este estudio de modelización predictiva de incidencia de cáncer de estómago en hombres del territorio inglés en concreto, si se pretende predecir a un futuro cercano, el modelo de interacción IV parece ser el más adecuado. Por el contrario, si las instituciones sanitarias carecen de información sobre incidencia

de la enfermedad a un futuro más lejano, se recomendaría más optar por modelos con interacción II o incluso interacción III.

Debido a que en la siguiente sección se van a ilustrar mediante mapas cartográficos las predicciones de tres años a futuro para las tasas crudas, se ha decidido optar por el modelo BYM-II para hacer dicho análisis, puesto que parece el modelo más completo.

4.5. Predicciones de tasas de incidencia del cáncer de estómago para los años 2018-2020

El objetivo final de esta sección es ilustrar la predicción de tasas de incidencia de cáncer de estómago hasta tres años a futuro. Para ello, se han obtenido las poblaciones correspondientes a cada región inglesa en los años 2018, 2019 y 2020 (proporcionadas en abierto por *Office for National Statistics*, [10]), permitiendo de esta manera realizar la predicción correspondiente mediante el modelo BYM-II, el cual hemos determinado que mostraba mayor capacidad predictiva a corto plazo.

A continuación, se muestran los mapas con la evolución temporal de las estimaciones de las tasas para todo el periodo 2001-2020. Evidentemente, las tasas ajustadas de los años 2018, 2019 y 2020 corresponden a las predicciones dadas por el modelo.

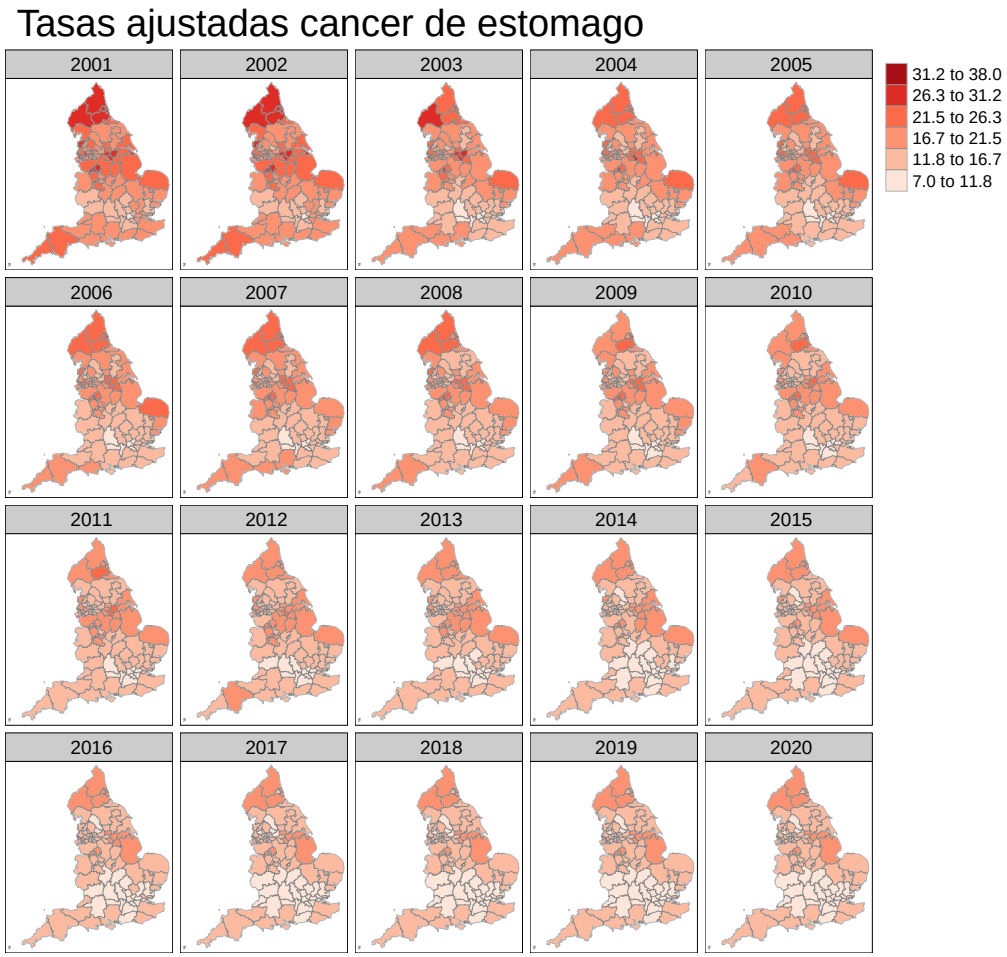


Figura 4.1: Tasas posteriores ajustadas por el modelo predictivo BYM-II para todo el periodo 2001-2020. Los años 2018-2020 han sido predichos.

En general, la evolución temporal de los mapas muestra un decrecimiento claro de las tasas ajustadas en cada una de las regiones inglesas. Se puede destacar un patrón espacio-temporal evidente, similar al obtenido para el modelo ajustado en la Figuras 3.16, el cual no consideraba los años 2018-2020. Respecto a estos últimos años, las predicciones realizadas muestran valores de acuerdo con el resto de estimaciones, y las tasas posteriores predichas parecen razonables. Por tanto, se puede concluir que el modelo BYM-II empleado para el proceso de predicción es, en un principio, una buena herramienta de estimación de riesgos a un futuro cercano para cuando no se disponga de datos suficientes.

Por último, se han escogido algunas de las 105 regiones totales del área de Inglaterra, y se ha analizado la evolución temporal de las tasas ajustadas por 100.000 habitantes estimadas por el modelo predictivo BYM-II para todo el periodo en cada una de ellas. Concretamente, se han seleccionado las regiones de Birmingham (identificada como E38000258/E38000259), Bristol (identificada como E38000222) y Lancashire (identificada como E38000200), puesto que son regiones con poblaciones muy diversas, y puede resultar de interés analizar el patrón temporal de cada una de ellas. Además de mostrar la línea temporal de

las medianas estimadas para cada región, se han dibujado también los intervalos de credibilidad correspondientes del 95 %, para poder apreciar la variabilidad de las tasas estimadas en regiones con menos habitantes. Para acabar, también se muestran las tasas crudas observadas en cada una de las tres regiones para cada año, y se comparan con los valores predichos.

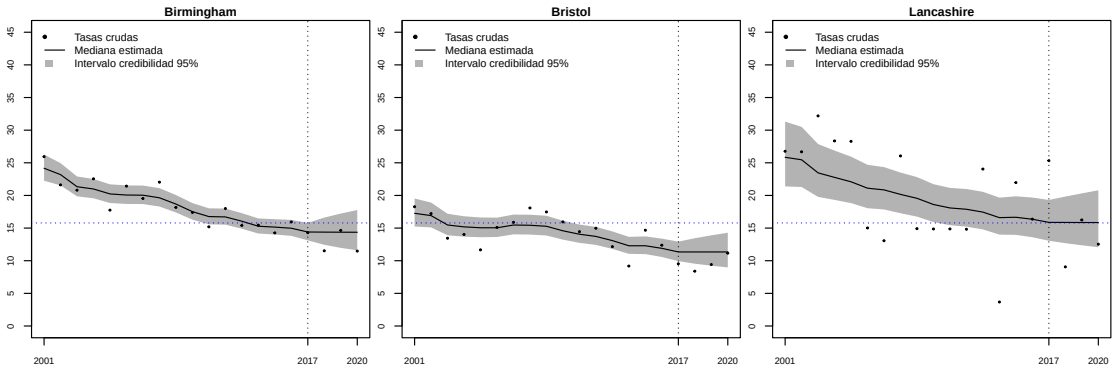


Figura 4.2: Medianas posteriores predictivas (línea contínua) por 100.000 habitantes para las regiones inglesas de Birmingham (izquierda), Bristol (centro) y Lancashire (derecha) con sus correspondientes intervalos de credibilidad del 95 % (en gris), durante el periodo 2001-2020. La línea vertical denota el año a partir del cual comienzan las predicciones, y la línea horizontal azul es el valor medio de las tasas ajustadas. Las tasas crudas se muestran como puntos negros.

Las tres regiones escogidas muestran un patrón temporal similar en lo que respecta a la evolución de las medianas estimadas en todo el periodo 2001-2020, y la línea contínua decrece del mismo modo que lo hacen las tasas crudas en la Figura 3.8, la cual analizaba las observaciones de incidencia y mortalidad en todo el área.

Por otro lado, los intervalos de credibilidad ilustrados en color gris muestran anchuras diferentes, lo cual es lógico al ser Birmingham la región más poblada y Lancashire la menos poblada. Es por esto que en Birmingham la línea contínua de la mediana está mejor ajustada a los valores reales (puntos negros), y a medida que la población en riesgo de una región decrece, aumenta la variabilidad como se puede apreciar en el gráfico correspondiente a la región de Lancashire. Por último, cabe destacar que al menos para las regiones seleccionadas como ejemplo, aunque las tasas crudas muestren tendencias diferentes (Bristol por ejemplo muestra una tendencia oscilatoria), el modelo predictivo simula claramente un patrón temporal decreciente de las tasas, lo cual cumple el objetivo principal del suavizado de tasas en regiones que no reflejan una tendencia real debido a su gran variabilidad.

Capítulo 5

Conclusiones y agradecimientos

El uso de la modelización espacio-temporal en el ámbito de la representación cartográfica de enfermedades o *disease mapping*, ha demostrado ser una potente herramienta. Puesto que se ha tratado con datos reales de cáncer de estómago, el cual es un cáncer con poca incidencia en general, y además los datos se han desagregado por regiones y años, las medidas clásicas de estimación como las tasas crudas pueden resultar bastante variables. En este contexto ha sido vital el desarrollo de algunos modelos que nos han permitido obtener patrones espaciales y temporales subyacentes, pudiendo ilustrar los resultados obtenidos. Es cierto, que el estudio de modelización de cáncer de estómago llevado a cabo solamente se ha presentado para el sexo masculino, y que además se ha tomado la población de cada región sin ser desagregada por grupos de edad, por lo que el análisis realizado puede generalizarse. Para ello, se deberían considerar propuestas de modelización alternativas que permitan inducir estructuras de correlación entre el sexo o los distintos grupos de edad.

Por otro lado, se ha argumentado el interés de crear un procedimiento de validación para seleccionar el mejor modelo espacio-temporal en términos de predicción. El método utilizado se basa en definir particiones temporales del periodo analizado, ajustando los distintos modelos en cada subconjunto y evaluando su capacidad predictiva. Para ello, se ha implementado en `R` una rutina que ha requerido un gran coste computacional debido a una razón principal además de la alta dimensionalidad de las matrices de estructura o precisión, y el número de restricciones de identificación necesarias en cada modelo: el elevado coste de estimar las distribuciones de cada uno de los hiperparámetros. Los resultados obtenidos para las medidas de validación respecto a las predicciones un año, dos y tres al futuro han sido bastante razonables, obteniendo, como era de esperar, peores resultados para predicciones más alejadas en el tiempo.

En relación a la reducción de costes computacionales, se ha implementado con éxito un procedimiento en `R-INLA` que permite ajustar los modelos de cada subconjunto de validación utilizando los hiperparámetros de un modelo ya ajustado con anterioridad en el periodo completo. De esta manera, se ha logrado reducir en aproximadamente un 50 % el coste inicial total.

Como futura línea de investigación, se plantea extender el proceso de validación desarrollado a otras propuestas de modelización para datos espacio-temporales que no se han considerado en el presente trabajo. Respecto a la técnica de reducción

de costes en concreto, está claro que ha superado las expectativas y se han logrado resultados muy positivos, por lo que se puede buscar profundizar más y extender los resultados en modo de líneas futuras, por ejemplo, integrando esta técnica en procesos estadísticos más complejos en forma de paquete en `R`.

5.1. Agradecimientos

Este Trabajo Fin de Máster ha sido realizado bajo la financiación de Ayudas de Iniciación a la Investigación de la Universidad Pública de Navarra en el ámbito de sus institutos de investigación durante el curso académico 2022-2023 (resolución 2359/2022).

Se agradece por un lado a la UPNA por facilitar la utilización de la aplicación *SSTCDApp*, utilizada para el ajuste y selección de los mejores modelos, previos a ser implementados en `R`. Se deben mencionar por consiguiente, [19],[20] y [17], por facilitar la comprensión de tal compleja teoría.

Por último, se mencionan también el equipo de `R` [21] y se destacan los artículos de desarrollo de `R-INLA` [3] y [22].

Apéndice A

Tiempos de ejecución de los modelos: ajuste clásico vs. argumento `control.mode()`

Subperiodo	Clásico	`restart=TRUE`	`restart=FALSE`	Modelo
2001-2013	9,04	6,87	6,51	BYM-I
2002-2014	7,31	6,79	6,76	BYM-I
2003-2015	7,44	6,90	6,42	BYM-I
2004-2016	7,38	6,79	6,52	BYM-I
2005-2017	7,32	6,58	6,73	BYM-I
2001-2013	10,85	9,85	7,53	BYM-II
2002-2014	13,74	12,19	8,12	BYM-II
2003-2015	12,42	10,13	8,17	BYM-II
2004-2016	16,61	10,26	7,87	BYM-II
2005-2017	12,77	9,79	7,83	BYM-II
2001-2013	11,27	10,16	7,22	BYM-III
2002-2014	12,66	10,32	7,37	BYM-III
2003-2015	11,74	9,49	7,40	BYM-III
2004-2016	10,83	10,79	7,11	BYM-III
2005-2017	11,08	9,86	7,10	BYM-III
2001-2013	21,24	15,19	9,58	BYM-IV
2002-2014	23,64	20,00	9,31	BYM-IV
2003-2015	58,80	29,44	16,92	BYM-IV
2004-2016	59,75	14,87	9,14	BYM-IV
2005-2017	24,56	15,14	9,02	BYM-IV

Tabla A.1: Tiempos de ejecución en segundos de los cuatro modelos BYM ajustados a cada subperiodo para cada método de ajuste.

Apéndice B

Resultados adicionales: ajuste clásico vs. argumento `control.mode()`

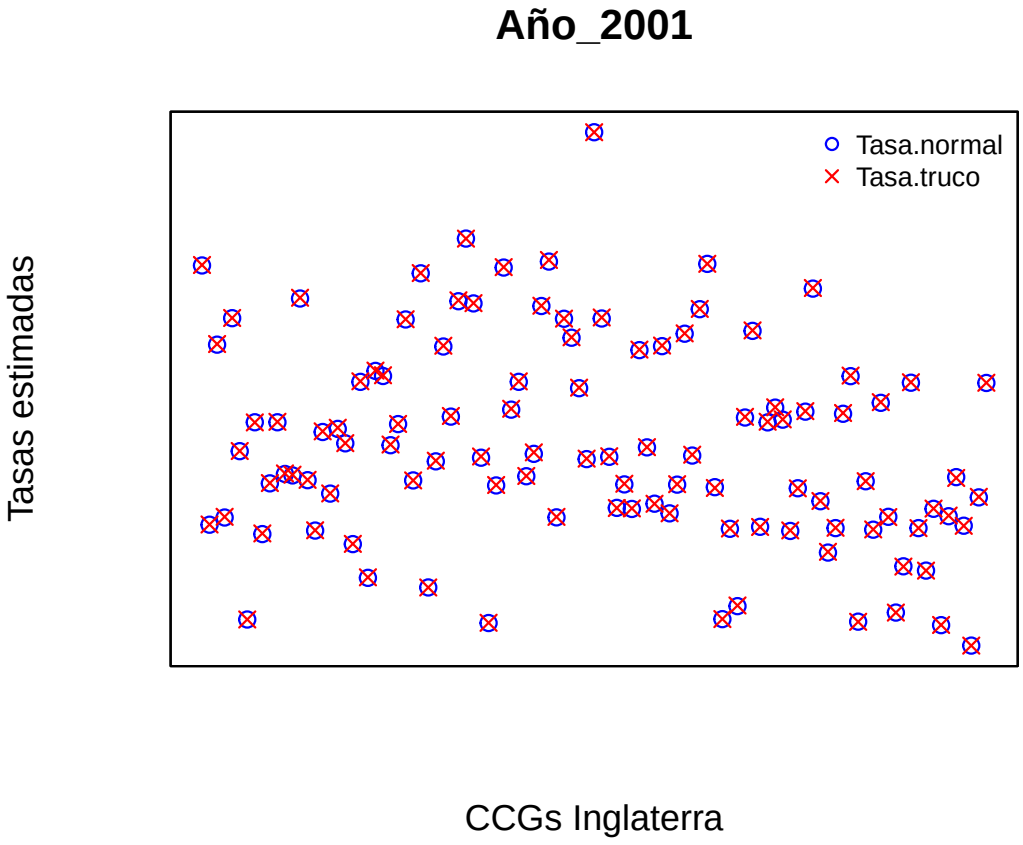


Figura B.1: Tasas estimadas en cada región o CCGs inglesa en el año 2001, por los modelos BYM-II ajustados al periodo 2001-2013. Las tasas estimadas por el modelo clásico aparecen como circulos azules, mientras que las tasas estimadas por el modelo que utiliza `Restart=TRUE` se muestran como cruces rojas.

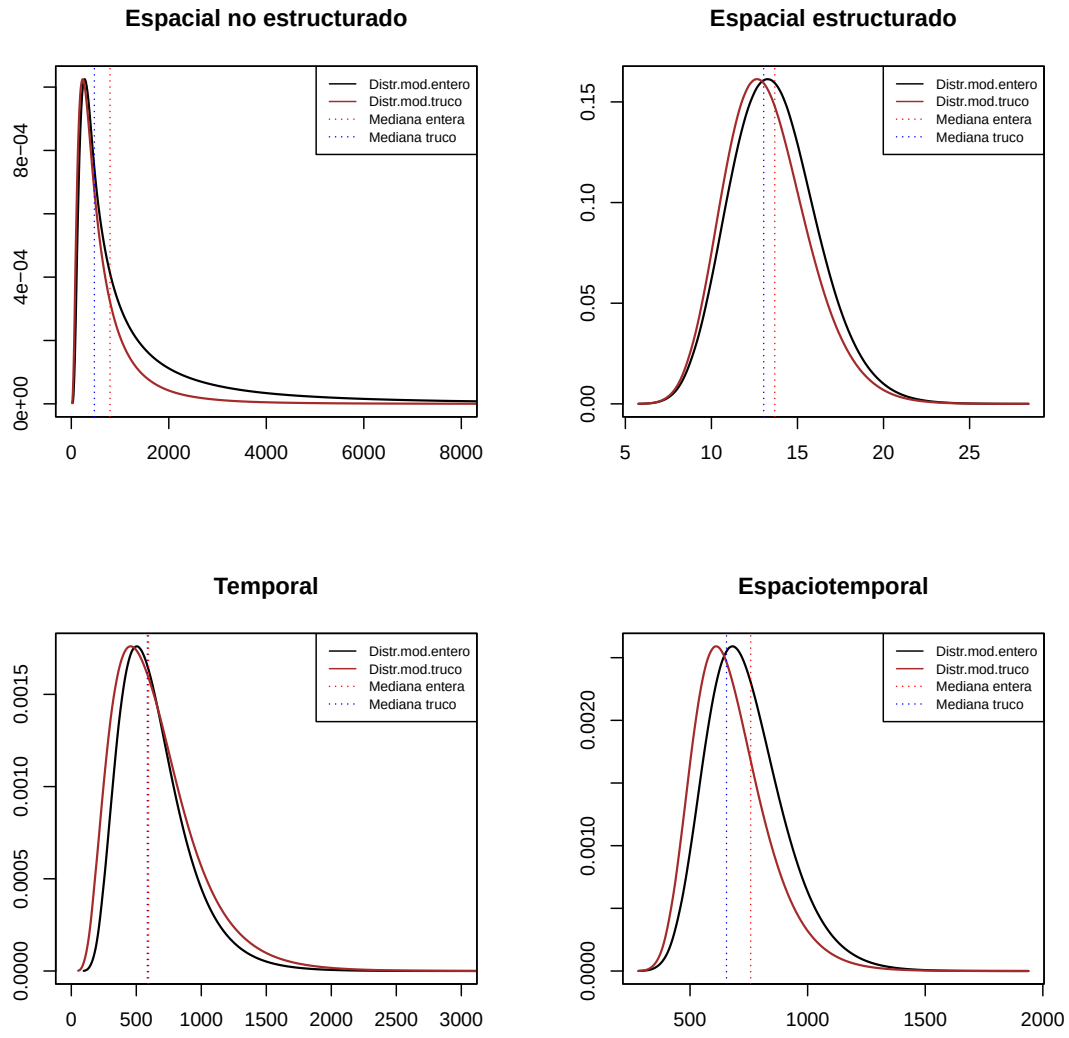


Figura B.2: Comparación entre las distribuciones *a posteriori* de los hiperparámetros del modelo BYM-II ajustado al periodo entero 2001-2017 de manera clásica (negro), y ajustado en el periodo 2001-2013 utilizando `restart=TRUE` (rojo). Las medianas de cada distribución aparecen como líneas verticales discontínuas.

Apéndice C

Comparación de modelos respecto a su DIC/WAIC en cada subconjunto de validación

	DIC BYM-I	DIC BYM-II	DIC BYM-III	DIC BYM-IV
2001-2013	6822,60	6811,41	6812,46	6800,87
2002-2014	6804,69	6784,23	6805,20	6787,11
2003-2015	6767,55	6746,04	6766,26	6747,00
2004-2016	6767,04	6763,27	6772,82	6765,21
2005-2017	6754,66	6760,11	6765,24	6762,01

Tabla C.1: Tabla de valores dados por el criterio de información DIC para los modelos de estructura espacial BYM ajustados con cada subconjunto de validación para todas las interacciones.

	WAIC BYM-I	WAIC BYM-II	WAIC BYM-III	WAIC BYM-IV
2001-2013	6820,23	6825,22	6822,28	6814,25
2002-2014	6806,63	6796,36	6818,09	6802,01
2003-2015	6766,55	6756,32	6774,96	6757,99
2004-2016	6768,64	6780,33	6787,13	6784,52
2005-2017	6759,56	6780,68	6784,72	6785,73

Tabla C.2: Tabla de valores dados por el criterio de información WAIC para los modelos de estructura espacial BYM ajustados con cada subconjunto de validación para todas las interacciones.

Bibliografía

[1] Etxeberria, J., Goicoa, T., and Ugarte, M.D. Evaluating space-time models for short-term cancer mortality risk predictions in small areas. *Biometrical Journal*, 56(3), 2013. https://doi.org/10.1002/bimj.201200259.

[2] Ugarte, M.D., Goicoa, T., Etxeberria, J., and Militino, A.F. Projections of cancer mortality risks using spatio-temporal P-spline models. *Sage Journals*, 21(5), 2012. https://doi.org/10.1177/0962280212446366.

[3] Rue, H., Martino, S., and Chopin, N. Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, 2009. https://doi.org/10.1111/j.1467-9868.2008.00700.x.

[4] Blangiardo, M. and Cameletti, M. *Spatial and Spatio-temporal Bayesian Models with R-INLA*. John Wiley & Sons, 2015.

[5] Gomez-Rubio, V. *Bayesian inference with INLA*. Chapman and Hall/CRC, 2020.

[6] Bakka, H., Rue, H., Fulstad, G.A., Riebler, A., Bolin, D., Illian, J., Krainksi, E., Simpson, D., and Lindgren, F. Spatial modeling with R-INLA: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(6), 2018. https://doi.org/10.1002/wics.1443.

[7] Rue, H., Riebler, A., Sørbye, S.H., Illian, J.B., Simpson, D.P., and Lindgren, F.K. Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017. https://doi.org/10.1146/annurev-statistics-060116-054045.

[8] R-INLA package: version 22.12.16. www.r-inla.org.

[9] National Cancer Registration and Analysis Service. https://www.cancerdata.nhs.uk/incidence_and_mortality.

[10] Office for National Statistics. https://www.ons.gov.uk/peoplepopulationandcommunity.

[11] Knorr-Held, L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19(17-18):2555–2567, 2000. https://doi.org/10.1002/1097-0258(20000915/30)19:17/18<2555::AID-SIM587>3.0.CO;2-%23.

[12] Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and Van Der Linde, A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*, 64(4):583–639, 2002. https://doi.org/10.1111/1467-9868.00353.

[13] Watanabe, S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11(Dec):3571–3594, 2010. https://doi.org/10.48550/arXiv.1004.2316.

[14] Gneiting, T. and Raftery, A.E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. https://doi.org/10.1198/016214506000001437.

[15] Martino, S. and Riebler, A. Integrated Nested Laplace Approximations (INLA). *John Wiley & Sons*, pages 10–12, 2019. https://doi.org/10.48550/arXiv.1907.01248.

[16] Orozco-Acosta, E., Riebler, A., Adin, A., and Ugarte, M.D. A scalable approach for short-term disease forecasting in high spatial resolution areal data. 2023. https://doi.org/10.48550/arXiv.2303.16549.

[17] Goicoa, T., Adin, A., Ugarte, M.D., and Hodges, J.S. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment*, 32:749–770, 2018. https://doi.org/10.1007/s00477-017-1405-0.

[18] Etxeberria, J., Goicoa, T., and Ugarte, M.D. Using mortality to predict incidence for rare and lethal cancers. *Biometrical Journal*, 2022. https://doi.org/10.1002/bimj.202200017.

[19] Adin, A., Goicoa, T., and Ugarte, M.D. Online relative risks/rates estimation in spatial and spatio-temporal disease mapping. *Computer Methods and Programs in Biomedicine*, 172:103–116, 2019. https://doi.org/10.1016/j.cmpb.2019.02.014.

[20] Ugarte, M.D., Adin, A., Goicoa, T., and Militino, A.F. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical Methods in Medical Research*, 23:507–530, 2014. https://doi.org/10.1177/0962280214527528.

[21] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. https://www.R-project.org/.

[22] Martins, T.G., Simpson, D., Lindgren, F., and Rue, H. Bayesian computing with INLA: new features. *Computational Statistics and Data Analysis*, 2013. https://doi.org/10.1016/j.csda.2013.04.014.