Máster en Modelización Matemática, Estadística y Computación

Trabajo fin de máster

Análisis espacial de riesgos de mortalidad por cáncer de próstata con INLA

Universidad de Zaragoza

Autora: Paula Navarro Esteban
Directora: María Dolores Ugarte Martínez
(Universidad Pública de Navarra)

3 de septiembre de 2012

Agradecimientos

En primer lugar, me gustaría expresar mi más sincero agradecimiento a mi directora María Dolores Ugarte Martínez por darme la oportunidad de trabajar con ella y por la gran cantidad de horas que me ha dedicado.

Me gustaría también darle las gracias a mis padres, a mi hermana y a mis amigos, por estar a mi lado apoyándome cuando más lo necesitaba.

Índice general

Indice de tablas

Índice de figuras

Prólogo

La estimación de riesgos de mortalidad por distintas causas es fundamental en salud pública, ya que generalmente existen diferencias regionales en la mortalidad y en la incidencia de enfermedades. Por lo tanto es crucial localizar las zonas o regiones que presenten riesgos extremos y tratar de corregir esas desigualdades mediante la aplicación de planes de salud adecuados.

A causa de la enorme variabilidad de los riesgos en regiones con pocos casos de mortalidad o poco pobladas, las medidas clásicas de estimación de riesgos como la razón de mortalidad estandarizada por edad no son fiables. Por ello se emplean modelos estadísticos más complejos que "suavizan" los riesgos. En este trabajo se utilizará el modelo de Besag, York, and Mollié (1991) que estima el método INLA (Integrated Nested Laplace Approximation) usando como herramienta el software libre "R" [1]. El modelo se aplicará al estudio de mortalidad por cáncer de próstata en las provincias españolas en dos periodos distintos, 1975-1991 y 1992-2008 [25].

Introducción

La estimación de riesgos de mortalidad es trascendental en salud pública. La existencia de desigualdades regionales en el ámbito de la mortalidad y la incidencia de enfermedades es un problema social conocido. Es por ello muy importante detectar las zonas o regiones que manifiestan riesgos extremos ya que hoy en día, existe evidencia suficiente que demuestra que las desigualdades en salud son evitables pues pueden reducirse mediante políticas públicas sanitarias y sociales y planes de salud adecuados.

La principal finalidad de este trabajo es estimar riesgos de mortalidad por cáncer de próstata en España mediante modelos de suavizado en dos periodos distintos: 1975-1991 y 1992-2008.

El cáncer de próstata es una enfermedad por la que se forman células malignas (cancerosas) en los tejidos de la próstata. Se ha visto que este cáncer es el cáncer más frecuente entre hombres que viven en España y se encuentra principalmente en hombres de edad avanzada, ya que a medida que los hombres envejecen, la próstata se puede agrandar y bloquear la uretra o la vejiga. Según la Asociación Española Contra el Cáncer (AECC) [31], se trata de un tumor con altísima incidencia, sobre todo en países desarrollados, pero con una mortalidad moderada, siendo la tercera causa de muerte por cáncer en el sexo masculino, tras el cáncer de pulmón y el colorrectal.

En general se puede representar la mortalidad de cualquier enfermedad usando mapas que muestren el patrón geográfico de mortalidad en determinadas regiones. Esto puede ser útil para epidemiólogos o investigadores del campo de la salud pública, ya que a partir de ellos pueden formular hipótesis referentes a la etiología de la enfermedad estudiada o asesorar sobre la distribución de fondos destinados a la salud. No obstante, los mapas sin tratar o las medidas directas no son muy creíbles sobre todo cuando se trata de una enfermedad poco común, con pocos casos detectados o se tiene una población muy pequeña. Por eso, hay que emplear métodos de suavizado, que consisten en recoger información de todas las áreas para hacer que las estimaciones de las tasas sean más estables.

Para estimar los riesgos de mortalidad por cáncer de próstata se van a emplear medidas clásicas de estimación de riesgos, como la conocida razón de mortalidad estandarizada por edad (RME o SMR en inglés, que se define como el número de casos totales observados de mortalidad en la

población de estudio entre el número de casos esperados o número de casos considerando que la población de estudio tiene las mismas tasas que la población estándar). Sin embargo, se sabe que esta medida es muy variable en las regiones que están poco pobladas o tienen pocos casos de mortalidad, en estos casos diremos que estamos trabajando con áreas pequeñas. Es decir, que entenderemos por áreas pequeñas aquellas regiones en estudio cuyo tamaño muestral supone en general un problema para la estimación de los parámetros de interés.

Para solventar estos inconvenientes se usan modelos estadísticos sofisticados que "suavizan" los riesgos. Uno de los más conocidos es el de Besag, York y Mollié (1991), que denotaremos por BYM. Este modelo es un modelo GMRF (en inglés "Gaussian Markov Random Field") que estima de manera aproximada el método INLA que, en inglés, se denomina "Integrated Nested Laplace Approximation".

En este trabajo se pretenden abordar los siguientes objetivos:

1. Estudiar detalladamente las medidas clásicas de estimación de riesgos, especialmente la razón de mortalidad estandarizada por edad, cuya principal ventaja es que involucra sólo el número total de casos de mortalidad, de modo que no es preciso saber en qué edad ocurren los casos de mortalidad en la población de estudio. Otra ventaja práctica es que la razón de mortalidad estandarizada no tiende a ser sensible a las inestabilidades numéricas en uno o dos índices específicos de la edad.

2. Investigar la potencialidad del método INLA para suavizar riesgos de mortalidad con el modelo de BYM. Este novedoso método es una alternativa computacionalmente más rápida que los métodos de MCMC (en inglés, Markov Chain Monte Carlo).

3. Aplicar los resultados obtenidos en 1 y 2 al análisis de datos reales de mortalidad por cáncer de próstata en España en dos periodos de tiempo: 1975-1991 y 1992-2008, y detectar a su vez, provincias cuyo riesgo de morir por cáncer de próstata es mayor que el riesgo global de España.

Para el análisis de los datos usaremos **R** (ver por ejemplo [3]), que es un software libre que permite realizar análisis estadísticos y gráficos, y en particular destacamos la utilización de la librería llamada *INLA*, ver [2], que dispone de todas las herramientas necesarias.

Este trabajo se estructura en cinco capítulos. En el primer capítulo se incluye información básica sobre el cáncer de próstata, qué es, cómo se diagnostica y cuáles son sus factores de riesgo. En el segundo capítulo, se calcula la razón de mortalidad estandarizada por edad para los datos de mortalidad por cáncer de próstata en España. Este capítulo debería estar previo a las conclusiones, pero se ha colocado antes para mostrar el orden real del trabajo realizado. En el tercer capítulo se profundiza sobre la razón de mortalidad estandarizada. En el cuarto capítulo, que es el núcleo de este trabajo, se explica el modelo de BYM y cómo el método

INLA lo aproxima. Estos conceptos se ilustran con la aplicación a los datos de mortalidad por cáncer de próstata. En este capítulo también se comparan los resultados obtenidos aplicando la estadística bayesiana con los obtenidos aplicando la estadística frecuentista. Por último, se recogen las conclusiones en el quinto capítulo.

Capítulo 1

Información general sobre el cáncer de próstata

Antes de profundizar en detalles matemáticos veamos qué es el cáncer de próstata y cómo se detecta.

1.1. ¿ Qué es el cáncer de próstata?

El cáncer de próstata es una enfermedad en la cual se forman células malignas en los tejidos de la próstata. Ésta es una glándula que pertenece al sistema reproductor masculino localizada justo por debajo de la vejiga y por delante del recto, tal y como se muestra en la Figura (1.1). Su tamaño es como el de una nuez y rodea una parte de la uretra. La glándula prostática produce un fluido que forma parte del semen.

El cáncer de próstata es el tercer tumor más frecuente en varones españoles y supone la tercera causa de muerte por cáncer en España [4]. Por ejemplo, en el año 2005 existía una tasa cruda de 25.89 casos por 100000 habitantes, en dicho año fallecieron 5500 personas por este cáncer [5].

Este cáncer constituye aproximadamente el 11 % de todas las neoplasias (cánceres) en los varones de Europa, y es el responsable del 9 % de las muertes por cáncer entre los hombres dentro de la Unión Europea [6].

Su incidencia (número de casos nuevos de cáncer que se diagnostican cada año) es muy variable en las distintas zonas del mundo pudiendo tener esto relación con distintos factores como son: exposición ambiental, dieta, estilo de vida y factores genéticos [7].

La probabilidad de padecer un cáncer de próstata se incrementa con la edad y en el 90 % de

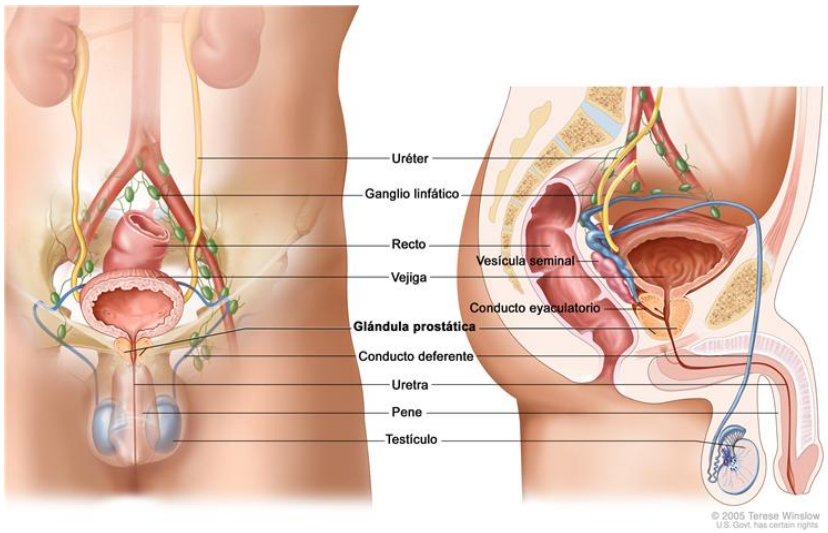

Figura 1.1: Anatomía del sistema reproductor masculino

los casos se produce en hombres de más de 65 años. Esto es debido a que a medida que los hombres envejecen, la próstata se puede agrandar y bloquear la vejiga o la uretra tal y como se muestra en la Figura (1.2). Esto puede ocasionar dificultades para orinar o interferir con la función sexual. La afección se llama hiperplasia prostática benigna (HPB).

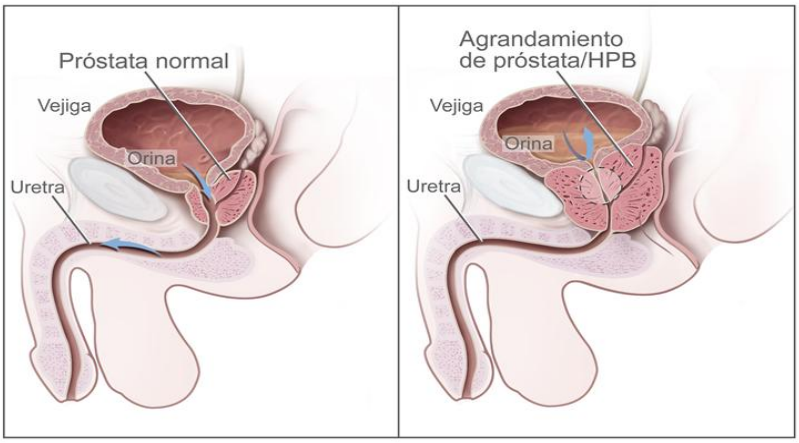

Figura 1.2: Próstata normal y HPB

En la raza negra existe un riesgo mayor de padecerlo y también se incrementa de forma importante en personas con parientes de primer grado afectados pudiendo llegar a ser siete veces superior en personas con dos o tres familiares de primer grado afectados [8].

La incidencia del cáncer de próstata está aumentando en países desarrollados entre otras causas por el aumento de la esperanza de vida y principalmente por el desarrollo de nuevas técnicas

diagnósticas, ver apartado 1.2.

Desde el punto de vista anatomo-patológico la mayor parte de los cánceres de próstata son adenocarcinomas, es decir, aparecen en las células glandulares que revisten órganos internos siendo frecuentemente de localización multifocal y de predominio en la zona periférica de la glándula prostática [9].

Algunos de los síntomas ocasionados por el cáncer de la próstata pueden ser:

- Disminución del calibre o interrupción del flujo urinario.
- Aumento de la frecuencia de la micción (especialmente por la noche).
- Dificultad para orinar.
- Dolor o ardor durante la micción.
- Presencia de sangre en la orina o en el semen.
- Dolor constante en la espalda, las caderas o la pelvis.
- Eyaculación dolorosa.

1.2. ¿Cómo se diagnostica?

Según [4] se pueden utilizar las siguientes pruebas y procedimientos:

- *Examen digital del recto (EDR)*: examen del recto. El médico o el enfermero inserta un dedo protegido por un guante lubricado en el recto y palpa la próstata a través de la pared del recto en busca de bultos o áreas anormales.
- *Prueba del antígeno prostático específico (APE)*: prueba de laboratorio que mide las concentraciones del APE en la sangre. El APE es una sustancia elaborada por la próstata que se puede encontrar en una mayor cantidad en la sangre de los hombres que tienen cáncer de próstata. La concentración de APE también puede ser elevada en los hombres que sufren una infección o una inflamación de la próstata, o que tienen HPB (próstata agrandada, pero no cancerosa). Esta prueba junto con la anterior son las más utilizadas.
- *Ecografía transrectal*: procedimiento por el cual se inserta en el recto una sonda que tiene aproximadamente el tamaño de un dedo para examinar la próstata. La sonda se utiliza para hacer rebotar ultrasonidos (ondas de sonido de alta energía) en los tejidos internos de la próstata y crear ecos. Los ecos forman una imagen de los tejidos corporales que se llama sonograma. La ecografía transrectal se puede usar durante una biopsia.

- *Biopsia*: extracción de células o tejidos realizada por un patólogo para observarlos bajo un microscopio. El patólogo examina la muestra en busca de células cancerosas. Hay dos tipos de biopsia utilizados para diagnosticar el cáncer de próstata.
    - *Biopsia transrectal*: se extrae tejido de la próstata mediante la inserción de una aguja fina a través del recto hasta la próstata. Este procedimiento se suele realizar mediante ecografía transrectal para ayudar a guiar la aguja.
    - *Biopsia transperineal*: se extrae una muestra de tejido prostático mediante la inserción de una aguja fina a través de la piel entre el escroto y el recto hasta la próstata.

1.3. Factores de riesgo

Un factor de riesgo es toda circunstancia o situación que afecta la posibilidad de tener una enfermedad, en nuestro caso el cáncer de próstata. Hay factores de riesgo que se pueden modificar, como beber alcohol, y otros, como la edad, que no se pueden cambiar. El tenerlos no es una condición necesaria ni suficiente para padecer la enfermedad, es decir, hay personas con uno o más factores de riesgo que nunca padecen cáncer, mientras que otras que lo padecen puede que hayan tenido pocos factores de riesgo conocidos o ninguno [10].

Todavía no se entiende completamente las causas del cáncer de próstata, pero se han encontrado varios factores que pueden cambiar el riesgo de padecer esta enfermedad. Son los siguientes:

- *Edad.* La probabilidad de tener cáncer de próstata aumenta rápidamente después de los 50 años, se conocen muy pocos casos antes de los 40. Casi dos de cada tres casos de cáncer de próstata se detectan en hombres mayores de 65 años.
- *Raza.* Los hombres de raza negra es el grupo étnico que con más frecuencia padece cáncer de próstata respecto a otras razas. Además éstos tienen una mayor probabilidad de ser diagnosticados en una etapa avanzada, y tienen más del doble de probabilidad de morir de cáncer de próstata en comparación con los hombres blancos. Al contrario, los hombres asiático-americanos y los hispanos o latinos tienen una menor frecuencia de padecer cáncer de próstata respecto a los hombres blancos. El porqué de estas diferencias no está claro todavía.
- *Nacionalidad.* El cáncer de próstata es más común en Norteamérica y en la región noroeste de Europa, Australia, y en las islas del Caribe. Es menos común en Asia, Centroamérica y Sudamérica. Estas diferencias tampoco están claras, es probable que se deba a los distintos estilos de vida o a que en los países desarrollados se usan más pruebas de detección.
- *Antecedentes familiares.* Se sabe que se duplica el riesgo de que un hombre padezca el cáncer si su padre o su hermano lo padecen también (siendo mayor el riesgo para los hombres que

tienen un hermano con la enfermedad que para aquellos con un padre afectado por este cáncer). Asimismo, el riesgo es mucho mayor en el caso de los hombres que tienen varios familiares afectados, especialmente si tales familiares eran jóvenes en el momento en el que se les diagnosticó el cáncer.

- *Otras causas.* Para factores como la alimentación, la obesidad, el tabaquismo, la inflamación de la próstata, las infecciones de transmisión sexual o la vasectomía, se han realizado estudios pero no se han encontrado conclusiones claras.

Capítulo 2

Creación de la base de datos

En este capítulo crearemos las bases de datos a partir del archivo "prostata.txt" (cedido por el Centro Nacional de Epidemiología), que recoge los casos de mortalidad por cáncer de próstata en cada provincia de España desde el año 1975 hasta 2008. Para nuestro caso vamos a considerar dos periodos de estudio, 1975-1991 y 1992-2008.

En primer lugar calcularemos los valores o muertes esperadas (mediante estandarización indirecta) según la mortalidad por cáncer de próstata en hombres, describiendo la sintaxis del software **R**. También calcularemos las tasas de mortalidad de referencia o tasa global de mortalidad en España, y finalmente obtendremos la Razón de Mortalidad Estandarizada (RME_i) o "Standardized Mortality Ratio" (SMR_i), en inglés, que definiremos matemáticamente más adelante.

Para entender los cálculos mencionados vamos a definir primeramente una serie de conceptos.

2.1. Definiciones básicas

2.1.1. Salud pública

La salud pública es la disciplina encargada de la protección de la salud a nivel poblacional. De ella se encargan especialistas como médicos, biólogos, estadísticos, veterinarios, etc, aunque su desarrollo depende de los gobiernos. Su objetivo es mejorar las condiciones de salud de las comunidades mediante la promoción de estilos de vida saludables, las campañas de concienciación, la educación y la investigación.

Los organismos de la salud pública deben evaluar las necesidades de salud de la población, los riesgos para la salud y analizar las causas de dichos riesgos. De acuerdo a lo detectado,

deben establecer las prioridades y desarrollar programas y planes que permitan responder a las necesidades de la población.

La salud pública también debe gestionar de manera óptima los recursos para asegurar que sus servicios lleguen a la mayor cantidad de gente posible, ya que parte de unos principios comunitarios y no personales.

2.1.2. Registro de cáncer

Una de las mayores necesidades de la salud pública es conocer, lo mejor posible, las características de salud y enfermedad de la población. Esta información es un instrumento fundamental para la planificación de las actuaciones en el ámbito de la salud en general y de la salud pública en particular.

Para ello se ha creado el sistema de registro de mortalidad, que es una fuente de datos exhaustiva, representativa tanto a nivel nacional como internacional, y que ha sido recopilada durante un largo periodo de tiempo. Todas las muertes ocurridas en una zona o país quedan recogidas en este sistema, junto con información de la persona, el tiempo, el lugar y la causa de la defunción.

Gracias a este registro se pueden construir indicadores de mortalidad que pueden ayudar a detectar problemas de salud importantes así como desigualdades entre regiones. No obstante, el objetivo más efectivo de los programas de salud pública es prevenir la causa que da origen a todos los demás trastornos o afecciones que conducen a la muerte. Con esa base se diseñó el Modelo Internacional de Certificado Médico de Causa de Defunción, en el que está basado el diseño del Boletín Estadístico de Defunción (B.E.D.).

Fue en 1967 cuando la Asamblea Mundial de la Salud estableció que las causas de defunción (definidas como "todas aquellas enfermedades, estados morbosos o lesiones que produjeron la muerte o contribuyeron a ella y las circunstancias del accidente o de la violencia que produjo dichas lesiones") fueran registradas en el certificado médico de causas de defunción. Más tarde se implantó la Clasificación Internacional de Enfermedades (CIE), por la cual las causas de muerte eran codificadas de una forma homogénea, una vez inscritas. De este modo se obtenía una fuente de datos universal.

2.2. Instalación y carga de librerías

Debemos definir un directorio de trabajo (en este caso una carpeta llamada "tfm" en la unidad D:/) donde se guardarán las bases de datos que creemos con ayuda del software **R**.

```
> setwd("D:/Paula/master/TFM/tfm")
```

Tras su instalación en **R**, cargamos las librerías necesarias con las siguientes instrucciones. (Nota: se ha escrito a continuación de cada librería una breve descripción.)

```
> library(plyr) # plyr es un conjunto de herramientas cuya finalidad es resolver
>               # un problema. Consiste en dividir un problema grande en varios
>               # más pequeños y manejables, para luego volverlos a unir y poder
>               # asi resolver el problema principal.
>
> library(reshape) # reshape facilita la reestructuración y la agregación de
>                  # datos usando solamente dos funciones: melt y cast.
>
> library(INLA) # INLA contiene funciones que permiten trabajar con modelos
>               # aditivos completamente bayesianos usando Integrated Nested
>               # Laplace Approximation.
>
> library(maptools) # maptools es un conjunto de herramientas que manipulan y
>                   # leen datos geográficos, en particular archivos ESRI.
>
> library(spdep) # spdep es una colección de funciones que sirven para crear
>                # matrices espaciales a partir de polígonos contiguos y de
>                # funciones que estiman intervalos y errores de modelos
>                # espaciales simultáneos autorregresivos (SAR) y de modelos
>                # espaciales de regresión (CAR), entre otros.
>
> library(RColorBrewer) # RColorBrewer proporciona paletas para colorear mapas
>                       # de acuerdo a una variable.
>
> library(Hmisc) # Hmisc contiene numerosas funciones útiles para el análisis de
>                # datos, gráficos de alto nivel, funciones para el cálculo del
>                # tamaño de una muestra, importar conjuntos de datos,etc.
>
> library(gplots) # gplots consta de varias herramientas para la representación
>                 # gráfica de los datos.
>
> library(pixmap) # pixmap contiene funciones para importar, exportar,
>                 # representar gráficamente y otras manipulaciones de imágenes
>                 # pixeladas.
>
> library(mvtnorm) # mvtnorm calcula probabilidades de normales multivariantes
>                  # y de t de student, así como cuantiles y densidades.
```

```
>
> library(numDeriv) # numDeriv ofrece métodos para calcular primeras y segundas
>                   # derivadas de forma numérica y precisa. La precisión de los
>                   # cálculos se realiza usando la extrapolación de Richardson,
>                   # aunque también se da el método de la diferencia simple.
>
> library(fields) # fields es para curvas, superficies, funciones como splines,
>                 # datos espaciales y estadística espacial.
>
> library(R.utils) # R.utils proporciona clases y métodos útiles cuando
>                  # programamos en R y desarrollamos paquetes de R.
```

2.3. Lectura de los datos

Cargamos el archivo que contiene los datos de mortalidad de cáncer de próstata en la población de España. Para facilitar el manejo de los datos se ha modificado dicho archivo, de tal manera que se ha introducido la variable "id" que indica de manera única a cada grupo de edad en cada año.

```
> mort<-read.table("prostata.txt",header=T)
```

Si observamos los datos vemos que hay 8 variables:

- **id**. Código de 5 cifras que identifica de manera única a cada grupo de edad en cada año. Por defecto esta variable no tenía ningún nombre asignado, así que se le ha dado éste.
- **prov**. Variable numérica que toma valores desde 1 hasta 50 e indica, por orden alfabético, la provincia a la que pertenece el grupo de edad en cada año. Es decir, que el 1 se identifica con Álava, el 2 con Albacete, el 3 con Alicante y así sucesivamente. Notar que se están considerando 50 provincias españolas (luego Ceuta y Melilla no se incluyen en este estudio).
- **provincia**. Variable con el nombre de la provincia.
- **sex**. Variable que toma el valor 1 en todos los casos. Esto indica que sólo hay hombres en la muestra.
- **edadgr**. Variable que indica el grupo de edad al que pertenece. Toma valores desde 1 hasta 18, es decir, que tenemos 18 grupos de edad, que por convenio el grupo 1 está formado por las personas de 0 a 4 años, el grupo 2 por las de 5 a 9 años, y así sucesivamente hasta el grupo 18 que está formado por las personas de más de 85 años. Ver [27].

- **casos**. Número de casos de mortalidad por cáncer de próstata en cada grupo.

- **pob**. Número de habitantes de cada grupo de edad en un año determinado (desde 1975 hasta 2008).

- **agno**. Año en que se recogieron los datos. Toma valores desde 1975 hasta 2008.

2.4. Cálculo de valores esperados

En este apartado calcularemos las tasas de referencia y los valores esperados de los casos de mortalidad por cáncer de próstata. Por convenio, las tasas de mortalidad por cáncer se expresan como casos por 100000 persona-años para evitar el uso de muchas cifras decimales. La expresión persona-años se usa para hacer referencia a que los datos se han recogido durante varios años, esto se debe a que en la práctica el cáncer es un fenómeno relativamente raro y para estudiarlo hay que observar poblaciones bastante grandes durante periodos de varios años.

Como nos interesa estudiar una causa de muerte específica tenemos que usar tasas de mortalidad estandarizadas por edad, ya que si comparamos tasas de mortalidad para un determinado tumor entre dos poblaciones diferentes, o en la misma población con el paso del tiempo se plantean problemas. Por ejemplo, la comparación de simples tasas brutas (definidas como el cociente del número total de casos entre el número total de persona-años de observación y multiplicado por 100000) puede llevar a dar conclusiones erróneas debido a las diferencias en la estructura de edad de las poblaciones a comparar. Si una población tiene una edad media menor que la otra, aunque las tasas específicas por edad (definidas como el cociente de el número de casos en un grupo de edad determinado entre los correspondientes personas-años de observación y multiplicados por 100000) sean las mismas en ambas poblaciones, aparecerán más casos en la población envejecida que en la más joven.

De este modo, al estudiar patrones de cáncer en una misma área a lo largo del tiempo hay que tener en cuenta si la estructura de edad de la población es distinta o está cambiando. Esto se consigue gracias a la estandarización por edad.

Según [29] existen dos métodos para dicha estandarización:

- Método de estandarización directa. Se define la tasa de mortalidad estandarizada por edad (TMEE) como la tasa que debería haber sido observada en una población con la misma estructura de edad que la de la población de referencia, llamada *población estándar*. Se calcula como una media ponderada de la tasa de mortalidad específica por edad (que es el número de muertes por cada 100000 personas restringido a cada grupo de edad) donde los pesos son la población estándar de cada grupo de edad, es decir,

Definición 2.4.1. *Definimos la tasa de mortalidad estandarizada por edad (TMEE) como*

$$TMEE = \frac{\sum_{j=1}^{J} a_j p_j}{\sum_{j=1}^{J} p_j}$$

donde a_j es la tasa específica por edad del j-ésimo grupo de edad, p_j es la población del j-ésimo grupo de edad de la población estándar y J es el número de grupos de edad que tenemos (en nuestro caso $J = 18$).

Se pueden comparar dos tasas de mortalidad estandarizada por edad si tienen la misma población estándar.

- Método de estandarización indirecta. La razón de mortalidad estandarizada (**SMR o RME**) se calcula como el cociente del número total de casos de muertes en la población de estudio entre el número esperado de casos. El número esperado de casos se obtiene aplicando las tasas de edad de la población estándar a la estructura de la población de estudio, es decir, calculamos los casos esperados considerando que la población de estudio tiene las mismas tasas que la población estándar.

  La principal ventaja de la SMR es que sólo involucra el número total de muertes, así que no es necesario conocer a qué edad ocurren las muertes en la población de estudio.

  Respecto a los inconvenientes, podemos comentar que no es apropiado calcular la razón de mortalidad estandarizada por edad como una medida de mortalidad única para una cierta región en un cierto periodo. Ya que si lo hiciéramos así estaríamos concentrando en una única medida toda la información correspondiente a la mortalidad, y como consecuencia, perderíamos gran cantidad de información. Por ejemplo, para unas determinadas enfermedades, uno esperaría que los efectos de la contaminación en una región tuvieran diferentes consecuencias en distintos grupos de edad. Por tanto, para cualquier análisis de datos de mortalidad las razones específicas por edad deberían ser siempre el punto de comienzo [29].

  Definición 2.4.2. *La razón de mortalidad estandarizada por edad interna para cada provincia se define como*

$$RME_i = \frac{y_i}{e_i}, \ \textit{con } i = 1, \ldots, 50$$

donde y_i es el n^o de casos de mortalidad por una causa determinada (cáncer de próstata) en hombres en la provincia i-ésima y e_i es el número de casos esperados que vienen definidos como

$$e_i = \sum_{j=1}^{18} n_{ij} R_j, \ \textit{siendo } R_j = (n^o \textit{ de muertes en el grupo j})/(\textit{población en el grupo j}),$$

es decir, R_j es la tasa global de mortalidad en España en el grupo de edad j y n_{ij} es la población en el grupo de edad j en la provincia i.

En la definición se ha usado la palabra "interna", con ello se pretende resaltar que cada provincia se compara con España en global y no se puede comparar provincias entre sí.

En el capítulo 3, mostraremos que la RME_i es el estimador máximo verosímil del riesgo de mortalidad (r_i) cuando suponemos que $y_i \sim \mathcal{P}(e_i r_i)$. Por tanto lo interesante es comprobar si la SMR es mayor que 1, es decir, si $y_i > e_i$, lo que significaría que el número de casos observados en la provincia i-ésima es mayor que el número esperado de casos en dicha provincia, entendiendo por el número de casos esperados como el número de casos que obtendríamos si la provincia i-ésima tuviera un comportamiento como el global de España. En otras palabras, si r_i>1 entonces la provincia i-ésima presenta un exceso de mortalidad comparado con la mortalidad de España. Mientras que si r_i<1, entonces la provincia i-ésima presenta un defecto de riesgo comparado con el riesgo global de España.

Frecuentemente las SMR se presentan multiplicándolas por 100 y en este trabajo se considerarán de esa forma, por tanto lo que en realidad comprobaremos será si r_i>100 o si r_i<100.

En general no se pueden comparar dos SMR de diferentes poblaciones.

2.4.1. Tasas de referencia o tasa global de mortalidad en España

Las tasas específicas de referencia las calcularemos teniendo en cuenta el propio país, en este caso España (esto se conoce como estandarización interna). Como sólo hemos recogido los datos de mortalidad por cáncer de próstata, sólo tendremos una causa y por lo tanto las tasas de mortalidad estarán contenidas en un vector de dimensión 18.

Para el análisis estadístico utilizaremos tasas específicas de referencia diferentes para cada periodo. Por tanto necesitamos definir la variable "periodo". Como ya habíamos comentado en la introducción, consideraremos dos periodos, el primero comprende los años desde 1975 hasta 1991, y el segundo desde 1992 hasta 2008.

```
> mort$periodo <-ifelse(mort$agno >= 1975 & mort$agno < 1992,1,2)
```

Además definimos 18 variables que corresponden a cada grupo de edad (recordar que estos grupos los hemos explicado en el apartado 2.2). Para ello, hacemos lo siguiente en R:

```
> mort$edad1 <- ifelse(mort$edadgr==1,mort$casos,0)
> mort$edad2 <- ifelse(mort$edadgr==2,mort$casos,0)
> mort$edad3 <- ifelse(mort$edadgr==3,mort$casos,0)
> mort$edad4 <- ifelse(mort$edadgr==4,mort$casos,0)
> mort$edad5 <- ifelse(mort$edadgr==5,mort$casos,0)
> mort$edad6 <- ifelse(mort$edadgr==6,mort$casos,0)
```

```
> mort$edad7 <- ifelse(mort$edadgr==7,mort$casos,0)
> mort$edad8 <- ifelse(mort$edadgr==8,mort$casos,0)
> mort$edad9 <- ifelse(mort$edadgr==9,mort$casos,0)
> mort$edad10 <- ifelse(mort$edadgr==10,mort$casos,0)
> mort$edad11 <- ifelse(mort$edadgr==11,mort$casos,0)
> mort$edad12 <- ifelse(mort$edadgr==12,mort$casos,0)
> mort$edad13 <- ifelse(mort$edadgr==13,mort$casos,0)
> mort$edad14 <- ifelse(mort$edadgr==14,mort$casos,0)
> mort$edad15 <- ifelse(mort$edadgr==15,mort$casos,0)
> mort$edad16 <- ifelse(mort$edadgr==16,mort$casos,0)
> mort$edad17 <- ifelse(mort$edadgr==17,mort$casos,0)
> mort$edad18 <- ifelse(mort$edadgr==18,mort$casos,0)
```

Calculamos el número de muertes del periodo para cada grupo de edad, es decir, los numeradores de cada elemento del vector:

```
> #renombramos la base de datos ya que sólo tenemos hombres
> hombres.ce <- mort[mort$sex==1,]
> #sumamos las muertes según grupos de edad por periodo
> hombres.ce<-aggregate(hombres.ce[,10:27],by=list(periodo=hombres.ce$periodo),
+                       FUN=sum,na.rm=T)
```

La opción *aggregate* divide el conjunto en subconjuntos. Observar que tenemos todas las combinaciones de periodo y grupo de edad. Ahora separamos los periodos 1 y 2, trasponemos para obtener una estructura parecida a nuestra matriz y nos quedamos con las filas que nos interesan

```
> hombres.ce.p1 <- hombres.ce[hombres.ce$periodo==1,] #seleccion periodo 1
> hombres.ce.p2 <- hombres.ce[hombres.ce$periodo==2,] #seleccion periodo 2
> #seleccionamos las columnas que nos interesan
> hombres.ce.p1 <- hombres.ce.p1[,2:19]
> hombres.ce.p2 <- hombres.ce.p2[,2:19]
> hombres.ce.p1 <- as.data.frame(t(hombres.ce.p1)) #trasponemos
> hombres.ce.p2 <- as.data.frame(t(hombres.ce.p2)) #trasponemos
> names(hombres.ce.p1) <- paste("total")
> #llamamos total al nº total de casos en cada grupo de edad en el periodo1
> names(hombres.ce.p2) <- paste("total")
> #llamamos total al nº total de casos en cada grupo de edad en el periodo2
```

Calculamos la población del periodo para cada grupo de edad, es decir, el denominador de

cada elemento del vector. Para ello, vamos a definir la variable *sección*, que no es mas que la variable *prov* del fichero original.

```
> ptempo_h<-mort
> ptempo_h$sex<-NULL #eliminamos la variable del sexo porque todos son hombres.
> pob.h <- ptempo_h
> seccion <- (pob.h$prov)
> pob.h <- data.frame(SECCION=seccion, pob.h)
```

Vamos a crear unas variables en el data frame que acabamos de crear para facilitar los cálculos. Dichas variables se llamarán *edadi* con $i = 1, \ldots, 18$ y contienen la población total según el grupo de edad.

```
> pob.h$edad1 <- ifelse(pob.h$edadgr==1,pob.h$pob,0)
> pob.h$edad2 <- ifelse(pob.h$edadgr==2,pob.h$pob,0)
> pob.h$edad3 <- ifelse(pob.h$edadgr==3,pob.h$pob,0)
> pob.h$edad4 <- ifelse(pob.h$edadgr==4,pob.h$pob,0)
> pob.h$edad5 <- ifelse(pob.h$edadgr==5,pob.h$pob,0)
> pob.h$edad6 <- ifelse(pob.h$edadgr==6,pob.h$pob,0)
> pob.h$edad7 <- ifelse(pob.h$edadgr==7,pob.h$pob,0)
> pob.h$edad8 <- ifelse(pob.h$edadgr==8,pob.h$pob,0)
> pob.h$edad9 <- ifelse(pob.h$edadgr==9,pob.h$pob,0)
> pob.h$edad10 <- ifelse(pob.h$edadgr==10,pob.h$pob,0)
> pob.h$edad11 <- ifelse(pob.h$edadgr==11,pob.h$pob,0)
> pob.h$edad12 <- ifelse(pob.h$edadgr==12,pob.h$pob,0)
> pob.h$edad13 <- ifelse(pob.h$edadgr==13,pob.h$pob,0)
> pob.h$edad14 <- ifelse(pob.h$edadgr==14,pob.h$pob,0)
> pob.h$edad15 <- ifelse(pob.h$edadgr==15,pob.h$pob,0)
> pob.h$edad16 <- ifelse(pob.h$edadgr==16,pob.h$pob,0)
> pob.h$edad17 <- ifelse(pob.h$edadgr==17,pob.h$pob,0)
> pob.h$edad18 <- ifelse(pob.h$edadgr==18,pob.h$pob,0)
```

Ahora sumamos la población según grupos de edad por periodo y provincia (en este apartado no utilizamos esta variable pero más tarde lo haremos).

```
> pob.h<-aggregate(pob.h[,10:27],list(periodo=pob.h$periodo,SECCION=pob.h$SECCION)
+                  ,FUN=sum)
```

Ahora (todos) a partir del objeto *pob.h* que hemos creado:

1975-1991		
	población	casos
edad1	23783116	3
edad2	26650153	0
edad3	28289365	3
edad4	27812155	10
edad5	25733004	14
edad6	23405963	11
edad7	21409388	12
edad8	19591498	23
edad9	18897184	46
edad10	18474795	170
edad11	18123387	527
edad12	16715265	1317
edad13	14107813	3276
edad14	11457767	6392
edad15	8747924	10793
edad16	5989069	13892
edad17	3335444	12847
edad18	1713562	9326

1992-2008		
	población	casos
edad1	17585302	0
edad2	17808776	0
edad3	19972259	0
edad4	23486777	3
edad5	27140860	5
edad6	29658756	3
edad7	29397065	8
edad8	27443972	11
edad9	24997207	47
edad10	22559250	154
edad11	20002762	568
edad12	18190253	1459
edad13	16964722	3555
edad14	15528149	7560
edad15	13273003	13217
edad16	9479828	18670
edad17	5670309	21230
edad18	3641927	25464

Tabla 2.1: Número de muertes y población total según grupos de edad para cada periodo

```
> pob.h.p1 <- pob.h[pob.h$periodo==1,] #separamos el primer periodo
> pob.h.p2 <- pob.h[pob.h$periodo==2,] #separamos el segundo periodo
> #calculamos la población total por grupos de edad
> hombres.ce.p1<-data.frame(poblacion=apply(pob.h.p1[,3:20],2,sum),hombres.ce.p1)
> hombres.ce.p2<-data.frame(poblacion=apply(pob.h.p2[,3:20],2,sum),hombres.ce.p2)
```

Por lo tanto para el periodo 1 y 2 respectivamente obtenemos la Tabla 2.1. A la vista de la Tabla 2.1 podemos comprobar que para cada periodo se verifica que aproximadamente el 90 % de los casos de cáncer de próstata se produce en hombres mayores de 65 años, es decir, a partir del grupo *edad14*, tal y como habíamos mencionado en el capítulo 1. Concretamente para el primer periodo el número de casos de cáncer de próstata en hombres mayores de 65 años representa un 90.77 % del total de casos en ese periodo y para el segundo periodo los hombres mayores de 65 años que padecen cáncer representan un 93.68 % del total de este mismo periodo.

Calculamos las tasas específicas por edad del país dividiendo los muertos por cáncer en cierto grupo de edad por la población en el grupo de edad.

```
> tasa.h.ce.p1 <- hombres.ce.p1[2]/hombres.ce.p1$poblacion
> tasa.h.ce.p2 <- hombres.ce.p2[2]/hombres.ce.p2$poblacion
```

2.4.2. Esperados

El número de muertes esperadas estarán contenidas en un vector de dimensión 18. Para calcular esta cantidad por estandarización indirecta debemos multiplicar la población de cada provincia por edad por las tasas de referencia en ese grupo de edad. Esto equivale a multiplicar la matriz de poblaciones, que contiene tantas filas como provincias y tantas columnas como grupos de edad por el vector de tasas calculado en el apartado anterior, esto es:

$$E_{ij} = \sum_j n_{ij} R_j$$

donde R_{ij} es la tasa específica por edad y n_{ij} es la población en la provincia i del grupo de edad j.

Hay que tener en cuenta que habrá 4 vectores de este tipo: los esperados para el periodo 1 tomando las tasas de referencia del periodo 1, los esperados para el periodo 1 tomando las tasas de referencia del periodo 2, los esperados para el periodo 2 tomando las tasas de referencia del periodo 1 y los esperados para el periodo 2 tomando las tasas de referencia del periodo 2. En la sintaxis que se muestra de R, los esperados tomando como referencia el periodo 1 están guardados en el objeto "esperados.p1" y los esperados tomando como referencia el periodo 2 en el objeto "esperados.p2". Se puede acceder al periodo en que se esperan a través de la variable *periodo*.

Pasamos a ver cómo se calculan los esperados. Vamos a ordenar las poblaciones por periodo y provincia para que nos quede así en los esperados.

```
> pob.h <- pob.h[order(pob.h$periodo, pob.h$SECCION),]
```

Multiplicamos las tasas específicas por las poblaciones, con esto calculamos el denominador de las RMEs, es decir, los esperados.

```
> #Esperados tomando como referencia las tasas del primer periodo
> esperados.p1 <- data.matrix(pob.h[,3:20])%*%data.matrix(tasa.h.ce.p1)
> esperados.p1 <- data.frame(periodo=pob.h$periodo,SECCION=pob.h$SECCION,
+                            esperados.p1)
> #Esperados tomando como referencia las tasas del segundo periodo
> esperados.p2 <- data.matrix(pob.h[,3:20])%*%data.matrix(tasa.h.ce.p2)
> esperados.p2 <- data.frame(periodo=pob.h$periodo,SECCION=pob.h$SECCION,
+                            esperados.p2)
```

Si pidiéramos los valores extremos de los esperados obtendríamos:

```
Tomando como referencia el primer periodo, el mínimo de los esperados es:
301.9957

Tomando como referencia el primer periodo, el máximo de los esperados es:
10804.71

Tomando como referencia el segundo periodo, el mínimo de los esperados es:
294.8144

Tomando como referencia el segundo periodo, el máximo de los esperados es:
10421.61
```

Luego los esperados son muy variables, con rangos 301.9957-10804.71, tomando como referencia el primer periodo, y 294.8144-10421.61, tomando como referencia el segundo periodo. Esta variabilidad se verá reflejada cuando calculemos las SMRs. Obtendremos SMRs altas en las provincias escasamente pobladas o en provincias donde el cáncer de próstata sea una enfermedad muy rara y haya muy pocos casos. Por eso tenemos que usar métodos de suavizado de los riesgos de mortalidad que usen la información de todas las provincias y así las estimaciones serán más fiables y los estimadores más estables.

2.4.3. Observados

Los observados estarán contenidos en una vector de igual estructura que los esperados pero en cada elemento deberemos tener los muertos observados por sección (provincia). Vamos a calcular los observados en cada uno de los periodos.

Creamos la variable "sección" (análogamente a lo anterior) y la incluimos en la base de datos de "mort".

```
> seccion <- mort$prov
> mort.h <- data.frame(mort[9], SECCION=seccion, mort[1:7],mort[10:27])
> #Notar que aqui no estamos cogiendo el año
```

Sumamos las muertes por sección (provincia) y periodo.

```
> mort.agr<-aggregate(data.frame(casos=mort.h$casos),
+                     list(periodo=mort.h$periodo,SECCION=mort.h$SECCION),
+                     FUN=sum)
```

Las ordenamos y las renombramos por "observados".

```
> observados <- mort.agr[order(mort.agr$periodo,mort.agr$SECCION),]
```

Guardamos todos los objetos que hemos ido creando.

```
> hombres_ce <- list(observados=observados, esperados.p1=esperados.p1,
+                    esperados.p2=esperados.p2)
> save(hombres_ce,file="espana_interna.RData")
```

Capítulo 3

Medidas clásicas de estimación de riesgos: razón de mortalidad estandarizada

La aproximación clásica asume que para el estudio de datos de mortalidad, el número de muertes en cada provincia y_i, sigue una distribución de Poisson. Estas distribuciones son independientes y vienen definidas como:

$$y_i \sim \mathcal{P}(\lambda_i = e_i r_i), \text{ para } i = 1, \ldots, 50 \tag{3.1}$$

donde e_i es el número esperado de muertes en la provincia i-ésima (que es conocido) y r_i es el riesgo relativo asociado a la provincia i-ésima (que es desconocido).

Para estimar el parámetro r_i que aparece en la distribución de probabilidad y_i, usaremos el método de estimación por máxima verosimilitud que elige como estimador a aquél valor del parámetro que tiene la propiedad de maximizar la probabilidad de la muestra aleatoria observada [13].

Definición 3.0.3. *Sea X una variable aleatoria discreta con función de probabilidad denotada por $p(X;\theta)$, es decir, que depende del parámetro θ y de la m.a.s. de tamaño n $X = (x_1, \ldots, x_n)$. La función de probabilidad conjunta viene dada por*

$$p(X;\theta) = \prod_{i=1}^{n} p(x_i;\theta).$$

*Cuando θ es conocido, esta función determina la probabilidad de aparición de cada muestra, sin embargo en un problema de estimación se conoce el valor de la muestra pero no el del parámetro. Luego se puede ver esta función de probabilidad como una función del parámetro desconocido. A esta nueva función se le llama **función de verosimilitud** y se denota por $L(\theta|X)$, donde $L(\theta|X) = p(X;\theta)$.*

Es decir, que la verosimilitud es una función que describe la dependencia de un/os parámetro/s en los valores de la muestra.

Al valor del parámetro que maximiza esta función se le llama ***estimador de máxima verosimilitud*** *o estimador MV y lo denotaremos por* $\hat{\theta}$.

En nuestro caso el parámetro desconocido es r_i y denotaremos a su estimador MV como $\hat{r}_i$.

Como los casos observados, y_i, siguen una distribución de Poisson de parámetro $\lambda_i = e_i r_i$, su función de probabilidad será:

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \text{ para } i = 1, \ldots, 50.$$

Además, suponemos que los y_i's son independientes y por tanto la función de probabilidad conjunta será el producto de las funciones de probabilidad marginales, es decir,

$$L(r_i|y_i) = \prod_{i=1}^{50} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} = \prod_{i=1}^{50} \frac{e^{-e_i r_i}(e_i r_i)^{y_i}}{y_i!}, \tag{3.2}$$

donde hemos tomado $n = 50$, ya que es el número de provincias que vamos a considerar.

Como el logaritmo es una función monótona creciente, maximizar L es lo mismo que maximizar $\log(L)$, y esto último es más sencillo. Por lo tanto, aplicando las propiedades del logaritmo a (3.2) obtenemos:

$$\log L(r_i|y_i) = \sum_{i=1}^{50} \left(-e_i r_i + y_i \log e_i r_i - \log y_i!\right).$$

Ahora bien, los posibles candidatos a estimadores MV son los valores que resuelven

$$\frac{\partial \log L}{\partial r_i} = 0.$$

Es decir,

$$\frac{\partial \log L}{\partial r_i} = -r_i + \frac{y_i e_i}{e_i r_i} = -r_i + \frac{y_i}{r_i} = 0.$$

Por tanto,

$$y_i = r_i e_i \Leftrightarrow \hat{r}_i = \frac{y_i}{e_i}.$$

Es importante notar que el $\hat{r}_i$ calculado es candidato a estimador MV, ya que el que la primera derivada sea igual a 0 es sólo una condición necesaria pero no suficiente para obtener un máximo. Por lo tanto vamos a calcular la segunda derivada:

$$\left.\frac{\partial^2 \log L}{\partial r_i^2}\right|_{r_i = \hat{r}_i} = -\frac{e_i^2}{y_i^2} < 0.$$

Por lo tanto $\hat{r}_i = \frac{y_i}{e_i}$ es el estimador MV de r_i que se denomina SMR_i.

Veamos cuál es la varianza de este estimador MV. Para ello tomamos varianzas en la igualdad anterior usando que e_i es constante y las propiedades de la varianza :

$$Var(\hat{r_i}) = Var\left(\frac{y_i}{e_i}\right) = \frac{Var(y_i)}{e_i^2}.$$

Como las y_i siguen una distribución de Poisson, su varianza es igual a la media, en este caso a $r_i e_i$, luego:

$$Var(\hat{r_i}) = \frac{r_i e_i}{e_i^2} = \frac{r_i}{e_i}.$$

La $Var(\hat{r_i})$ depende de r_i, luego un estimador de la varianza es:

$$\widehat{Var}(\hat{r_i}) = \frac{\hat{r_i}}{e_i} = \frac{y_i}{e_i^2}. \tag{3.3}$$

La varianza de este estimador será grande si e_i es pequeño, por tanto para enfermedades muy raras o áreas pequeñas la SMR podría ser muy inestable. Por ello se consideran modelos más sofisticados que suavizan los riesgos en el sentido de que los hacen más estables, es decir, menos variables.

3.1. Cálculo de la SMR del cáncer de próstata en España

A continuación vamos a determinar la SMR del cáncer de próstata en España. Como ya hemos comentado en el capítulo dos, esta razón es el cociente entre las muertes observadas y las esperadas en cada provincia. Frecuentemente las SMR suelen aparecer multiplicadas por 100, y así lo vamos a considerar a continuación.

Para su cálculo y representación vamos a tener en cuenta los esperados según cada periodo de tiempo. La explicación de cómo se realiza el mapa de España en **R** se muestra en el capítulo 4.

3.1.1. Cálculo de la SMR del primer periodo respecto del primer periodo

Seleccionamos los datos del primer periodo y calculamos la SMR correspondiente.

```
> observados1<- observados[observados$periodo==1,]
> esperados1<- esperados.p1[esperados.p1$periodo==1,]
> SMRperiodo1<-observados1$casos/esperados1$total*100
```

Calculamos los intervalos de confianza del riesgo relativo al 95 % (que es el más habitual).

$$IC_{95} = \left[\hat{r_i} - 1.96\sqrt{\widehat{Var}(\hat{r_i})}, \hat{r_i} + 1.96\sqrt{\widehat{Var}(\hat{r_i})}\right]$$

Para calcular la $\widehat{Var}(\hat{r_i})$ usamos la expresión (3.3). En **R** se calcularía como sigue:

```
> var.periodo1 <-observados1$casos/(esperados1$total)^2 #varianza estimada
> l.i.periodo1 <-SMRperiodo1-1.96*(var.periodo1)^(1/2)
> #extremo inferior del intervalo de confianza
> u.i.periodo1 <-SMRperiodo1+1.96*(var.periodo1)^(1/2)
> #extremo superior del intervalo de confianza
```

Recogemos la SMR y los intervalos de confianza en la Tablas (3.1) y (3.2), en las cuales podemos obervar que no hay ningún intervalo de confianza que contenga al 100, no obstante hay algunos que están muy próximos a 100, como es el caso de la provincia de La Coruña y de Segovia. En estas tablas se han escrito con rojo las provincias que tienen un exceso de riesgo mayor que el global de España, es decir, son provincias que tienen una SMR>100 y cuyo intervalo de confianza no contiene al 100. Las provincias que presentan un defecto de riesgo menor que el global de España, es decir, provincias con un SMR<100 y cuyo intervalo de confianza no contiene al 100, se han escrito con verde.

En la Figura (3.1) se muestran las SMR calculadas para el primer periodo tomando como valores esperados los valores del primer periodo.

3.1.2. Cálculo de la SMR del segundo periodo respecto del segundo periodo

Seleccionamos los datos del segundo periodo y calculamos la SMR correspondiente.

```
> observados2<- observados[observados$periodo==2,]
> esperados2<- esperados.p2[esperados.p2$periodo==2,]
> SMRperiodo2<-observados2$casos/esperados2$total*100
```

Calculamos los intervalos de confianza del riesgo relativo al 95 %, para ello calculamos la $\widehat{Var}(\hat{r}_i)$ usando la expresión (3.3).

```
> var.periodo2 <-observados2$casos/(esperados2$total)^2 #varianza estimada
> l.i.periodo2 <-SMRperiodo2-1.96*(var.periodo2)^(1/2)
> #extremo inferior del intervalo de confianza
> u.i.periodo2 <-SMRperiodo2+1.96*(var.periodo2)^(1/2)
> #extremo superior del intervalo de confianza
```

Recogemos la SMR y los intervalos de confianza para el periodo 1992-2008 en las Tablas (3.3) y (3.4), en los cuales podemos obervar qué intervalos contienen al 100 y cuáles no. Análogamente al periodo 1, se han escrito con rojo las provincias que tienen un exceso de riesgo mayor que el global de España, es decir, son provincias que tienen una SMR>100 y cuyo intervalo de confianza

periodo 1975-1991			
Provincia	SMR	Extremo inferior del IC	Extremo superior del IC
1 Álava	92.02732	91.92318	92.13146
2 Albacete	110.44717	110.36488	110.52947
3 Alicante	108.21751	108.16860	108.26641
4 Almería	94.07112	93.99299	94.14924
5 Ávila	77.47897	77.39882	77.55912
6 Badajoz	106.66730	106.60720	106.72740
7 Baleares	106.66730	115.30053	115.42398
8 Barcelona	106.98268	106.95751	107.00784
9 Burgos	82.65539	82.58920	82.72159
10 Cáceres	89.90307	89.83803	89.96811
11 Cádiz	111.22778	111.16298	111.29257
12 Castellón	114.90356	114.83234	114.97477
13 Ciudad Real	102.64547	102.57750	102.71345
14 Córdoba	84.74963	84.69627	84.80299
15 La Coruña	100.08295	100.03636	100.12953
16 Cuenca	78.47246	78.39963	78.54529
17 Gerona	112.44738	112.37658	112.51818
18 Granada	86.55596	86.50169	86.61024
19 Guadalajara	81.88871	81.80049	81.97693
20 Guipúzcoa	87.32163	87.25971	87.38355
21 Huelva	109.73550	109.65302	109.81798
22 Huesca	101.55544	101.47279	101.63809
23 Jaén	79.34486	79.29164	79.39807
24 León	96.61224	96.55350	96.67099
25 Lérida	99.03057	98.96212	99.09903

Tabla 3.1: SMR e intervalos de confianza al 95 % para el periodo 1975-1991. Parte 1.

periodo 1975-1991			
Provincia	SMR	Extremo inferior del IC	Extremo superior del IC
26 La Rioja	104.46859	104.37730	104.55989
27 Lugo	91.68496	91.62891	91.74101
28 Madrid	84.94824	84.92421	84.97226
29 Málaga	114.26644	114.20826	114.32462
30 Murcia	98.91722	98.86487	98.96956
31 Navarra	103.56994	103.50316	103.63672
32 Orense	86.05517	85.99728	86.11306
33 Princip. de Asturias	103.55412	103.50848	103.59975
34 Palencia	109.25338	109.14914	109.35762
35 Las Palmas	121.92257	121.84378	122.00135
36 Pontevedra	107.17513	107.11678	107.23348
37 Salamanca	91.50711	91.44293	91.57128
38 Sta Cruz de Tenerife	114.49665	114.42482	114.56848
39 Cantabria	110.51444	110.44314	110.58573
40 Segovia	99.75085	99.64795	99.85375
41 Sevilla	97.49931	97.45369	97.54492
42 Soria	77.81567	77.71618	77.91516
43 Tarragona	104.72063	104.65695	104.78430
44 Teruel	87.43344	87.34678	87.52011
45 Toledo	86.86749	86.80919	86.92578
46 Valencia	103.24719	103.21145	103.28293
47 Valladolid	114.45912	114.38035	114.53790
48 Vizcaya	117.09059	117.03538	117.14580
49 Zamora	93.84106	93.76312	93.91899
50 Zaragoza	106.98786	106.93640	107.03931

Tabla 3.2: SMR e intervalos de confianza al 95 % para el periodo 1975-1991. Parte 2.

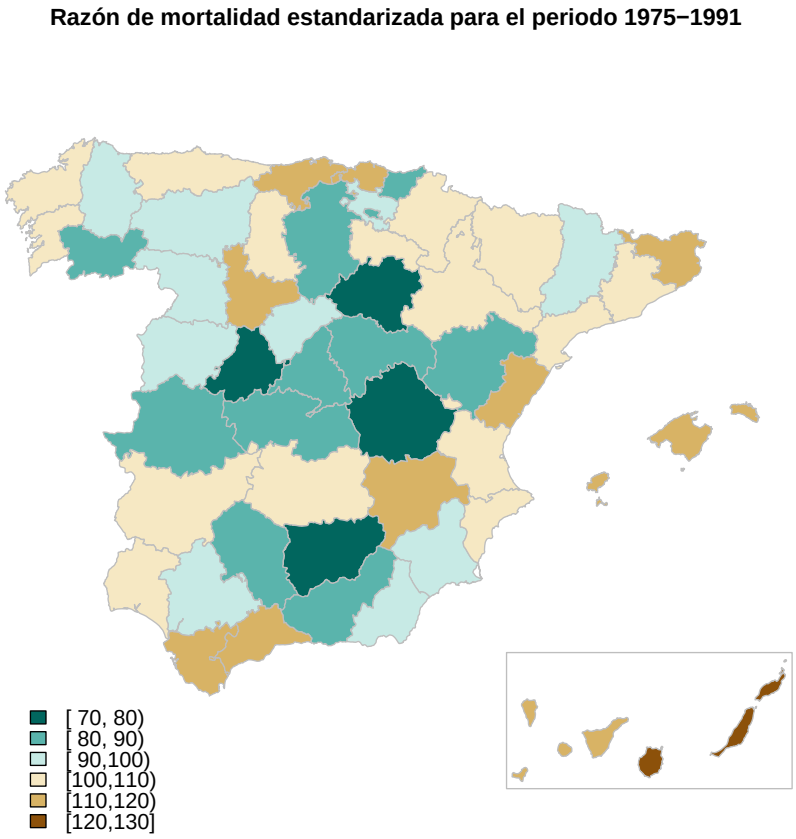

Figura 3.1: Razón de mortalidad estandarizada para el periodo 1975-1991

no contiene al 100. Las provincias que presentan un defecto de riesgo menor que el global de España, es decir, provincias con un SMR<100 y cuyo intervalo de confianza no contiene al 100, se han escrito con verde.

En la Figura (3.2) se muestran las SMR calculadas para el segundo periodo tomando como valores esperados los valores del segundo periodo.

3.1.3. Cálculo de la SMR del segundo periodo respecto del primer periodo

Seleccionamos los datos necesarios para este caso y calculamos la SMR correspondiente.

```
> observados2resp1<- observados[observados$periodo==2,]
> esperados2resp1<- esperados.p1[esperados.p1$periodo==2,]
> SMRperiodo2resp1<-observados2resp1$casos/esperados2resp1$total*100
```

periodo 1992-2008			
Provincia	SMR	Extremo inferior del IC	Extremo superior del IC
1 Álava	98.40416	98.32482	98.48350
2 Albacete	107.01264	106.94577	107.07950
3 Alicante	97.61050	97.57573	97.64526
4 Almería	91.84615	91.78508	91.90722
5 Ávila	97.15825	97.08350	97.23300
6 Badajoz	93.59586	93.54755	93.64417
7 Baleares	106.82080	106.77256	106.86903
8 Barcelona	92.91271	92.89420	92.93122
9 Burgos	98.38001	98.32103	98.43898
10 Cáceres	87.91058	87.85606	87.96509
11 Cádiz	91.03101	90.98525	91.07677
12 Castellón	117.89997	117.83914	117.96080
13 Ciudad Real	94.03099	93.97770	94.08428
14 Córdoba	81.05476	81.01172	81.09781
15 La Coruña	113.05015	113.01057	113.08972
16 Cuenca	85.87793	85.81206	85.94380
17 Gerona	106.07055	106.01717	106.12393
18 Granada	93.51888	93.47302	93.56475
19 Guadalajara	85.60331	85.52728	85.67934
20 Guipúzcoa	99.20028	99.14948	99.25107
21 Huelva	97.05806	96.99331	97.12282
22 Huesca	115.85890	115.78420	115.93361
23 Jaén	87.86206	87.81444	87.90967
24 León	99.48571	99.43780	99.53363
25 Lérida	102.92020	102.86216	102.97824

Tabla 3.3: SMR e intervalos de confianza al 95 % para el periodo 1992-2008. Parte 1.

periodo 1992-2008			
Provincia	SMR	Extremo inferior del IC	Extremo superior del IC
26 La Rioja	112.27201	112.19697	112.34705
27 Lugo	102.42395	102.37269	102.47521
28 Madrid	93.84833	93.82921	93.86746
29 Málaga	93.14765	93.10805	93.18726
30 Murcia	101.20645	101.16449	101.24841
31 Navarra	103.22139	103.16846	103.27432
32 Orense	110.90768	110.85200	110.96337
33 Princip. de Asturias	111.14995	111.11203	111.18787
34 Palencia	100.58111	100.49750	100.66471
35 Las Palmas	130.46746	130.40475	130.53016
36 Pontevedra	119.30135	119.25268	119.35002
37 Salamanca	104.76178	104.70400	104.81955
38 Sta Cruz de Tenerife	103.98666	103.93451	104.03882
39 Cantabria	106.59053	106.53559	106.64547
40 Segovia	99.36707	99.28223	99.45191
41 Sevilla	98.88642	98.84995	98.92289
42 Soria	84.76124	84.67141	84.85107
43 Tarragona	93.95004	93.90210	93.99799
44 Teruel	98.08611	98.00608	98.16614
45 Toledo	96.24431	96.19473	96.29388
46 Valencia	104.17185	104.14259	104.20112
47 Valladolid	109.90651	109.84679	109.96624
48 Vizcaya	102.99567	102.95598	103.03536
49 Zamora	100.88207	100.81456	100.94957
50 Zaragoza	108.63811	108.59601	108.68022

Tabla 3.4: SMR e intervalos de confianza al 95 % para el periodo 1992-2008. Parte 2.

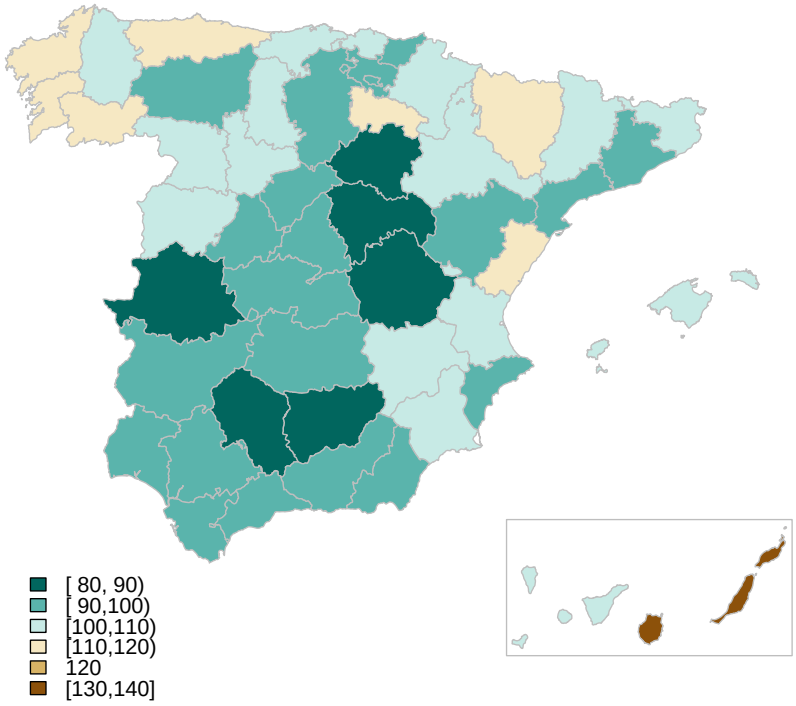

Figura 3.2: Razón de mortalidad estandarizada para el periodo 1992-2008

Calculamos los intervalos de confianza del riesgo relativo al 95 %, para ello calculamos la $\widehat{Var}(\hat{r}_i)$ usando la expresión (3.3).

```
> var.periodo2resp1 <-observados2resp1$casos/(esperados2resp1$total)^2
> #varianza estimada
> l.i.periodo2resp1 <-SMRperiodo2resp1-1.96*(var.periodo2resp1)^(1/2)
> #extremo inferior del intervalo de confianza
> u.i.periodo2resp1 <-SMRperiodo2resp1+1.96*(var.periodo2resp1)^(1/2)
> #extremo superior del intervalo de confianza
```

Recogemos la SMR y los intervalos de confianza en las Tablas (3.5) y (3.6), en los cuales podemos obervar qué intervalos contienen al 100 y cuáles no. Análogamente al periodo 1 y 2, se han escrito con rojo las provincias que tienen un exceso de riesgo mayor que el global de España, es decir, son provincias que tienen una SMR>100 y cuyo intervalo de confianza no contiene al 100. Las provincias que presentan un defecto de riesgo menor que el global de España, es decir, provincias con un SMR<100 y cuyo intervalo de confianza no contiene al 100, se han escrito con verde.

periodo 1992-2008 respecto al periodo 1975-1991			
Provincia	SMR	Extremo inferior del IC	Extremo superior del IC
1 Álava	94.94350	94.86695	95.02004
2 Albacete	103.74169	103.67687	103.80651
3 Alicante	93.32588	93.29264	93.35911
4 Almería	88.06494	88.00639	88.12350
5 Ávila	96.48705	96.41282	96.56129
6 Badajoz	90.42922	90.38255	90.47590
7 Baleares	104.06972	104.02273	104.11671
8 Barcelona	89.61832	89.60047	89.63617
9 Burgos	96.59573	96.53783	96.65364
10 Cáceres	85.49558	85.44256	85.54860
11 Cádiz	86.38933	86.34590	86.43276
12 Castellón	114.78843	114.72920	114.84765
13 Ciudad Real	90.79416	90.74270	90.84561
14 Córdoba	78.10929	78.06781	78.15077
15 La Coruña	109.98124	109.94274	110.01974
16 Cuenca	84.59675	84.53187	84.66164
17 Gerona	103.072428	103.02055	103.12428
18 Granada	89.60168	89.55773	89.64562
19 Guadalajara	85.16259	85.08695	85.23822
20 Guipúzcoa	95.34348	95.29465	95.39230
21 Huelva	93.03777	92.97570	93.09985
22 Huesca	114.54132	114.46746	114.61517
23 Jaén	84.42009	84.37434	84.46584
24 León	97.40539	97.35847	97.45230
25 Lérida	100.89467	100.83777	100.95157

Tabla 3.5: SMR e intervalos de confianza al 95 % para el periodo 1992-2008 respecto al periodo 1975-1991. Parte 1.

periodo 1992-2008 respecto al periodo 1975-1991			
Provincia	SMR	Extremo inferior del IC	Extremo superior del IC
26 La Rioja	109.33051	109.25744	109.40358
27 Lugo	101.75593	101.70501	101.80685
28 Madrid	90.44334	90.42491	90.46177
29 Málaga	88.82532	88.78755	88.86308
30 Murcia	96.89576	96.85559	96.93593
31 Navarra	100.72587	100.67422	100.77752
32 Orense	109.80886	109.75372	109.86399
33 Princip. de Asturias	107.60661	107.56990	107.64332
34 Palencia	98.59275	98.51080	98.67470
35 Las Palmas	125.42646	125.36618	125.48674
36 Pontevedra	115.54346	115.49632	115.59060
37 Salamanca	104.13039	104.07296	104.18782
38 Sta Cruz de Tenerife	100.51233	100.46191	100.56274
39 Cantabria	103.45516	103.40184	103.50849
40 Segovia	98.54571	98.46158	98.629855
41 Sevilla	94.39442	94.35961	94.42924
42 Soria	85.33804	85.24760	85.42849
43 Tarragona	91.56986	91.52312	91.61659
44 Teruel	97.07215	96.99295	97.15136
45 Toledo	93.93642	93.88803	93.98480
46 Valencia	100.16652	100.13838	100.19466
47 Valladolid	107.23896	107.18068	107.29723
48 Vizcaya	98.40796	98.37004	98.44589
49 Zamora	100.23226	100.16519	100.29933
50 Zaragoza	105.76047	105.71949	105.80146

Tabla 3.6: SMR e intervalos de confianza al 95 % para el periodo 1992-2008 respecto al periodo 1975-1991. Parte 2.

En la Figura (3.3) se muestran las SMR calculadas para el segundo periodo tomando como valores esperados los valores del primer periodo.

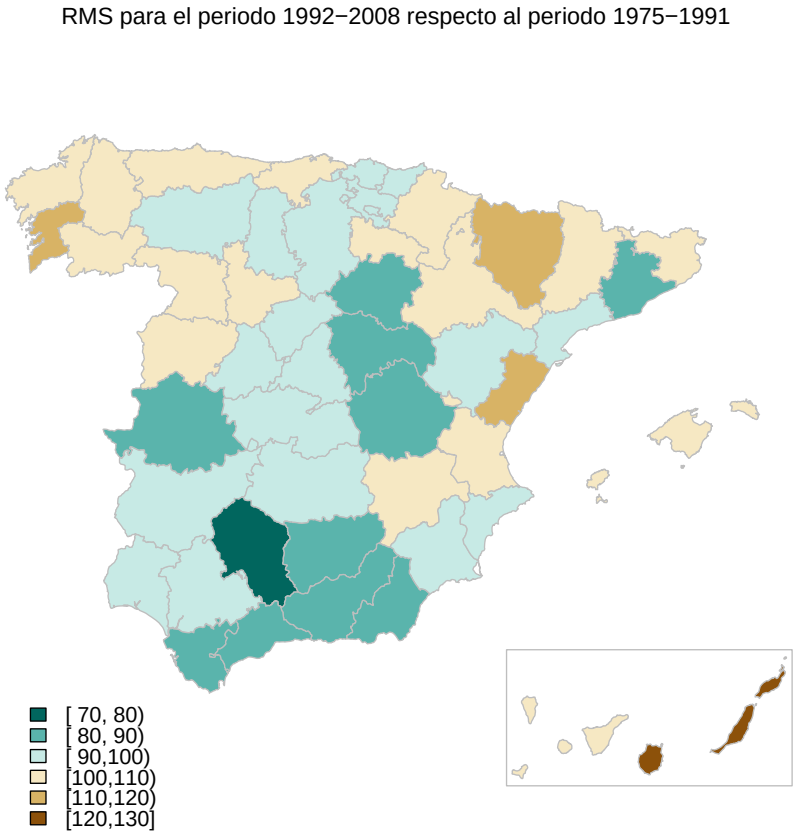

Figura 3.3: Razón de mortalidad estandarizada para el periodo 1992-2008 respecto al periodo 1975-1991

Capítulo 4

Aplicación del modelo BYM con INLA

Los modelos más utilizados son los modelos jerárquicos bayesianos, concretamente el propuesto por Besag, York y Mollie (BYM) [12] el cual supone la principal referencia en métodos de suavización de riesgos e incluye dos efectos aleatorios: uno recoge la dependencia espacial (se sabe que regiones vecinas presentan riesgos parecidos debido a que tienen estilos de vida similares, así como mismo tipo de alimentación o igual contaminación atmosférica, por ejemplo) y el otro recoge la dependencia no estructurada. Dicho modelo es estimado por el método INLA. La estimación obtenida mediante el modelo BYM corresponde a una suavización de los riesgos de mortalidad que será más pronunciada en aquellas áreas con menor población.

Este modelo establece una ponderación entre dos fuentes de variación para obtener una estimación del riesgo relativo de cada área. La primera de estas fuentes de variación, de estructura espacial, comparte información entre unidades vecinas para modelizar la dependencia geográfica de los riesgos. Consideraremos que un área es vecina de otra si comparte frontera.

Este efecto hace posible que regiones vecinas tengan estimaciones similares y de esta forma el riesgo varíe geográficamente de forma suave. El segundo de los efectos, espacialmente heterogéneo, toma valores independientes en todas las unidades geográficas lo que permite que localizaciones vecinas presenten riesgos diferentes. En el caso de que el área pequeña tenga una población de gran tamaño tendrá mayor peso la información proporcionada por este área; en cambio, si presenta una población de reducido tamaño tendrá mayor peso la información del resto de áreas (o áreas vecinas). Mediante la ponderación de ambos tipos de información el modelo BYM minimiza el problema relativo a la estabilidad de los riesgos estimados. A la estimación del riesgo obtenida a partir del modelo BYM se la denominará de aquí en adelante Razón de Mortalidad Estandarizada Suavizada (RMEs) [32].

En este capítulo vamos a exponer las ideas básicas de la modelización bayesiana antes de aplicar el modelo de BYM.

4.1. Descripción del modelo BYM

4.1.1. Estadística bayesiana

Fundamentalmente la estadística bayesiana permite que los parámetros de cualquier función de probabilidad tengan distribuciones, es decir, sean variables aleatorias. A estas distribuciones se les llama *distribuciones a priori*. Debido a esto, los parámetros de las distribuciones a priori mencionadas pueden ser también variables aleatorias, de aquí que se establezca una jerarquía. Estos modelos jerárquicos son la base de la inferencia bayesiana.

Combinando la verosimilitud con las distribuciones a priori de los parámetros, se forma una distribución llamada *distribución a posteriori* que describe el comportamiento de los parámetros después de haber visto los datos.

Resumiendo, la distribución a priori es la idea que tenemos de la variación de los datos. Los datos se modelizan con ayuda de la función de verosimilitud y a través de la distribución a posteriori actualizamos el conocimiento que tenemos de la variación de los datos. Además esta distribución a posteriori puede convertirse en la distribución a priori de los parámetros antes de experimentar con los siguientes datos.

Como ventajas de la estadística bayesiana podemos decir que permite la formulación de modelos más complejos y flexibles que la estadística frecuentista y que el software bayesiano ha alcanzando en la actualidad un grado de madurez que permite la utilización cotidiana de esta metodología. No obstante, la estadística bayesiana es criticada por los estadísticos frecuentistas por conciderarla subjetiva ya que dicen que la estadística bayesiana se encuentra en el observador, mientras que la estadística clásica o frecuentista es un concepto objetivo, que se encuentra en la naturaleza. De este modo, en estadística clásica sólo se toma como fuente de información las muestras finitas obtenidas. En el caso bayesiano, sin embargo, además de la muestra también juega un papel fundamental la información previa de los fenómenos que se tratan de modelizar [16].

Veamos las diferencias entre ambas:

- En estadística frecuentista la probabilidad de un suceso es la proporción de veces que se daría dicho suceso si se repitiera indefinidamente un experimento. La probabilidad es objetiva y es igual para todos los observadores, mientras que para un bayesiano la probabilidad de un suceso es una medida de la credibilidad de los posibles valores que puede tomar dicho

suceso. La probabilidad es subjetiva y depende del observador [17].

- En estadística clásica los parámetros de un modelo estadístico se suponen constantes que habremos de determinar. Sin embargo en estadística bayesiana los parámetros de un modelo estadístico se suponen variables aleatorias cuyas distribución de probabilidad habremos de determinar.

- La inferencia de la estadística frecuentista se basa en la función de verosimilitud, la cual mide la credibilidad de los valores de los parámetros dados los datos que hemos observado: Verosim(Parámetros)=P(datos|Parámetros). La principal herramienta de la estadística bayesiana es el **teorema de Bayes** que actualiza la información, a priori, de los parámetros con la información proporcionada por los datos.

$$\pi(\theta|datos) = \frac{\pi(datos|\theta)\pi(\theta)}{\pi(datos)} \propto \pi(datos|\theta)\pi(\theta),$$

donde $\pi(datos|\theta)$ es la función de verosimilitud, que ya hemos definido anteriormente, $\pi(\theta)$ es la distribución a priori y $\pi(\theta|datos)$ es la distribución a posteriori que vamos a definir a continuación.

Definición 4.1.1. *Sean y los datos observados y θ el/los parámetro/s del modelo. Sea $\pi(\theta)$ la distribución a priori y $\pi(y|\theta)$ la verosimilitud. Se define la distribución a posteriori como*

$$\pi(\theta|y) = \pi(y|\theta)\pi(\theta)/C, \text{ donde } C = \int \pi(y|\theta)\pi(\theta)d\theta \tag{4.1}$$

Notación: $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$.

Observaciones.

- En la inferencia bayesiana, las distribuciones a posteriori aparecen continuamente. Por ejemplo, los cuantiles o momentos se pueden escribir en función de la media a posteriori de funciones de θ, siendo la media a posteriori de una función $f(\theta)$:

$$E[f(\theta|y)] = \frac{\int f(\theta)\pi(\theta)\pi(y|\theta)d\theta}{\int \pi(\theta)\pi(y|\theta)d\theta}.$$

No obstante, el cálculo de estas integrales ha sido uno de los principales problemas prácticos en la inferencia bayesiana, sobre todo en dimensiones altas. Como dicho cálculo es imposible realizarlo de forma exacta, se aproxima numéricamente mediante aproximaciones de Laplace o integración de Monte Carlo, como veremos más adelante [21].

- La verosimilitud informa sobre los parámetros a través de los datos, mientras que las distribuciones a priori informan a través de las asunciones, por eso cuando haya una gran cantidad de datos, la verosimilitud contribuirá en mayor medida a la estimación del riesgo. Sin embargo cuando la muestra de datos sea pequeña, la/s distribución/es a priori dominarán el análisis.

4.1.2. Modelo de BYM

Vamos a escribir el modelo de BYM basándonos en [12] y [15].

Sean $x_i = \log(r_i)$ el logaritmo del riesgo relativo, que es desconocido, en la provincia i-ésima para $i = 1, \ldots, 50$ e y_i el número de casos observados de muertes durante el periodo del estudio.

Cuando la enfermedad no es contagiosa y rara, se suele suponer que los y_i's son variables de Poisson independientes con medias $e_i r_i$, donde e_i es el número de casos esperados en la provincia i-ésima suponiendo riesgo constante, es decir, basado sólo en la tasa de incidencia global y en la población de riesgo en la provincia i-ésima ajustada por edad.

Definición 4.1.2. *El **modelo de BYM** se define en dos etapas:*

$$\begin{aligned} y_i | r_i &\sim P(e_i r_i) \\ \log(r_i) &= t + u_i + v_i, \ \textit{para } i = 1, \ldots, 50, \end{aligned} \tag{4.2}$$

donde y_i es el número de casos observados de muertes durante el periodo del estudio, r_i es el riesgo relativo en la provincia i-ésima, e_i es el número de casos esperados en la provincia i-ésima, t es un término estándar asociado a las covariables medidas que son conocidas o sospechadas para ser relevantes en la enfermedad que normalmente tiene la forma de un modelo lineal $t = A\theta$ con al menos A desconocido, u_i y v_i son términos adicionales que se pueden interpretar como sustitutos para variables desconocidas o no observadas. En particular, u_i muestra la estructura espacial en la que dos áreas que son vecinas tendrían valores más "parecidos" que cualquier par arbitrario de áreas que tomásemos, mientras que v_i representa las variabilidad no estructurada, es decir, como un "ruido" que recoge los efectos aleatorios no espaciales (llamado heterogenidad).

La segunda etapa de este modelo también se puede expresar como:

$$r_i = e^{t+u_i+v_i}, \ \text{para } i = 1, \ldots, 50.$$

Por simplicidad y sin pérdida de información podemos ignorar t y denotaremos a u_i por u y a v_i por v.

Hasta el momento no tenemos mas información, así que podemos suponer que u y v son independientes y que $v \sim \mathcal{N}(0, \sigma_v^2)$ siendo σ_v^2 su varianza que es desconocida. Para u se elige una función de densidad de la siguiente familia:

$$p(u) \propto \exp\left(-\sum_{i<j} \omega_{ij} \phi(u_i - u_j) \right) \ \text{con } u \in \mathbb{R},$$

donde ω_{ij} son los pesos no negativos tales que $\omega_{ij} = 0$ excepto si i y j están en áreas vecinas y ϕ es una función par de z creciente con $|z|$.

Por lo tanto la densidad condicional de u_i es:

$$p_i(u_i|\cdots) \propto \exp\left(-\sum_{j\in\partial_i}\omega_{ij}\phi(u_i-u_j)\right), \tag{4.3}$$

donde ∂_i denota a las áreas vecinas de i.

Por simplicidad tomamos $\omega_{ij}=1$ y $\phi(z)=z^2/(2\sigma_u^2)$ con k constante positiva desconocida.

Así (4.3) se transforma en:

$$p(u|\sigma_u^2) \propto \frac{1}{(\sigma_u^2)^{n/2}}\exp\left(-\frac{1}{2\sigma_u^2}\sum_{i\sim j}(u_i-u_j)^2\right),$$

donde $i\sim j$ indica que i y j son vecinos. Notar que esta aproximación especifica la dependencia espacial como un promedio del efecto de sus áreas vecinas limitando la vecindad a sus áreas contiguas.

Esto recibe el nombre de *autorregresión intrínseca gaussiana* y una de sus ventajas es que los momentos condicionales están definidos como funciones simples de los "valores vecinos" y el número de vecinos.

$$E(u_i|\ldots)=\bar{u_i}=\sum_{i\sim j}u_i/n_i,\ Var(u_i|\ldots)=\sigma_u^2/n_i$$

siendo n_i la cardinalidad de ∂_i, es decir, el número de vecinos de i.

Como resultado final se obtienen para cada provincia, la media a posteriori de la distribución del riesgo así como la probabilidad de que ésta sea mayor de 100.

Para calcular las distribuciones a posteriori no se usa la definición previa sino que, en la práctica, se aproximan con los métodos de Monte Carlo.

En nuestro caso, la densidad a posteriori de u, v, σ_u^2 y σ_v^2 en la cual basamos las inferencias es:

$$P(u,v,\sigma_u^2,\sigma_v^2|y) \propto \prod_{i=1}^{n}(\exp(-c_ie^{x_i})(c_ie^{x_i})^{y_i}/y_i!)\times\sigma_u^{2-n/2}\exp\left(-\frac{1}{2\sigma_u^2}\sum_{i\sim j}(u_i-u_j)^2\right)$$
$$\times\sigma_v^{2-n/2}\exp\left(-\frac{1}{2\sigma_v^2}\sum_{i=1}^{n}v_i^2\right)\times prior(\sigma_u^2,\sigma_v^2)$$

donde $prior(\sigma_u^2,\sigma_v^2)$ es la densidad a priori de los dos hiperparámetros.

Para evitar singularidades y lograr que la distribución sea correcta realizamos la siguiente modificación:

$$prior(\sigma_u^2,\sigma_v^2) \propto \exp^{-\epsilon/2\sigma_u^2}\exp^{-\epsilon/2\sigma_v^2},\ \text{con}\ \sigma_u^2,\sigma_v^2>0$$

donde ϵ es una constante pequeña positiva que tomamos ϵ=0.001.

Las densidades condicionales de u_i y v_i no se ven modificadas, mientras que las de σ_u^2 y σ_v^2 permanecen en la familia de distribuciones gamma-inversas. Es decir,

$$\sigma_v^2 \sim Gamma^{-1}(0.001, 0.001); \sigma_u^2 \sim Gamma^{-1}(0.001, 0.001).$$

Un resumen de este modelo se recoge en la Figura (4.1), en la cual E_i representa los casos observados, en vez de e_i.

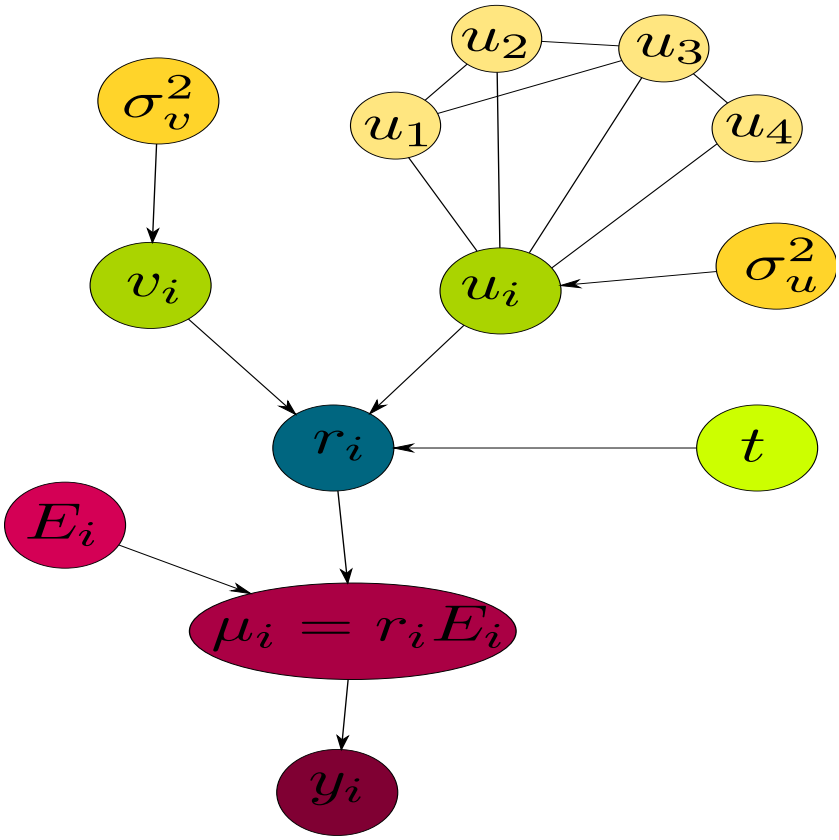

Figura 4.1: Modelo BYM

En la práctica una vez obtenidas las estimaciones a posteriori de los parámetros, se estiman los riesgos de cada provincia a través de la siguiente expresión:

$$RME = (\exp(u_i + v_i)) * 100$$

De esta forma se excluye al parámetro t de la fórmula (4.2) de la estimación, por lo que la RME compara el riesgo de cada provincia con el nivel medio de España.

En esta sección hablaremos también de la matriz de vecindades ya que el modelo BYM suaviza

los riesgos prestando información entre regiones vecinas en lugar de hacer uso de todas las unidades, independientemente de su ubicación geográfica.

4.2. Lectura de datos y cartografía

En el siguiente paso importaremos la cartografía necesaria para calcular la matriz de vecindades y representar gráficamente los mapas. Veamos primero la importancia de los mapas en epidemiología.

4.2.1. Importancia de los mapas

Tradicionalmente, los mapas o atlas se han utilizado para representar la geografía de un país o de una región. Pueden ser políticos, físicos, económicos,... En cambio, los mapas o atlas que vamos a usar, ilustrarán las diferencias de mortalidad que existen entre las distintas provincias españolas con respecto a España.

Este trabajo se encamina en el ámbito de la representación cartográfica de enfermedades o "disease mapping" (en inglés) [30]. En este campo se suelen aplicar modelos a nivel de área, donde un área puede ser una sección censal, una provincia (como es nuestro caso), un país, etc. Generalmente se asume que el número de eventos (casos de mortalidad, por ejemplo) localizados en cada una de estas áreas sigue una distribución de Poisson con media igual al riesgo de mortalidad (r_i) por el número de casos esperados (e_i). Actualmente el riesgo se estima mediante modelos jerárquicos bayesianos, siendo el modelo de BYM [12] uno de los más utilizados para el estudio de la distribución geográfica de riesgos en áreas pequeñas. Una vez realizada la estimación, los valores del riesgo estimado se representan en un mapa, que permite identificar las zonas geográficas con mayor riesgo de mortalidad [22].

Los primeros mapas sanitarios que se realizaron fueron a finales del siglo XVIII y principios del XIX, cuando la fiebre amarilla era uno de los mayores problemas en Salud Pública de Estados Unidos. Para conocer la causa de esta enfermedad se representó el mapa de Nueva York junto con los casos detectados de esta enfermedad. Se realizaron más mapas como éste en los años sucesivos y gracias a ellos se detectó el patrón geográfico de la fiebre amarilla donde se mostraba que esta enfermedad era más común en las zonas en donde las temperaturas eran altas, tenían una humedad elevada y había una gran concentración de insectos en el aire [22].

Desde entonces se ha seguido realizando este tipo de mapas con resultados exitosos, como el descubrimiento del origen del cólera en Londres a mital del siglo XIX, cuando se observó (gracias a los mapas que dibujó el doctor Snow) que la falta de drenaje en un determinado distrito era la causa de esta epidemia.

En nuestros mapas se van a representar los riesgos estimados de mortalidad por cáncer de próstata que existen entre las distintas provincias españolas con respecto a España. Éstos permitirán apreciar las diferencias existentes entra cada provincia y España en global. Se van a considerar como riesgos de mortalidad estimados, la media de las distribuciones a posteriori. Como ya hemos comentado anteriormente, si SMR>100 en una determinada provincia, dicha provincia presenta un exceso de riesgo respecto a España, mientras que si SMR<100 en una provincia entonces esa provincia tendrá un defecto de riesgo comparado con España.

4.2.2. Cartografía

Para abrir la cartografía debemos disponer de los archivos en formato ESRI Shapefile (SHP). Éstos son tres y deben estar en la misma carpeta: *esp_ prov.shp, esp_ prov.dbf* y *esp_prov.shx*. Leeremos la cartografía y la asignaremos al objeto *Carto* mediante la siguiente sintaxis:

```
> Carto<-readShapePoly("D:/Paula/master/TFM/tfm/esp_prov.shp")
```

Podemos dibujar la cartografía con sus ejes de coordenadas tal y como muestra en la Figura (4.2).

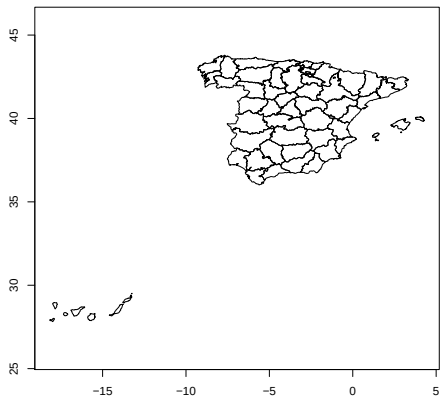

Figura 4.2: Mapa de España sin modificar

Para que quede más estético, vamos a acercar las islas Canarias mediante el siguiente código y mostrando el resultado en la Figura (4.3).

```
> #Mediante el argumento IDvar ordenamos la cartografia según la variable CODIGO
> xName <- readShapePoly("esp_prov.shp",IDvar="ESP_PROV_I",
+                        proj4string=CRS("+proj=longlat +ellps=clrk66"))
```

```
> xNameTras <- xName[xName$ESP_PROV_I !=51 & xName$ESP_PROV_I !=52,]
> #Ïndice de los polinomos a trasladar
> idx<-c(35, 38)
> #plot(xName[idx,])
>
> #Traslación que aplicamos a long/lat
> #shft<-c(6, 7) #NUEVO
> #shft<-c(5, 8) #NUEVO
> shft<-c(18, 7) #NUEVO
> #Bucle sobre las provincias a trasladar
> for(prov.index in idx)
+ {
+     prov<-slot(((xName [prov.index,])), "polygons")
+     prov.polygons <-slot(prov[[1]], "Polygons")
+
+     new.polygons <-as.list(rep(NA, length(prov.polygons)))
+     for(pol.index in 1:length(prov.polygons))
+     {
+         xx<-  slot(prov.polygons[pol.index][[1]], "coords")
+         xx<-xx+matrix(shft, byrow=TRUE, ncol=2, nrow=nrow(xx))
+         #points(xx)
+         new.polygons[pol.index]<-Polygon(xx)
+     }
+
+     xNameTras@polygons[prov.index]<-Polygons(new.polygons, ID=slot(prov[[1]],
+                                                                    "ID"))
+ }
> #Aumentar un poco los márgenes del bounding box del mapa total
> bbox.xNameTras <- bbox(xNameTras[(1:50),])                    #
> bbox.xNameTras[c(1,2)] <- bbox.xNameTras[c(1,2)] - 0.1
> # Bounding box del mapa total
> bbox.xNameTras[c(3,4)] <- bbox.xNameTras[c(3,4)] + 0.1        #
> xNameTras@bbox<-bbox.xNameTras
> #Determinar bounding box de las islas
> bbox.islas <- bbox(xNameTras[(1:50)[idx],])
> #Aumentar un poco los márgenes del rectángulo
> bbox.islas[c(1,2)] <- bbox.islas[c(1,2)] - 0.1
> bbox.islas[c(3,4)] <- bbox.islas[c(3,4)] + 0.1
> #Construir rectángulo para las islas.
> rect.islas <- rbind(c(bbox.islas[1],bbox.islas[2]),c(bbox.islas[1],
```

```
+                bbox.islas[4]),c(bbox.islas[3],bbox.islas[4]),
+                c(bbox.islas[3],bbox.islas[2]),c(bbox.islas[1],bbox.islas[2]))
> #Construir objetos Polygon y Polygons con el rectángulo para las islas.
> polygon.islas <- Polygon(rect.islas)
> polygons.islas <- Polygons(list(polygon.islas), "Cuadro")
> #Incluir el nuevo polígono en xNameTras
> num.rect <- length(xNameTras@polygons)+1
> xNameTras@polygons[[num.rect]] <- polygons.islas
> #Determinar el orden en que se dibujará el nuevo polígono,
> #así se dibujará primero.
> xNameTras@plotOrder <- c(as.integer(num.rect),xNameTras@plotOrder)
> xtodas <- xNameTras
> #Incluir un nuevo nivel para el rectángulo de Canarias
> levels(xtodas$NAME)<- cbind(xtodas$NAME, " ")
> #Añadir la nueva fila al data frame correspondiente al rectángulo de Canarias,
> #con valores no válidos
> df <- xtodas@data
> df[num.rect,] <- c(num.rect, as.factor(NA), as.factor(" "),
+                    rep(NA, dim(df)[2]-3))
> xtodas@data <- df
> CartoMod<-xtodas
> plot(CartoMod, axes=T)
> Carto<-xName[xName$ESP_PROV_I !=51 & xName$ESP_PROV_I !=52,]
```

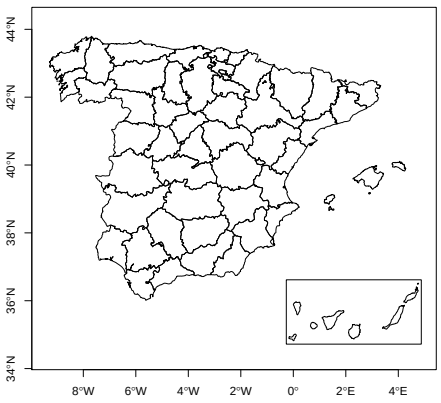

Figura 4.3: Mapa de España con las islas Canarias trasladadas

4.2.3. Construcción de la matriz de vecindad

Para continuar con el análisis, vamos a crear un archivo que contenga toda la información de las vecindades entre áreas. Como ya hemos comentado anteriormente un área se considerará vecina de otra si comparten frontera.

Usando la función *poly2nb* de la librería *maptools* obtenemos la información de las vecindades.

```
> x.nb<-poly2nb(Carto)
```

Para la construcción del fichero nos ayudaremos de la función *nb2inla* (autor: Virgilio Gómez-Rubio) que pasa un objeto "x.nb" a otro con formato de INLA.

```
> nb2inla <-function(file, nb)
+ {
+
+     n<-length(nb)
+
+     if(!file.create(file))
+ {
+         stop("Cannot open file")
+     }
+
+     txt<-paste(n, "\n", sep="")
+     cat(txt, file=file, append = TRUE)
+
+     for(i in 1:length(nb))
+     {
+         txt<-paste(c(i, length(nb[[i]]), nb[[i]]), collapse=" ")
+         txt<-paste(txt, "\n",sep="")
+         cat(txt, file=file, append = TRUE)
+     }
+ }
```

Ahora ejecutamos la función *nb2inla* a la lista de vecinos

```
> nb2inla(file="esp_prov_nb_SinMod.inla", x.nb)
```

Modificamos manualmente la matriz de vecindades "esp_ prov_ nb_ SinMod.inla" y generamos "esp_ prov_ nb.inla". Estos cambios son necesarios porque las islas no tienen vecinos y la función

INLA da problemas cuando existen regiones sin ningún vecino. Las modificaciones que realizamos son las siguientes:

- Sustituimos: la fila 7 1 0 por 7 2 8 46
- Sustituimos: la fila 8 3 17 25 43 por 8 4 7 17 25 43
- Sustituimos: la fila 35 1 0 por 35 2 11 38
- Sustituimos: la fila 11 2 29 41 por 11 4 29 35 38 41
- Sustituimos: la fila 38 1 0 por 38 2 11 35

La idea es conectar a Baleares con Barcelona y Valencia, conectar a Las Palmas con Cádiz y Santa Cruz de Tenerife, y conectar a Santa Cruz de Tenerife con Cádiz y Las palmas. A esta nueva matriz de vecindades la llamamos "esp_ prov_ nb" y está formada por 50 "vectores" que corresponden a cada provincia. Cada vector comienza con ", le sigue un número que identifica a la provincia (las cuales están numeradas por orden alfabético), el siguiente número indica el número de vecinos que tiene esa provincia y las siguientes cifras indican quienes son sus vecinos. Por ejemplo, el vector "1 5 9 20 26 31 48" indica con el 1 que se trata de la provincia de Álava, el 5 quiere decir que tiene 5 vecinos que son Burgos (9), Guipúzcoa (20), La Rioja (26), Navarra (31) y Vizcaya (48).

```
> esp_prov_nb <- c(50,"1 5 9 20 26 31 48","2 7 3 13 16 18 23 30 46",
+                  "3 3 2 30 46","4 2 18 30","5 6 10 28 37 40 45 47",
+                  "6 6 10 13 14 21 41 45","7 2 8 46","8 4 7 17 25 43",
+                  "9 8 1 26 34 39 40 42 47 48","10 5 5 6 13 37 45",
+                  "11 4 29 35 38 41","12 3 43 44 46",
+                  "13 7 2 6 10 14 16 23 45","14 6 6 13 18 23 29 41",
+                  "15 2 27 36","16 7 2 13 19 28 44 45 46","17 2 8 25",
+                  "18 6 2 4 14 23 29 30","19 6 16 28 40 42 44 50",
+                  "20 3 1 31 48","21 2 6 41","22 3 25 31 50",
+                  "23 4 2 13 14 18","24 7 27 32 33 34 39 47 49",
+                  "25 5 8 17 22 43 50","26 5 1 9 31 42 50",
+                  "27 5 15 24 32 33 36","28 5 5 16 19 40 45",
+                  "29 4 11 14 18 41","30 4 2 3 4 18",
+                  "31 5 1 20 22 26 50","32 4 24 27 36 49","33 3 24 27 39",
+                  "34 4 9 24 39 47","35 2 11 38","36 3 15 27 32",
+                  "37 4 5 10 47 49","38 2 11 35","39 5 9 24 33 34 48",
+                  "40 6 5 9 19 28 42 47","41 5 6 11 14 21 29",
+                  "42 5 9 19 26 40 50","43 5 8 12 25 44 50",
+                  "44 6 12 16 19 43 46 50","45 6 5 6 10 13 16 28",
```

```
+                  "46 6 2 3 7 12 16 44","47 7 5 9 24 34 37 40 49",
+                  "48 4 1 9 20 39","49 4 24 32 37 47",
+                  "50 8 19 22 25 26 31 42 43 44")
```

4.3. Suavización de Riesgos

4.3.1. INLA

Integrated Nested Laplace Approximation (INLA) se usa como una herramienta para la inferencia bayesiana. INLA es una alternativa prometedora a los métodos Markov Chain Monte Carlo (MCMC), que proporciona resultados adecuados en un corto tiempo computacional [14].

Hasta hace poco, las técnicas MCMC (que sirven para obtener estimaciones fiables de las distribuciones a posteriori) han sido usadas para la inferencia bayesiana, pero estos métodos son muy costosos computacionalmente ya que los modelos espacio-temporales de enfermedades cartográficos son muy complejos. Recientemente (en torno al año 2009) se ha propuesto como alternativa a estos métodos, INLA [25]. Esta metodología ofrece aproximaciones fiables a las marginales a posteriori en un corto periodo de tiempo computacional. Además de su fiabilidad y de su rapidez computacional, INLA es muy fácil de aplicar ya que existe un paquete en **R** que permite aplicarlo.

Antes de profundizar en el funcionamiento de INLA necesitamos previamente introducir los siguientes conceptos:

Markov Chain Monte Carlo (MCMC)

Generalmente, no es posible evaluar el denominador de la distribución a posteriori (4.1) de forma analítica. Así que se hace necesario el tratamiento numérico, aproximado del problema, (salvo para unos determinados casos). Además si la dimensión del espacio paramétrico es mayor que 1, la integración es aún más complicada.

Estos complicados problemas de integración han constituido la principal dificultad del análisis bayesiano. Hay distintos métodos de integración numérica mediante aproximaciones determinísticas, ver [18], [19], [20]. Pero estos métodos no tienen en cuenta la aleatoriedad del problema o que las funciones implicadas sean densidades probabilísticas, por ejemplo.

Gracias a los métodos basados en simulación Monte Carlo mediante Cadenas de Markov, MCMC, problemas que hasta hace aproximadamente 20 años no eran analíticamente tratables y que precisaban distintas aproximaciones numéricas para las integrales implicadas, se pueden ahora resolver. La idea de estos métodos es simular muestras con algún tipo de dependencia que

converjan (bajo ciertas condiciones de regularidad) a la distribución de interés $\pi(\theta|y)$. De este modo, estos métodos permiten muestrear la distribución a posteriori, aunque ésta sea desconocida, gracias a la construcción de una cadena de Markov, mediante simulación Monte Carlo, cuya distribución estacionaria sea, precisamente $\pi(\theta|y)$.

> "MCMC es, esencialmente, integración Monte Carlo, haciendo correr por largo tiempo una inteligentemente construida cadena de Markov ."[21]

Precisando aún más, vamos a introducir MCMC como un método para evaluar expresiones del siguiente tipo:

$$E[f(X)] = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx},$$

donde X un vector de k variables aleatorias con distribución $\pi(.)$, en nuestro caso X es el vector de los parámetros del modelo y $\pi(.)$ es su distribución a posteriori.

Para ello, vamos a describir las dos partes que componen este método: integración Monte Carlo y Cadenas de Markov [21].

Cadena de Markov

Una *cadena de Markov* es una sucesión de variables aleatorias, $\{X_1, X_2, \ldots, X_t, \ldots\}$ tal que

$$\forall t \geq 0, P(X_{t+1}|X_t, X_{t-1}, \ldots, X_1) = P(X_{t+1}|X_t).$$

Es decir, dado X_t, el siguiente estado X_{t+1} no depende de la historia de la cadena $\{X_1, \ldots, X_{t-1}\}$.

Integración Monte Carlo

Es un método que mediante modelos matemáticos intenta reproducir el comportamiento aleatorio de sistemas reales no dinámicos. Veamos cómo lo logra.

La integración Monte Carlo evalúa $E[f(X)]$ extrayendo muestras $\{X_t, t = 1, \ldots, n\}$ de $\pi(.)$ y entonces aproxima

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t).$$

Es decir, que la media poblacional de $f(X)$ se estima a través de la media muestral. Si las muestras $\{X_t\}$ son independientes, por la ley de los grandes números podemos asegurar que las aproximaciones pueden ser tan precisas como queramos, basta con aumentar el tamaño de la muestra, n. Notar que aquí el valor de n lo decide el analista.

En general, no es factible extraer muestras $\{X_t\}$ independientes, sin embargo no es necesario que las muestras sean independientes. Las $\{X_t\}$ pueden ser generadas por cualquier proceso, el cual extraiga muestras a través de un conjunto de puntos de $\pi(.)$, donde ésta no se anule, en las proporciones adecuadas. Una forma de hacer esto es a través de las cadenas de Markov, teniendo $\pi(.)$ como su distribución estacionaria. Esto es "Markov Chain Monte Carlo".

Observación. Ley de los grandes números. Sea (Ω, F, P) un espacio de probabilidad. Sean $X_1, X_2, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas definidas en dicho espacio de probabilidad y con media μ, $E[X_i] = \mu$. Entonces se tiene:

$$P\left(\omega \in \Omega : \lim_{n \to +\infty} \frac{X_1 + \cdots X_n}{n} = \mu\right) = 1.$$

Gaussian Markov random field (GMRF)

A *Gaussian Markov random field (GMRF)*, $\mathbf{x} = \{x_i | i \in \nu\}$, es un vector aleatorio normal $|\nu|$-dimensional que satisface propiedades de independencia condicional de Markov. Para simplificar podemos asumir que $\nu = \{1, \ldots, n\}$.

Las propiedades de independencia condicional se pueden representar usando un grafo no dirigido $G = (\nu, \epsilon)$ con ν nodos y ϵ arcos. Diremos que x_i e x_j son condicionalmente independientes si y sólo si $i, j \notin \epsilon$, es decir, si no existe ningún arco entre los nodos x_i e x_j. Entonces diremos que $\mathbf{x}$ es un GMRF con respecto a G. Cuando $\{i, j\} \in \epsilon$ diremos que i y j son vecinos [24].

Los GMRFs son también conocidos en inglés como "conditional auto-regressions (CARs)" siguiendo el trabajo de Besag [23]. Esta familia de modelos cubre un buen número de modelos jerárquicos bayesianos, incluyendo la mayoría de los usados en estadística espacial. Además, las propiedades de Markov pueden ser usadas para modelizar la dependencia local.

Modelo de variable latente

Un modelo de variable latente es un modelo estadístico que relaciona un conjunto de variables (las llamadas variables manifiestas) con un conjunto de varibles latentes (que son las variables que no son empíricamente observables).

INLA

Recientemente se ha implementado el método conocido en inglés como Integrated Nested Laplace Approximation(INLA) para estimar modelos del tipo "Latent Gaussian Markov Random Fields", en inglés [25]. Entre estos modelos que puede estimar INLA se encuentra el modelo BYM,

para el cual proporciona una buena aproximación a las distribuciones posteriores marginales de los parámetros en el modelo.

Entendemos por distribuciones posteriores marginales, denotándolas por $\pi(\theta_i|y)$, a aquellas distribuciones a posteriori que dependen de alguno de los parámetros del modelo. En nuestro caso, como queremos estimar el riesgo relativo en diferentes provincias, no nos interesa calcular la distribución a posteriori conjunta sino que es suficiente con calcular las marginales. De esta forma es más fácil hacer las aproximaciones o cálculos necesarios que si tomásemos la conjunta.

Además de proporcionar una aproximación de las marginales posteriores de los parámetros de un modelo, INLA puede calcular también criterios para la selección del modelo, como el DIC (*Deviance Information Criterion*, en inglés), que es el análogo al AIC de la estadística frecuentista, o el histograma PIT (*Probability Integral Transform*, en inglés), que sirve para detectar valores atípicos y para comparar y validar los modelos.

A continuación vamos a explicar qué aproximaciones usa INLA.

En primer lugar, supongamos que hemos observado n variables $y_i, i = 1, \ldots, n$ con una distribución (en nuestro caso Poisson) con media λ_i que está relacionada con un predictor lineal η_i a tavés de una función. A su vez, η_i se modeliza como sigue:

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

donde $f^{(j)}$ representa una función no lineal o efecto aleatorio (estamos suponiendo que hay n_f)en un conjunto de covariables $\mathbf{u}$, β_k son los coeficientes para los efectos lineales de un vector de covariables $\mathbf{z}$ y ϵ_i son los términos no estructurados. Los efectos latentes $\mathbf{x} = \{\{\eta_i\}, \alpha, \{\beta_k\}, \ldots\}$ se suponen normales con media 0 y matriz de precisión $\mathbf{Q}(\theta_1)$, donde θ_1 es vector de hiperparámetros. Por lo tanto, las observaciones tendrán una verosimilitud que dependerá de los efectos latentes $\mathbf{x}$ y de los parámetros θ_2. De aquí que las obervaciones y_i serán supuestas independientes dados $\mathbf{x}$ y θ_2.

Este tipo de modelos GMRF son los que puede estimar INLA. En particular, para los modelos espaciales (como es nuestro caso), las $f^{(j)}(u_{ji})$ representan los efectos aleatorios en una localización espacial i.

Se supone que el cuerpo latente tiene propiedades de independencia condicionales, lo que convierte a x en un GMRF.

Recordar que el objetivo de INLA es calcular las distribuciones a posteriori marginales. Éstas se pueden escribir como:

$$\pi(x_i|y) \propto \int \pi(x_i|\theta, y)\pi(\theta|y)d\theta$$

y

$$\pi(\theta_j|y) \propto \int \pi(\theta|y)d\theta_{-j},$$

donde θ_{-j} denota θ menos la componente j.

Las aproximaciones se harán de las integrales de los miembros de la derecha de las anteriores expresiones. Dichas aproximaciones serán posibles si la dimensión de θ es pequeña, pero esto es lo que suele suceder en la práctica. Notar que también se necesita una aproximación de $\pi(\theta|y)$.

De $\pi(\mathbf{x}, \theta, y) = \pi(x|\theta, y)\pi(\theta|y)\pi(y)$ se sigue que la marginal posterior $\pi(\theta|y)$ de los hiperparámetros se puede obtener usando la aproximación de Laplace:

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

donde $\tilde{\pi}_G(x|\theta, y)$ es la aproximación gaussiana (es decir, aproximamos la distribución de una variable no normal a una gaussiana uniendo la moda y la curvatura de la moda [24]) a la condicional de $\mathbf{x}$ y $x^*(\theta)$ es la moda de la condicional dado un valor de θ.

Por lo tanto, las marginales que nos interesan se pueden calcular usando la siguiente suma finita evaluada en los puntos θ_k :

$$\tilde{\pi}(x_i|y) = \sum_k \tilde{\pi}(x_i|\theta_k, y) \times \tilde{\pi}(\theta_k|y) \times \Delta_k,$$

donde Δ_k representa los pesos para cada vector de los valores de θ_k en la red, $\tilde{\pi}(x_i|\theta, y)$ y $\tilde{\pi}(\theta|y)$ representan las aproximaciones de $\pi(x_i|\theta, y)$ y $\pi(\theta|y)$, respectivamente [14].

Los puntos θ_k se pueden elegir de dos maneras: usando la estrategia "GRID" o bien usando "CCD" (Central Composite Design). La primera se recomienda si la dimensión del vector de los hiperparámetros, h, es media ($h = 6/12$), y lo que hace es para definir una red de puntos que cubre el área donde está la mayor parte de la masa de $\tilde{\pi}(\theta_k|y)$. La segunda estrategia, CCD, tiene su origen en el campo del diseño y consiste en colocar una pequeña cantidad de "puntos" en un espacio m-dimensional para estimar la curvatura de $\tilde{\pi}(\theta_k|y)$. Para más información ver [25]. Dicho artículo recomienda usar la estrategia CCD para problemas cuyo vector de hiperparámetros tiene una dimensión h grande, siendo $h > 12$, aproximadamente.

Para calcular la aproximación $\tilde{\pi}(x_i|\theta_k, y)$ existen tres formas posibles: aproximaciones gaussianas, totales de Laplace y simplificadas de Laplace. Cada una de ellas presenta características diferentes, así como tiempos de cálculo y precisión. Aunque la aproximación de Gauss es la más rápida computacionalmente, puede generar errores en la localización de la media posterior y/o errores debido a la falta de asimetría [24]. Por otra parte, la aproximación de Laplace es la más precisa, pero es muy costosa computacionalmente. Por último, la aproximación simplificada de Laplace es la más rápida para calcular y por lo general es lo suficientemente precisa.

Para concluir, INLA es una alternativa a los métodos MCMC (Markov Chain Monte Carlo) computacionalmente más rápida que proporciona resultados satisfactorios en un corto periodo

de tiempo computacional [14]. Esto se debe principalmente a que hace uso de aproximaciones de Laplace para obtener las distribuciones marginales a posteriori y utiliza computación en paralelo.

Veamos ahora la aplicación de INLA en nuestro caso particular. En primer lugar, realizaremos un análisis descriptivo donde estudiaremos cada periodo de estudio (1975-1991 y 1992-2008) de forma independiente, es decir, consideraremos como si tuviéramos dos estudios de diseño transversal distintos, uno para cada periodo.

4.3.2. Análisis del primer periodo (estandarización interna con tasas específicas de referencia del primer periodo)

Vamos a obtener las RME suavizadas para el periodo 1975-1991 y estandarizaremos según las tasas específicas por edad del mismo periodo. Como las tasas específicas se han calculado teniendo en cuenta toda España (estandarización interna), las RMEs tendrán como referencia la mortalidad de todo el país.

Creamos la matriz de datos necesaria para ejecutar el modelo. Seleccionamos los datos del primer periodo (1975-1991):

```
> obs <- observados[observados$periodo==1,]
> esp <- esperados.p1[esperados.p1$periodo==1,]
```

Para simplificar la notación, asignamos los valores observados y esperados a los objetos O y E, respectivamente. Además creamos el objeto "region" que necesitaremos para el análisis:

```
> O <- obs$casos
> E <- esp$total
> region <- 1:length(x.nb)
```

Finalmente, creamos un objeto "data.frame" con todos los datos necesarios, para ejecutar el primer modelo:

```
> Datos <- data.frame(region=region, region.struct=region, O=O, E=E)
```

Definimos la fórmula del modelo (formula1) y ejecutamos la función "inla" que estima el modelo BYM, comentado anteriormente. Así, obtendremos los riesgos suavizados para provincia. También obtendremos la probabilidad de que dichos riesgos sea mayor que 100 (probabilidad de exceso de riesgo) para cada provincia. Todos los resultados de este modelo se guardarán en el objeto "resultado.p1".

```
> formula1 <- O ~ f(region.struct,model="besag",graph=esp_prov_nb,
+                   param=c(1,0.001))+f(region,model="iid", param=c(1,0.01))
> resultado.p1<-inla(formula1, family="poisson", data=Datos, E=E,
+                    control.compute=list(dic=T, cpo=TRUE),
+                    control.predictor=list(compute=TRUE,cdf=c(log(1))),
+                    control.inla=list(strategy="laplace",int.strategy="grid"))
```

En *graph* hemos escrito la matriz de vecindad. El efecto de suavizado es especificado en "f()", siendo *model="iid"* para un efecto aleatorio i.i.d. y *"besag"* para un IGMRF (Intrinsic Gaussian Random Field). Luego estamos ante un modelo BYM ya que tenemos la suma de un "iid" más un "besag". Además con hyper=c(1,0.001) indicamos los parámetros a y b de la distribución a priori (LogGamma) del hiperparámetro (Log-Precision).Por defecto: a=1, b=0.001. Como *dic=TRUE* y *cpo=TRUE*, se calcularán el DIC y las medidas de validación CPO y PIT, respectivamente. Con *compute=TRUE* se calculan las marginales del predictor lineal $\log(\lambda_i)$ y con *cdf=c(log(1))* se calcula para el predictor lineal la "cumulative distribution function" (cdf), es decir, se calculan las probabilidades $\text{Prob}(\log(\lambda_i) < x(p))$. En este caso $x(p) = \log(1)$. Con *strategy=laplace* se utiliza la estrategia para obtener las aproximaciones de las distribuciones marginales a posteriori, como es "laplace" tiene mayor precisión aunque también más coste computacional que si hubiéramos usado "gaussian" o "simplified.laplace". Por último con "int.strategy" indicamos la estrategia de integración utilizada, puede ser "ccd", "grid" (más precisa pero con más coste computacional) y "eb" (empirical bayes).

Si quisiéramos ver todas las componentes que podemos obtener con inla basta con escribir:

```
> names(resultado.p1)
```

```
 [1] "names.fixed"                "summary.fixed"
 [3] "marginals.fixed"            "summary.lincomb"
 [5] "marginals.lincomb"          "size.lincomb"
 [7] "summary.lincomb.derived"    "marginals.lincomb.derived"
 [9] "size.lincomb.derived"       "mlik"
[11] "cpo"                        "model.random"
[13] "summary.random"             "marginals.random"
[15] "size.random"                "summary.linear.predictor"
[17] "marginals.linear.predictor" "summary.fitted.values"
[19] "marginals.fitted.values"    "size.linear.predictor"
[21] "summary.hyperpar"           "marginals.hyperpar"
[23] "internal.summary.hyperpar"  "internal.marginals.hyperpar"
[25] "si"                         "total.offset"
[27] "model.spde2.blc"            "summary.spde2.blc"
```

```
[29] "marginals.spde2.blc"          "size.spde2.blc"
[31] "misc"                         "dic"
[33] "mode"                         "neffp"
[35] "joint.hyper"                  "nhyper"
[37] "version"                      "Q"
[39] "graph"                        "cpu.used"
[41] ".args"                        "call"
[43] "model.matrix"
```

La ventaja de este modelo es que permite obtener las distribuciones a posteriori marginales de la suma del efecto espacial y el heterogéneo.

Los riesgos suavizados de cada área las obtenemos mediante la siguiente sintaxis:

```
> RMEs.p1 <- resultado.p1$summary.fitted.values$mean*100
```

Es decir, que las RMEs se muestran con su valor multiplicado por 100. Por consiguiente, aquellas provincias con valores de RMEs superiores a 100 indican un exceso de riesgo de mortalidad respecto a la mortalidad de la población española, mientras que valores inferiores evidencian un menor riesgo.

Para obtener la probabilidad de que los riesgos suavizados sea superior a 100 para cada área ejecutaremos la siguiente sintaxis:

```
> PRP.p1 <- 1-resultado.p1$summary.linear.predictor[,"0 cdf"]
```

Representaremos los riesgos suavizados mediante mapas de cortes fijos donde los colores verdes representan las áreas con mayor defecto de mortalidad, mientras que las áreas con mayor exceso de mortalidad son representadas en tonos marrones.

Podemos observar que tras la suavización, las zonas despobladas no presentan comportamientos anómalos y que las RME ahora no reflejan la distribución de la población en la región de estudio sino el riesgo estudiado.

En la Figura (4.4) se muestran los riesgos suavizados de mortalidad por cáncer de próstata para España (concretamente, la media de las distribuciones). Como ya hemos mencionado anteriormente observamos las provincias cuyos riesgos estimados estén por encima de 100, lo que indicará que tienen un exceso de riesgo de mortalidad respecto a la mortalidad de la población española, y las provincias que presenten riesgos menores de 100, lo que se traducirá como un menor riesgo respecto a la mortalidad de la población española. En este caso se observa que se dan más casos de mortalidad de cáncer de próstata, comparando con la media de España, en

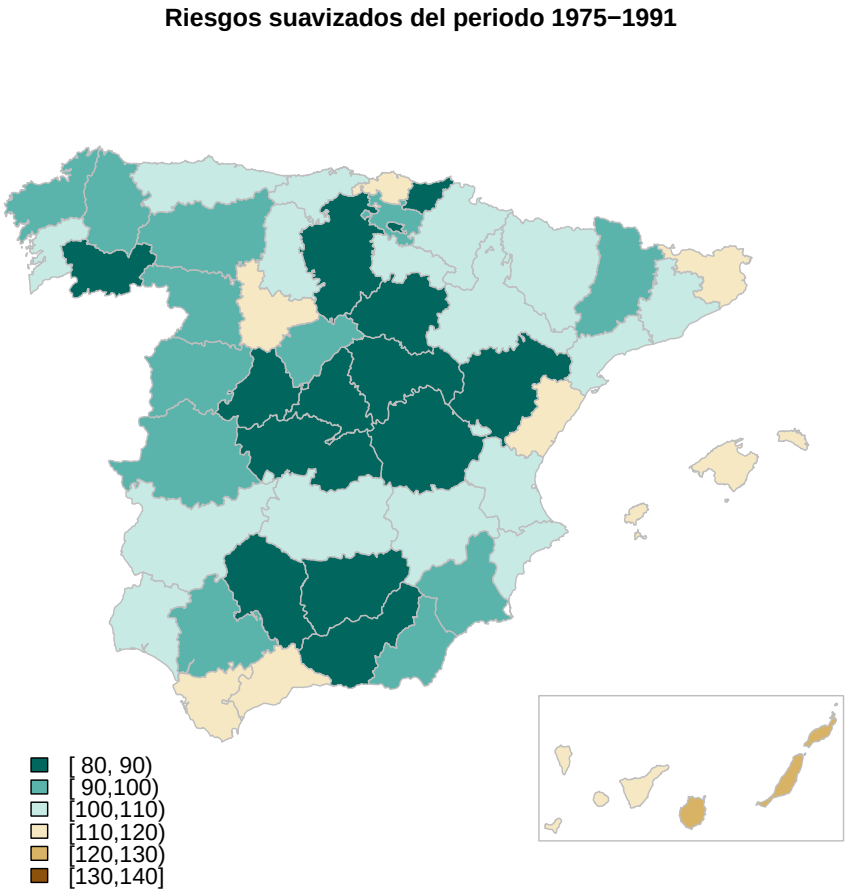

Figura 4.4: Riesgos suavizados para el periodo 1975-1991

las Palmas, seguido por Valladolid, Baleares, Santa Cruz de Tenerife y en algunas zonas de la costa como Castellón, Gerona, Vizcaya, Cádiz y Málaga. Las provincias de Pontevedra, Asturias, Cantabria, Palencia, La Rioja, Navarra, Huesca, Zaragoza, Barcelona, Tarragona, Valencia, Alicante, Albacete, Ciudad Real, Badajoz y Huelva tienen un riesgo muy poco por encima de 100 (entre 100 y 110), luego tienen un mayor riesgo que en España en global. En cambio puede observarse un patrón dominado por las provincias ubicadas en el interior (Ávila, Burgos, Soria, Madrid, Guadalajara, Cuenca, Toledo, Teruel, Córdoba, Jaén, etc) donde los riesgos son menores que en España en su totalidad, ya que recordar que el valor 100 indica que el riesgo de la provincia es como el de España.

Si comparamos la Figura (4.4) con la Figura (3.1), realizada en el capítulo 3, vemos que prácticamente los mapas que representan la SMR y los riesgos suavizados son muy parecidos, aunque se encuentran diferencias en las provincias de Soria (42), Ávila (5), Cuenca (16) y Jaén (23) que tienen una SMR entre 70 y 80, mientras que sus riesgos suavizados varían entre 80 y

90. Cantabria tiene una SMR entre 110 y 120, mientras que sus riesgos suavizados varían entre 100 y 110.Esto se debe a que estas provincias no son muy pobladas o tienen pocos casos de mortalidad por cáncer. En efecto, si calculamos con **R** el número de observados y esperados para estas provincias en el primer periodo obtenemos:

```
Ávila: nº de observados: 359 , nº de esperados es: 463.3516

Población total de Ávila en el año 1991: 87608

Cuenca: nº de observados: 446 , nº de esperados es: 568.3522

Población total de Cuenca en el año 1991: 102195

Soria: nº de observados: 235 , nº de esperados es: 301.9957

Población total de Soria en el año 1991: 47149

Jaén: nº de observados: 854 , nº de esperados es: 1076.314

Población total de Jaén en el año 1991: 315168
```

Con el objeto de cuantificar la evidencia estadística que proporcionan las estimaciones del riesgo en cada provincia crearemos mapas de la probabilidad de exceso de riesgo (RMEs>100). Para representar estas probabilidades en el mapa consideraremos la siguiente segmentación: (0,1; 0,2; 0,8 y 0,9). Utilizaremos tonalidades verdes para los riesgos relativos con baja probabilidad de ser superior a 100, rojas para los riesgos relativos con alta probabilidad (áreas con exceso de riesgo, es decir, provincias donde el riesgo es mayor que el riesgo global de España) y el rango intermedio lo representaremos en amarillo. Consideraremos que una probabilidad es alta cuando sea mayor que 0.8.

El mapa representado en la Figura (4.5) representa las probabilidades de las RMEs en el periodo 1975-1991. Veamos en él, qué probabilidades tienen las provincias que en el mapa representado en la Figura (4.4) presentaban un riesgo estimado por encima de 100 (que eran entre otras: las Palmas, Santa Cruz de Tenerife, Valladolid, Baleares, Castellón, Gerona, Vizcaya, Cádiz, Málaga, Pontevedra, Cantabria, Asturias, Palencia, Zaragoza, Barcelona, Valencia, Alicante, Albacete, Huelva y Badajoz). Observamos que todas estas provincias tienen una probabidad superior a 0.9, lo que significa que con una alta probabilidad el riesgo de mortalidad por cáncer de próstata de estas provincias está por encima del riesgo global de España. Además con alta probabilidad, entre 0.8 y 0.9, tenemos también Navarra.

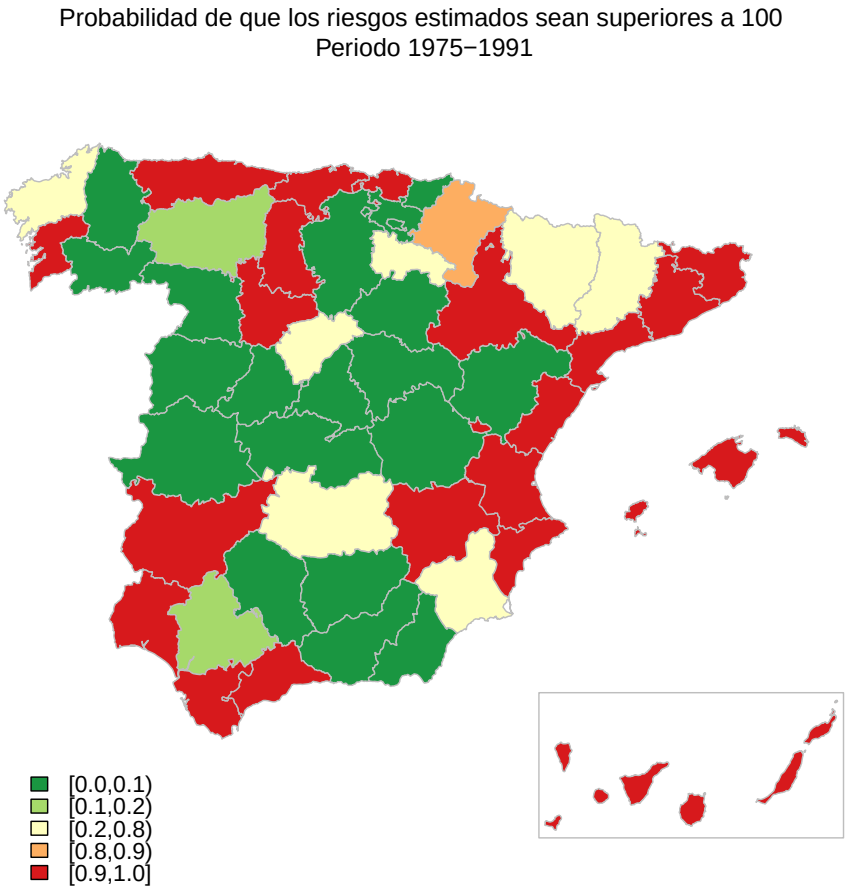

Figura 4.5: Probabilidad de los riesgos estimados del periodo 1975-1991 sean superiores a 100

Para observar la forma de la distribución de la RMEs representaremos su función de densidad en la Figura (4.6), ya que bajo la filosofía Bayesiana no se obtiene una estimación puntual del riesgo sino toda la distribución de probabilidad. En su interior se muestran los colores correspondientes a los grupos de valores utilizados en la representación geográfica de las RMEs. La función de densidad permite apreciar con detalle el rango de valores de la RMEs, el cual no es visible en el mapa.

La forma de la función de densidad de las RMEs es como la de una distribución normal de media en torno a 115 pero con dos máximos, uno local y otro absoluto. También vemos que la cola derecha de esta función es más larga que la de la izquierda.

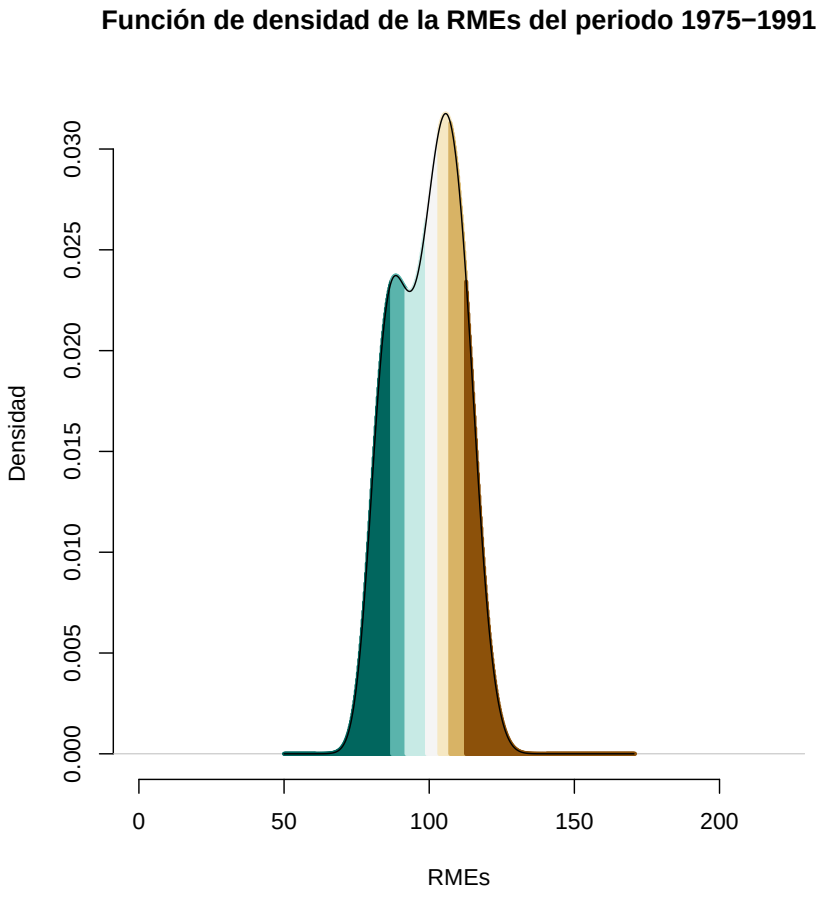

Figura 4.6: Función de densidad de la RMEs del periodo 1975-1991

4.3.3. Análisis del segundo periodo (estandarización interna con tasas específicas de referencia del segundo periodo)

De forma totalmente análoga debemos proceder al análisis del segundo periodo (1992-2008). Seleccionamos los datos del segundo periodo (1992-2008):

```
> obs <- observados[observados$periodo==2,]
> esp <- esperados.p2[esperados.p2$periodo==2,]
```

Seleccionamos la causa de mortalidad que vamos a analizar, en este caso (total). Para simplificar la notación, asignamos los valores observados y esperados a los objetos O y E respectivamente. Además creamos el objeto "region" que necesitaremos para el análisis:

```
> O <- obs$casos
> E <- esp$total
> region <- 1:length(x.nb)
```

Finalmente, creamos un objeto "data.frame" con todos los datos necesarios, para ejecutar este segundo modelo:

```
> Datos <- data.frame(region=region, region.struct=region, O=O, E=E)
```

Para este modelo utilizaremos la misma fórmula que en la sección anterior (formula1) y ejecutaremos la función "inla". Así, obtendremos los riesgos suavizados para cada provincia. También obtendremos la probabilidad de que dichos riesgos sea mayor que 100 (probabilidad de exceso de riesgo) para cada provincia. Todos los resultados de este modelo se guardarán en el objeto "resultado.p2".

```
> formula1 <- O ~ f(region.struct,model="besag",graph=esp_prov_nb,
+                   param=c(1,0.001))+f(region,model="iid",param=c(1,0.01))
> resultado.p2<-inla(formula1, family="poisson", data=Datos, E=E,
+                    control.compute=list(dic=T, cpo=TRUE),
+                    control.predictor=list(compute=TRUE, cdf=c(log(1))),
+                    control.inla=list(strategy="laplace", int.strategy="grid"))
```

Los riesgos suavizados de cada área las obtenemos mediante la siguiente sintaxis:

```
> RMEs.p2 <- resultado.p2$summary.fitted.values$mean*100
```

Para obtener la probabilidad de que los riesgos suavizados sea superior a 100 para cada área ejecutaremos la siguiente sintaxis:

```
> PRP.p2 <- 1-resultado.p2$summary.linear.predictor["0 cdf"]
```

Representamos gráficamente los resultados del segundo periodo 1992-2008, estandarizados (internamente) según las tasas específicas del mismo periodo 1992-2008, siguiendo el mismo esquema que en el apartado anterior.

Análogamente a lo anterior, construimos los intervalos para asignar distintos colores y poder observar qué provincias tienen un riesgo por encima de 100 (representadas por colores marrones) y cuales por debajo de 100 (representadas en colores verdes).

A la vista del mapa representado en la Figura (4.7) observamos que las provincias de las Palmas, Castellón, Huesca, La Rioja, Asturias, La Coruña, Pontevedra y Orense tienen un

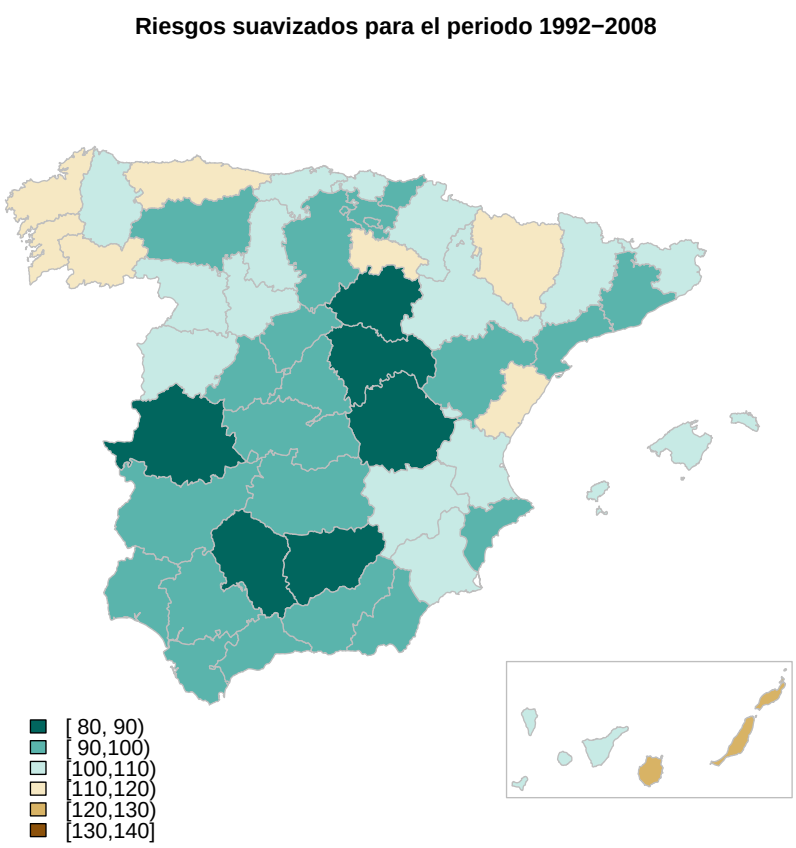

Figura 4.7: Riesgos suavizados para el periodo 1992-2008

riesgo de mortalidad por encima de 100, es decir, que el riesgo de morir por cáncer de próstata en esas provincias es mayor que el riesgo de España. También podríamos citar las provincias de Lugo, Cantabria, Vizcaya, Navarra, Zaragoza, Lérida, Gerona, Baleares, Valencia, Albacete, Murica, Santa Cruz de Tenerife, Palencia, Valladolid, Zamora, Salamanca, las cuales presentan un riesgo por encima de 100 (en particaular, entre 100 y 110).

Si comparamos la Figura (4.7) con la Figura (3.2), realizada en el capítulo 3, vemos que,al igual que para el periodo 1975-1991, prácticamente los mapas que representan la SMR y los riesgos suavizados son iguales, sólamente se encuentran diferencias en la provincia de Las Palmas (35) que tiene una SMR entre 130 y 140, mientras que sus riesgos suavizados varían entre 120 y 130. Esto puede deberse a que esta provincia tiene pocos casos de mortalidad por cáncer. Veamos cuántos casos de mortalidad tiene Las Palmas y que población tiene en el año 2008.

```
Las Palmas: nº de observados: 1663 , nº de esperados es: 1274.647
```

```
Población total de las Palmas en el año 2008: 535774
```

A continuación veremos con qué probabilidad podemos asegurar estas afirmaciones. Prácticamente las provincias restantes presentan un riesgo menor de 100, es decir, que evidencian un riesgo menor respecto a la mortalidad de la población española.

Análogamente al primer periodo, realizamos un mapa para el segundo periodo (1992-2008) que indique la probabilidad de que el riesgo estimado sea superior a 100. Para ello seguiremos usando los mismos colores que anteriormente, es decir, que utilizaremos tonalidades verdes para los riesgos relativos con baja probabilidad de ser superior a 100, rojas para los riesgos relativos con alta probabilidad y el rango intermedio lo representaremos en amarillo.

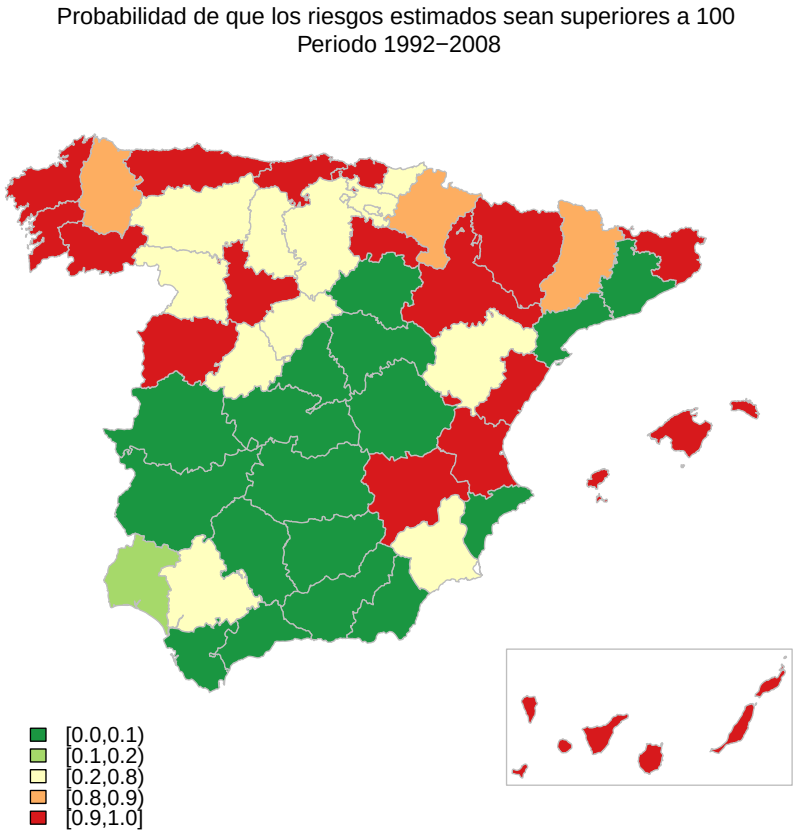

Figura 4.8: Probabilidad de que los riesgos estimados del periodo 1992-2008 sean superiores a 100

El mapa de las probabilidades de los riesgos suavizados en el segundo periodo que comprende los años 1992-2008, Figura (4.8), muestra las provincias de La Coruña, Pontevedra, Orense, Asturias, Cantabria, Vizcaya, La Rioja, Huesca, Zaragoza, Gerona, Baleares, Castellón, Valencia, Albacete, Valladolid, Salamanca, Santa Cruz de Tenerife y Las Palmas, que presentaban un exceso de riesgo de mortalidad por cáncer respecto a la mortalidad por cáncer de la España tienen una probabilidad por encima de 0.9, es decir, que es "muy probable" que las personas de

estas provincias tengan más riesgo de morir por cáncer de próstata que en España en global. Con una probabilidad de entre 0.8 y 0.9 están las provincias de Lugo, Navarra y Lérida, que también mostraban un riesgo de mortalidad mayor de 100 en el mapa representado en la Figura (4.7). Por lo tanto las provincias de Murcia, Palencia y Zamora que antes mostraban un riesgo de mortalidad por encima de 100, observamos que en este mapa tienen una probabilidad de entre 0.2 y 0.8, luego como estamos considerando que una probabilidad es alta si está por encima de 0.8, podemos decir que estás provincias tienen un riesgo alto comparado con el riesgo de España con una probabilidad baja.

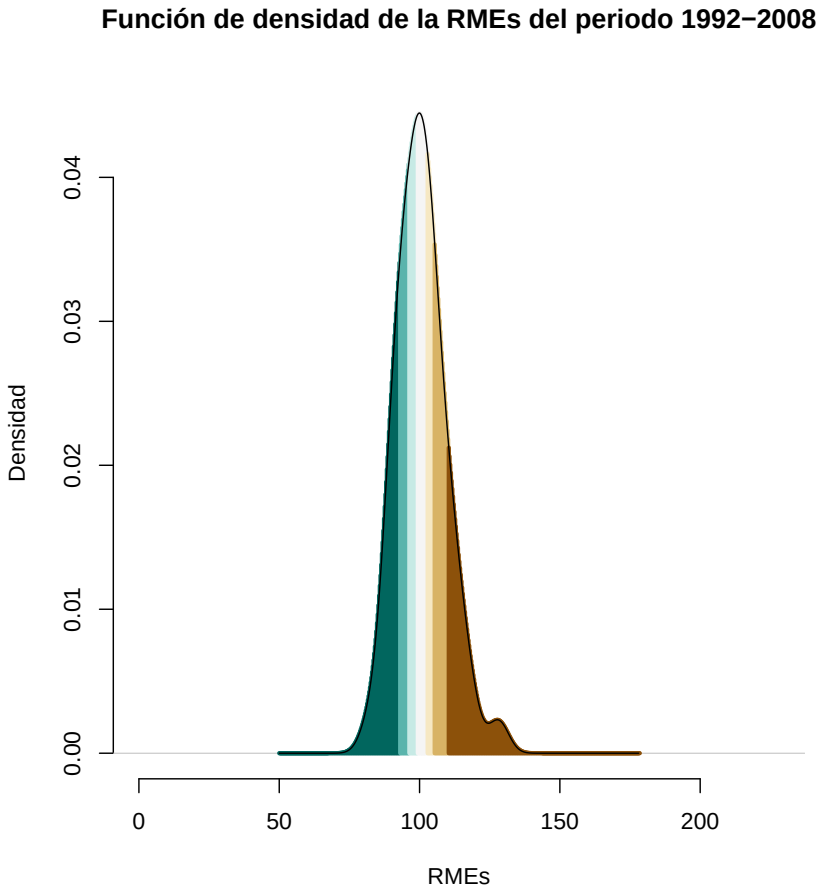

Figura 4.9: Función de densidad de la RMEs del periodo 1992-2008

La función de densidad para el segundo periodo se ha representado en la Figura (4.9). Al igual que la función para el primer periodo, también presenta dos máximos, uno local y otro global. La diferencia es que ahora el local es mucho menor que el global y su coordenada en el eje horizontal se encuentra a la derecha del global.

4.3.4. Análisis del segundo periodo (estandarización interna con tasas específicas de referencia del primer periodo)

Vamos a obtener los riesgos suavizados para el periodo 1992-2008, a continuación estandarizaremos según las tasas específicas por edad del periodo anterior 1975-1991. Recordar que las tasas específicas las calculamos teniendo en cuenta todo el país de España (estandarización interna), así las RMEs tendrán como referencia la mortalidad en toda España en el periodo anterior.

En primer lugar, creamos la matriz de datos necesaria para ejecutar el modelo, seleccionando los valores observados del segundo periodo y los valores esperados para el segundo periodo estandarizados según las tasas específicas del primer periodo. De este modo las RMEs obtenidas estarán en referencia a la mortalidad media en el primer periodo y podremos ver si la mortalidad ha aumentado o disminuido en cada área, ya que los mapas que hemos representado en la s Figuras (4.4) y (4.7) no se pueden comparar.

```
> obs <- observados[observados$periodo==2,]
> esp <- esperados.p1[esperados.p1$periodo==2,]
```

Seleccionamos la causa de mortalidad que vamos a analizar, en este caso "total". Para simplificar la notación, asignamos los valores observados y esperados a los objetos O y E respectivamente. Además creamos el objeto "region" que necesitaremos para el análisis:

```
> O <- obs$casos
> E <- esp$total
> region <- 1:length(x.nb)
```

Finalmente, creamos un objeto "data.frame" con todos los datos necesarios, para ejecutar el primer modelo:

```
> Datos <- data.frame(region=region, region.struct=region, O=O, E=E)
```

Para este modelo utilizaremos la misma fórmula que en las secciones anteriores (formula1) y ejecutaremos la función "inla". Así, obtendremos los riesgos suavizados para cada área. Todos los resultados de este modelo se guardarán en el objeto "resultado.p2.resp.p1".

```
> formula1 <- O ~ f(region.struct,model="besag",graph=esp_prov_nb,
+                   param=c(1,0.001))+f(region,model="iid", param=c(1,0.01))
> resultado.p2.resp.p1<-inla(formula1, family="poisson", data=Datos,E=E,
+                            control.compute=list(dic=T, cpo=TRUE),
+                            control.predictor=list(compute=TRUE, cdf=c(log(1))),
+                            control.inla=list(int.strategy="grid"))
```

Los riesgos suavizados de cada área las obtenemos mediante la siguiente sintaxis:

```
> RMEs.p2.resp.p1 <- resultado.p2.resp.p1$summary.fitted.values$mean*100
```

Los riesgos suavizados del segundo periodo (1992-2008) teniendo en cuenta las tasas del primer periodo se representan en la Figura (4.10). En ella se representan los mapas de cortes fijos donde los colores verdes representan las áreas con mayor defecto de mortalidad, mientras que los tonos marrones representan las áreas con mayor exceso. Recordemos que los valores de los riesgos suavizados están en referencia a la media de la mortalidad en todo el país para el primer periodo (1975-1991).

```
> Cortes.fijos <- cut2(RMEs.p1, g=7, onlycuts=T)
> Cortes.fijos[1] <- min(RMEs.p1,RMEs.p2.resp.p1)
> Cortes.fijos[length(Cortes.fijos)] <- max(RMEs.p1,RMEs.p2.resp.p1)
```

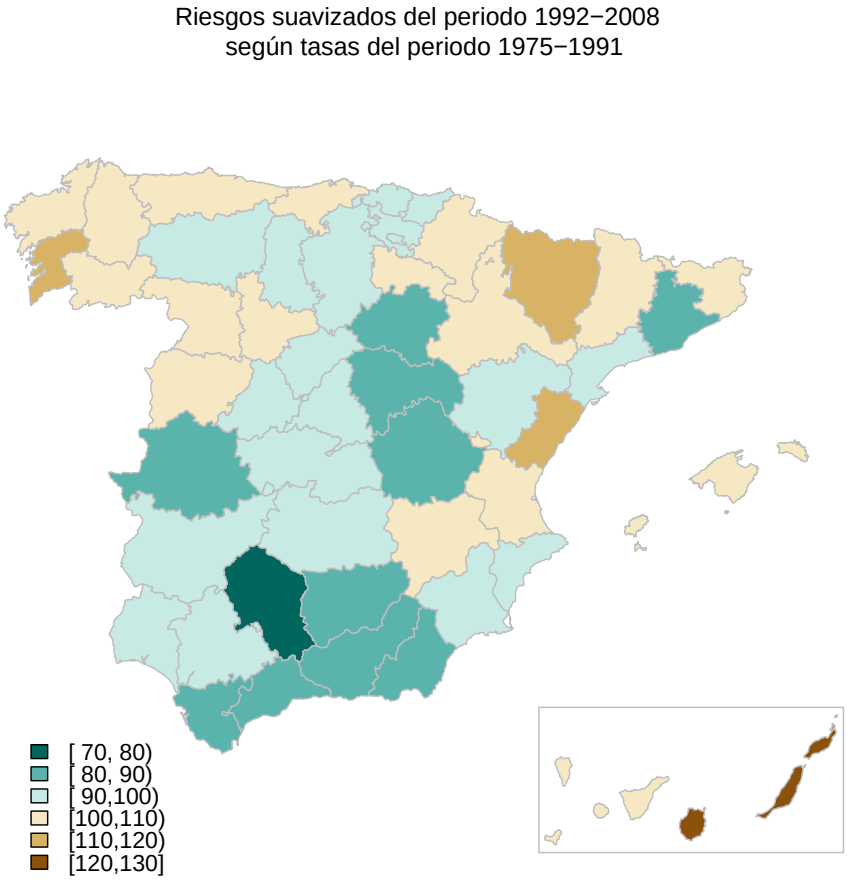

Figura 4.10: Riesgos suavizados para el periodo 1992-2008 según las tasas del periodo 1975-1991

Notar que en la Figura (4.10) donde se han representado los riesgos suavizados del segundo periodo (1992-2008) según las tasas del primer periodo, se han fijado los cortes según los cortes fijos de los riesgos suavizados del primer periodo, ver Figura (4.4). De este modo se puede comparar la tendencia en la distribución de los riesgos estimados entre los dos periodos analizados. Se utiliza la misma gama de colores divergente que la utilizada en la representación geográfica de los riesgos suavizados de las figuras anteriores. En esta figura observamos que Las Palmas, Huesca, Castellón y Pontevedra presentan el mayor exceso de riesgo de mortalidad (concretamente mayor de 110), mientras que las provincias de Lugo, Orense, La Coruña, Asturias, Cantabria, Navarra, La Rioja, Zaragoza, Lérida, Gerona, Baleares, Valencia, Albacete, Valladolid, Zamora, Salamanca y Santa Cruz de Tenerife tienen también un exceso de riesgo pero en menor medidad (concretamente entre 100 y 110).

Si comparamos la Figura (4.10) con la Figura (3.3) observamos que son exactamente iguales. Esto puede deberse al aumento de la población o al descenso del número de casos de mortalidad

por cáncer de próstata. Veamos, por ejemplo, como ha variado la población de cada provincia. Para ello mostraremos como ejemplo la población en el año 1991 y en el año 2008 en las Tablas 4.1 y 4.2. Estas tablas se han realizado usando la siguiente orden en **R**:

```
> poblaciontotal<-aggregate(mort$pob,by=list(provincia=mort$provincia,
+                                            agno=mort$agno),FUN=sum,na.rm=T)
> casostotal<-aggregate(mort$casos,by=list(provincia=mort$provincia,
+                                          agno=mort$agno),FUN=sum,na.rm=T)
```

En estas tablas aparecenen azul las provincias que han aumentado su población, como puede verse la mayoría de ellas la ha incrementado.

A continuación se han realizado las tablas 4.3 y 4.4, donde se muestran los casos de mortalidad por cáncer en el año 1991 y en el año 2008 por provincias. Se han señalado con azul las provincias cuya mortalidad por cáncer ha disminuido. Observamos que el descenso se ha producido en un número muy pequeño de provincias. No obstante, en general las provincias que en el año 2008 tienen más casos de mortalidad que en el año 1991 han aumentado su población de forma más significativa que el número de casos de muertes. Cabe destacar también, que las provincias de Ávila, León y Lugo presentan en el año 2008 menos casos de mortalidad que en 1991, asi como número de habitantes. En cambio, Asturias tiene menor número de habitantes (ha pasado de 527875 a 506429 habitantes) y mayor número de casos (en 1991 tenía 133 y en 2008 tenía 196). Eso mismo les ha pasado a las provincias de Orense, Palencia, Salamanca, Soria, Vizcaya y Zamora. Sin embrago, a Barcelona, Navarra y Valladolid les ha ocurrido lo contrario, han aumentado su población pero han logrado reducir el número de muertes por cáncer.

provincia	pob. año 1991	pob. año 2008
Álava	135846	152290
Albacete	170574	197900
Alicante	635766	934825
Almería	227068	348428
Asturias	527875	506429
Ávila	87608	85107
Badajoz	321359	333706
Barcelona	2267372	2619950
Burgos	177000	184488
Cáceres	203664	201717
Cádiz	537052	597355
Cantabria	257997	280573
Castellón	221145	293266
Ciudad Real	232734	256289
Córdoba	369279	385648
Coruña	528875	538275
Cuenca	102195	108062
Girona	253635	364375
Granada	388961	443619
Guadalajara	74240	118748
Guipúzcoa	332616	339503
Huelva	219204	248844
Huesca	104774	112492
Islas Baleares	351142	530773
Jaén	315168	324254

Tabla 4.1: Población total según provincias y periodos. Parte 1.

provincia	pob. año 1991	pob. año 2008
Las Palmas	385546	535774
León	257647	235557
Lleida	176165	215834
Lugo	188022	168060
Madrid	2389806	3031261
Málaga	571728	764375
Murcia	517273	725827
Navarra	258524	304285
Ourense	170094	157489
Palencia	92003	84218
Pontevedra	430940	454793
Rioja	130613	157673
Salamanca	174739	169100
Segovia	73647	81527
Sevilla	797993	903410
Soria	47149	47142
Sta. Cruz de Tenerife	358889	495613
Tarragona	270012	396502
Teruel	72134	74997
Toledo	245004	327341
Valencia	1034610	1237337
Valladolid	243236	255239
Vizcaya	564772	551484
Zamora	105214	96242
Zaragoza	410467	463864

Tabla 4.2: Población total según provincias y periodos. Parte 2.

provincia	casos año 1991	casos año 2008
Álava	21	39
Albacete	39	65
Alicante	137	181
Almería	40	52
Asturias	133	196
Ávila	33	29
Badajoz	75	92
Barcelona	583	524
Burgos	46	69
Cáceres	52	63
Cádiz	81	114
Cantabria	69	69
Castellón	70	74
Ciudad Real	59	71
Córdoba	68	89
Coruña	129	161
Cuenca	28	45
Girona	75	88
Granada	70	75
Guadalajara	32	36
Guipúzcoa	80	112
Huelva	36	63
Huesca	47	54
Islas Baleares	102	115
Jaén	51	95

Tabla 4.3: Casos de mortalidad según provincias y periodos. Parte 1.

provincia	casos año 1991	casos año 2008
Las Palmas	71	113
León	98	95
Lleida	56	62
Lugo	75	71
Madrid	432	544
Málaga	97	147
Murcia	107	172
Navarra	84	83
Ourense	79	91
Palencia	27	32
Pontevedra	109	130
Rioja	43	53
Salamanca	51	56
Segovia	21	28
Sevilla	111	201
Soria	14	21
Sta. Cruz de Tenerife	81	110
Tarragona	61	70
Teruel	26	31
Toledo	60	106
Valencia	220	253
Valladolid	58	57
Vizcaya	98	162
Zamora	37	45
Zaragoza	132	147

Tabla 4.4: Casos de mortalidad según provincias y periodos. Parte 2.

No obstante, hay que calcular la probabilidad de que los riesgos estimados del segundo periodo según las tasas del primer periodo sean mayores que 100 para cada provincia. Para ello ejecutamos la siguiente sintaxis en **R**:

```
> PRP.p2.resp.p1<-1-resultado.p2.resp.p1$summary.linear.predictor[,"0 cdf"]
```

En la Figura (4.11) se representa la probabilidad de que los riesgos estimados del segundo periodo según las tasas del primer periodo sean mayores que 100 para cada provincia, siguiendo las tonalidades, las segmentaciones y las interpretaciones que en las Figuras (4.5) y (4.8). Se observa que las provincias de las Palmas, Baleares, Castellón, Huesca, Zaragoza, la Rioja, Valladolid, Asturias, la Coruña, Pontevedra y Orense, que en la Figura (4.10) tenían un riesgo suavizado por encima de 100, resultan tener una probabilidad de que el riesgo estimado sea mayor que 100, mayor de 0.9. Es decir, que estas provincias han incrementado su riesgo de que sus habitantes mueran por cáncer de próstata en el periodo de 1992-2008 frente al periodo 1975-1991. También hay que destacar que Gerona, Cantabria, Salamanca y Lugo tienen una probabilidad de que el riesgo estimado sea mayor que 100, mayor que 0.8 y menor que 0.9. Por tanto, estas provincias también han incrementado su riesgo de morir por cáncer en el segundo periodo frente al primero. Respecto a las provincias de Navarra, Lérida, Valencia, Albacete y Zamora, que ya hemos visto que tenían un exceso de riesgo en comparación con el riesgo de España, vemos en este mapa que tienen una probabilidad baja (menor de 0.8), así que no podemos asegurar que su riesgo sea mayor comparado con el de España en global en el periodo anterior. Por lo tanto, las provincias que han incrementado su riesgo de que sus habitantes mueran por cáncer de próstata en el primer periodo frente al segundo periodo son distintas desde el punto de vista bayesiano que desde el punto de vista frecuentista (calculadas en el capítulo 3). Concretamente, Santa Cruz de Tenerife, Baleares, Valencia, Albacete, Lérida, Navarra y Zamora tienen una SMR entre 100 y 110, pero no se puede asegurar que sus riesgos suavizados sean mayores que 100 ya que no tienen una probabilidad mayor que 0.8.

En resumen, el mapa representado en la Figura (4.11) muestra que el riesgo de mortalidad por cáncer de próstata se ha incrementado en el segundo periodo respecto del primero en la mitad norte de España. Posiblemente sea como resultado de un incremento mayor de la esperanza de vida en la mitad norte que en la mitad sur de España en los últimos años [26] y un aumento de la población, ya que como hemos comentado en el primer capítulo, uno de los principales factores de riesgo de padecer cáncer de próstata era la edad.

En la Figura (4.12) representamos la función de densidad de los riesgos suavizados del periodo 1975-1991 y del periodo 1992-2008, estandarizadas según las tasas específicas del periodo 1975-1991 (estandarización interna). El eje de las y's está en escala logarítmica. En esta gráfica observamos que, para los riesgos suavizados por encima de 100, la funcion de densidad del primer periodo limita un área mayor que la función de densidad del segundo periodo.

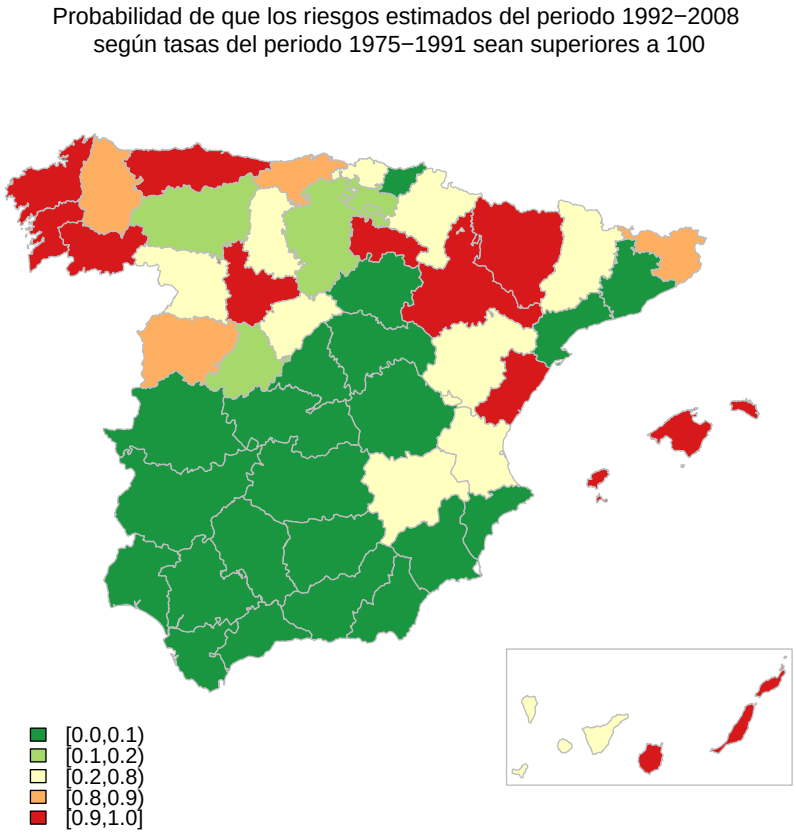

Figura 4.11: Probabilidad de que los riesgos estimados del periodo 1992-2008 según tasas del periodo 1975-1991 sean superiores a 100

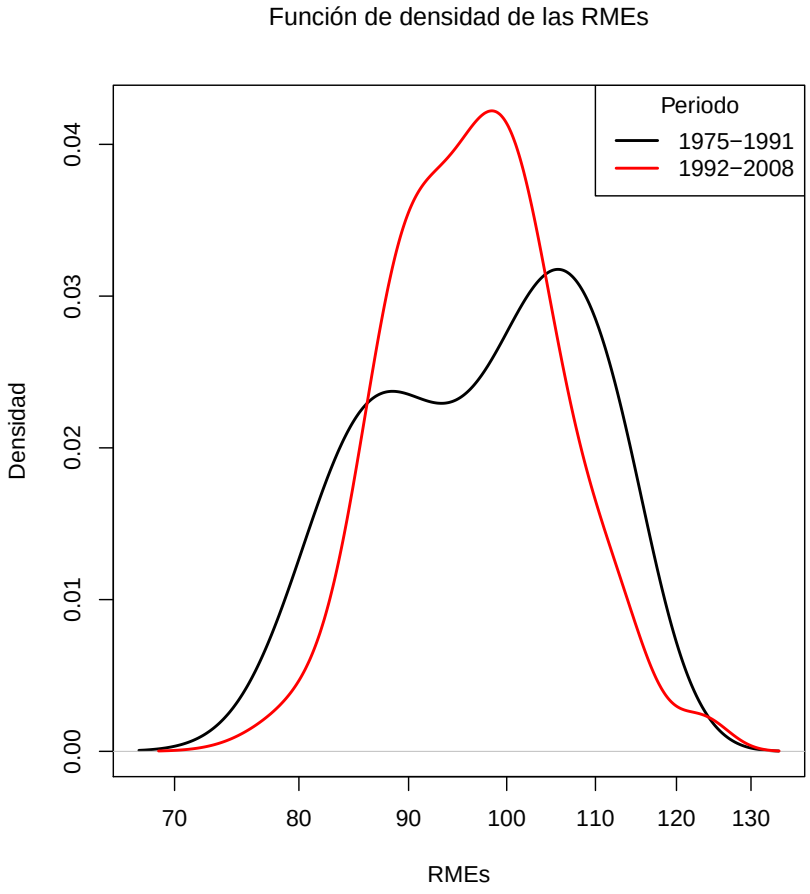

Figura 4.12: Función de densidad de las RMEs calculadas durante el periodo 1975-1991 y el periodo 1992-2008

Capítulo 5

Conclusiones

En este proyecto se han estimado los riesgos de mortalidad por cáncer de próstata en España durante dos periodos de tiempo: 1975-1991 y 1992-2008. Para ello se ha estimado un modelo estadístico espacial propuesto por Besag, York y Mollié [12]. Este modelo suaviza los riesgos estimados teniendo en cuenta la dependencia espacial (provincias lindantes tendrán riesgos parecidos debido a un estilo de vida, alimentación, etc, similares) y la dependencia no estructurada (factores de riesgo que no dependen del espacio geográfico y que influyen en la mortalidad). Debido a la complejidad de este modelo jerárquico bayesiano lo vamos a estimar mediante el método INLA ("Integrated Nested Laplace Approximation", en inglés).

El método INLA se publicó por primera vez en la prestigiosa revista "Journal of the Royal Statistical Society: Series B" en el año 2009 [25], lo que indica que es un método muy novedoso y actualmente en auge. INLA es un instrumento que ofrece una serie de aproximaciones de las marginales posteriores de los parámetros de un modelo jerárquico bayesiano en el cual los efectos o variables latentes son un campo gausiano markoviano aleatorio ("Gaussian Markov Random Field").

Además es un método computacionalmente más rápido que los "Markov Chain Monte Carlo" (MCMC). Cabe destacar que la llegada de los MCMC supuso un gran avance en la aplicación de la estadística bayesiana, ya que permitieron aproximar integrales de problemas que hasta entonces eran impensables de resolver. Es decir, INLA es la última y más potente herramienta de la estadística bayesiana.

El análisis espacial realizado en este trabajo ha mostrado que comparando el segundo periodo de tiempo (desde 1992 hasta 2008) con el primero (desde 1975 hasta 1991), el riesgo de mortalidad por cáncer de próstata se ha incrementado en el segundo periodo respecto del primero en la mitad norte de España. Concretamente han sido las provincias de las Palmas, Baleares, Castellón, Huesca, Zaragoza, la Rioja, Valladolid, Asturias, la Coruña, Pontevedra, Orense, Gerona,

Cantabria, Salamanca y Lugo. Estos resultados son consecuencia de un aumento de la población en la mayoría de las provincias en el segundo periodo frente al primero. Posiblemente otra causa pueda ser un incremento mayor de la esperanza de vida en la mitad norte que en la mitad sur de España en los últimos años, debido a que uno de los principales factores de riesgo de padecer cáncer de próstata es la edad.

Además si se comparan los resultados de la estadística frecuentista o clásica frente a los obtenidos con la estadística bayesiana, se observa que las provincias cuyo riesgo de mortalidad se ha incrementado en el segundo periodo respecto del primero, citadas en el párrafo anterior, son las mismas aplicando la estadística bayesiana o la frecuentista. Sin embargo, con la estadística clásica se obtiene además que las provincias de Santa Cruz de Tenerife, Baleares, Valencia, Albacete, Lérida, Navarra y La Rioja presentan un incremento de riesgo en el segundo periodo respecto al primero. Estas diferencias se deben a que estas provincias tienen riesgos menos estables, es decir, más variables y por eso con la estadística bayesiana no aparecen ya que ésta suaviza los riesgos.

La importancia de estos resultados reside en su aplicación, ya que permiten establecer por parte de epidemiólogos y profesionales de la salud pública planes de salud adecuados para eliminar posibles desigualdades en la salud, y pueden proporcionar información acerca de los posibles factores de riesgo que determinan esta enfermedad.

Bibliografía

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

[2] Rue, H., Martino, S., and Lindgren, F., (2009). *INLA: Functions which allow to perform a full Bayesian analysis of structured (geo-)additive models using Integrated Nested Laplace Approximaxion*. R package version 0.0.

[3] Ugarte, M.D., Militino, A.F., and Arnholt. A. *Probability and Statistics with R*. CRC Press, 2008.

[4] Página web de atención primaria en la red: `http://www.fisterra.com/guias-clinicas/cancer-prostata/#1751`.

[5] Cabanes, A., Pérez-Gómez, B., Aragonés, N., Pollán, M., López-Abente, G. *La situación del cáncer en España, 1975-2006*. Instituto de Salud Carlos III. Madrid, 2009.

[6] Heidenreich A, Aus G, Bolla M, et al. *EAU guidelines on prostate cancer*. Eur Urol 2008; 53:68-80.

[7] Schaid, D. J. *Pooled genome linkage scan of aggressive prostate cancer: results from the International Consortium for Prostate Cancer Genetics*, Human Genetics. November 2006, Volume 120, Issue 4, pp 471-485.

[8] Zelefsky, M. J., Eastham, J. A., Sartor, O. A., Kantoff, P. (2008). *Cancer of the prostate*. In V.T.DeVita, T. S. Lawrence, S. A. Rosenberg (Eds.), Cancer: Principles Practice of Oncology (pp. 1393-1452). Philadelphia: Lippincoot Williams and Wilkins.

[9] Gulley, J. L.; Arlen, PM; Bastian, A; Morin, S; Marte, J; Beetham, P; Tsang, KY; Yokokawa, J et al. (2005). *Combining a Recombinant Cancer Vaccine with Standard Definitive Radiotherapy in Patients with Localized Prostate Cancer*. Clinical Cancer Research 11 (9): 3353-62.

[10] Página web de la Sociedad Americana de Cáncer: `http://www.cancer.org/index`.

[11] Marí-Dell'Olmo, M., Gotseus, M., Palència, L., y el Grupo de Metodología del Proyecto Medea. *Estudio de tendencias. Manual para el análisis de datos. Desigualdades socioeconómicas y medioambientales en la mortalidad en ciudades de España*. Medea.

[12] Besag, J., York, J., Mollié, A. (1991). *Bayesian image restoration with two applications in spatial statistics*. Annals of the Institute of Statistical Mathematics, 43, 1-59.

[13] Ugarte, M.D., Militino, A.F. *Estadística Aplicada con S-PLUS. 2ª Edición Revisada*. Universida Pública de Navarra, 2002.

[14] Schrödle, B., Held, L., 2011. *A primer on disease mapping and ecological regression using INLA*. Comput Stat, 26, 241-258. DOI 10.1007/s00180-010-0208-2.

[15] Lawson, A.B. *Bayesian Disease Mapping, Hierarchical Modeling in Spatial Epidemiology*. Chapman Hall/CRC, 2009.

[16] Gelman, A.; Carlin, J. B.; Stern, H. S. and Rubin, D. B.*Bayesian data analysis (2ndedition)*. Chapman Hall/CRC, 2004.

[17] Lee, P. M.*Bayesian Statistics: An Introduction(3rd. ed.)*. Edward Arnold , 2004.

[18] Bernardo, J. M. and Smith, A. F. M. *Bayesian Theory*. Chichester, UK: Wiley, 1994.

[19] O'Hagan, A., 1994. *Kendall's advanced theory of statistics. Vol. 2B. Bayesian* . Arnold, London.

[20] Robert, C. and Casella, G. (1999). *Monte Carlo Statistical Method*. Springer, New York.

[21] Gilks W R, Richardson S and Spiegelhalter D J (Eds.) *Markov chain Monte Carlo in Practice*. Chapman and Hall, London, UK (1996).

[22] Ocaña-Riola R, Sánchez-Cantalejo C. *Epidemiología y análisis espacial*. Información Estadística y Cartográfica de Andalucía 2012; 2: 146-153.

[23] Besag, J. (1974). *Spatial interaction and the statistical analysis of lattice systems*. Journal of the Royal Statistical Society B, 36, 192-225.

[24] Rue, H. and S. Martino (2007). *Approximate Bayesian inference for hierarchical Gaussian Markov random fields models*. Journal of Statistical Planning and Inference 137 (10), 3177-3192. Special Issue: Bayesian Inference for Stochastic Processes.

[25] Rue H., Martino S. and Chopin N. (2009) *Approximate Bayesian Inference for Latent Gaussian Models Using Integrated Nested Laplace Approximations (with discussion)*. Journal of the Royal Statistical Society, Series B, 71, 319-392.

[26] Página web del Instituto Nacional de Estadística: `http://www.ine.es/jaxi/menu.do?L=0&divi=IDB&his=0&type=db`.

[27] *Registros de Cáncer-Principios y métodos* (capítulo 9). IARC, Publicaciones científicas nº95, Lyon (Francia) 1995.

[28] Breslow, N.E., Day, N.E. *Statistical methods in Cancer Research, volume II the design and analysis of cohort studies* (chapter 2). IARC Scientific Publications nº82, IARC Lyon 1987.

[29] Ugarte, M.D. (2009) *Mortality* en Encyclopedia of Medical Decision Making; VOL 2, Kattan, MW (Editor), Thousand Oaks, CA: SAGE Publications pp 788-792.

[30] Ugarte, M. D., Ibáñez, B., Militino, A. F. (2006). Modelling risks in disease mapping. Statistical Methods in Medical Research, 15, 21-35.

[31] Página web del Instituto Nacional del Cáncer: `http://www.cancer.gov/espanol/pdq/tratamiento/prostata/Patient/page1`.

[32] Borrell, C., Serral, G., Martínez-Beneito, M.A., Marí-Dell'Olmo, M., Rodríguez-Sanz, M., y grupo MEDEA (2009). *Atlas de mortalidad en ciudades de España (1996-2003)*, 2ª edición. ODS 2010-03-29.