UNIVERSITAT DE VALÈNCIA

Programa de Doctorat en Estadística i Optimització

SOME CONTRIBUTIONS IN DISEASE MAPPING MODELING

by

Francisca Corpas Burgos

PhD Thesis
in Statistics and Optimization

Supervised by

Miguel Ángel Martínez Beneito

May 2020

This thesis has been supported by predoctoral contract UGP-15-156 from the Fundación para el Fomento de la Investigación Sanitaria y Biomédica de la Comunitat Valenciana (FISABIO).

Acknowledgements

Después de varios años de trabajo, ha llegado el momento de concluir mi tesis doctoral escribiendo este apartado de agradecimientos. Echando la vista atrás, jamás imaginé que esto hubiera podido ocurrir algún día, pero aquí estoy, dejándome llevar por la oportunidad que se me ha brindado. A lo largo de estos años, he conocido personas maravillosas y he vivido muchos buenos momentos gracias a ellas, es por ello que me gustaría agradecer a todas aquellas personas que me han ayudado y apoyado, durante el tiempo de desarrollo de este trabajo. Este tiempo ha supuesto un período de aprendizaje intenso para mí, tanto a nivel profesional como personal.

En primer lugar, me gustaría comenzar agradeciendo la gran ayuda y apoyo que he recibido durante la realización de este trabajo, por parte de mi director de tesis, Miguel Ángel. Sin él, este trabajo no hubiera sido posible. Miguel Ángel y yo comenzamos a trabajar juntos allá por finales de 2014, cuando me recibió para realizar unas prácticas profesionales en el Área de Desigualdades en Salud de la Fundación FISABIO y, posteriormente, dirigir mi trabajo final de máster. Desde entonces hasta ahora, hemos continuado trabajando juntos en diferentes proyectos. Durante todos estos años, no ha habido ni un sólo momento en el que no resolviera todas aquellas dudas que me iban surgiendo. A lo largo de todo este tiempo, ha demostrado ser un gran profesional de la estadística y también una gran persona. No hay suficientes palabras

para agradecer todo lo que he podido aprender a su lado, el apoyo que he recibido, su buen trato, buenos consejos y su confianza en mí desde el primer momento. Además de ser mi director y jefe, también ha sido un gran amigo.

En segundo lugar, me gustaría agradecer el apoyo de mis compañeros del Área de Desigualdes en Salud de la Fundación FISABIO, los que están ahora y los que alguna vez estuvieron, y del Servicio de Epidemiología de la DGSP. Los momentos duros, tanto a nivel laboral como personal, han sido mucho más llevaderos a vuestro lado, siempre me habéis apoyado y habéis estado ahí, para ayudarme cuando lo he necesitado. Hemos demostrado ser un gran equipo y nunca os olvidaré. También a la Fundación FISABIO, por su contratación en 2016 para el desarrollo de esta tesis, gracias a su II Convocatoria de Ayudas para contratos predoctorales de formación en investigación. Sin esta ayuda, todo hubiera sido mucho más difícil.

Quisiera incluir en estos agradecimientos a la Dra. Paula Moraga, por recibirme en su departamento en Lancaster University, donde realicé una estancia de doctorado en 2018. Sin lugar a dudas, esta estancia ha sido una gran experiencia para mí, que volvería a repetir una y otra vez. Gracias Paula por ayudarme y estar disponible siempre que lo necesité. También a todas las personas que tuve la oportunidad de conocer durante mi estancia. Cada una de ellas contribuyó a que todo saliera bien y a hacer de mi estancia una de las mejores experiencias que he vivido.

Más allá del ámbito profesional, me gustaría extender mis agradecimientos a mi familia, mis padres y mi hermana, por todo el apoyo y ánimo que me han ofrecido durante estos años. En particular:

A mi abuelo Juan, por querernos tanto y cuidarnos siempre. Junto a mis padres, una de las personas más especiales e importantes de mi

vida. A mi abuela Encarnación, por preocuparse siempre por nosotros y enseñarnos que la humildad es algo valioso en la vida. A mi hermana, por demostrarme que siempre puedo contar con ella, en los buenos y en los malos momentos, siempre nos tendremos la una a la otra. Y, en especial, a mi madre y a mi padre, por su constante sacrificio y duro trabajo para que hoy yo haya podido llegar hasta aquí. Por ser unos luchadores, por sus sabios consejos, por motivarme a lograr mis metas y por confiar siempre en mí. Para mí, sois los mejores padres que podría tener y, gracias a vosotros, hoy soy la persona que soy. Aunque algunos ya no estéis físicamente con nosotros, siempre estaréis en mi corazón. Este trabajo va por vosotros.

Y, por último, a mis amigos y más que amigos, en especial a Joaquín, Noelia y Víctor. Gracias a vosotros nunca me he sentido sóla a pesar de tener lejos a mi familia. Siempre he podido contar con vosotros, en los buenos y no tan buenos momentos, y siempre habéis estado para apoyarme. Espero que nuestro vínculo perdure para siempre y continuemos compartiendo muchos buenos momentos, como lo hemos hecho hasta ahora.

GRACIAS A TODOS DESDE EL CORAZÓN

Dedicado a

Mi familia, en especial, a mis padres

ALGUNAS CONTRIBUCIONES A LA MODELIZACIÓN EN ESTUDIOS DE MAPEO DE ENFERMEDADES

UNIVERSITAT DE VALÈNCIA
Programa de Doctorat en Estadística i Optimització

Resumen amplio

La epidemiología espacial es la disciplina científica que persigue el estudio de la distribución geográfica de eventos relacionados con la salud, tales como la incidencia o las muertes por alguna enfermedad, así como de sus factores determinantes en la población (Last, 2001). Sus principales objetivos son describir, cuantificar y explicar las variaciones geográficas de las enfermedades, evaluar la asociación entre la incidencia de enfermedades y posibles factores de riesgo e identificar agrupaciones geográficas de las enfermedades (Elliott et al., 2000). La posibilidad de contar con datos de salud y población referenciados geográficamente, los avances en la computación y el desarrollo de metodologías estadísticas adecuadas han hecho posible el crecimiento de esta displicina.

El mapeo de enfermedades tiene una larga tradición como rama de la epidemiología espacial. Los mapas de enfermedades proporcionan un resumen visual de información geográfica compleja y permiten identificar patrones geográficos de las enfermedades que, simplemente observando los datos organizados en tablas, podrían pasar desapercibidos. De hecho, estas herramientas se utilizan con própositos descriptivos para la vigilancia de la salud pública, a fin de:

- Detectar aquellas localizaciones que muestran un mayor riesgo.

- Generar hipótesis etiológicas para identificar los factores de riesgo que influyen en la frecuencia de aparición de las enfermedades.
- Ayudar en la definición de políticas de salud y de asignación de recursos para paliar las desigualdades geográficas encontradas.

Los mapas de enfermedades representan indicadores epidemiológicos calculados a partir de los datos de salud disponibles. Estos datos pueden disponerse en diferentes formatos dependiendo de la unidad de análisis estudiada. En general, podemos distinguir entre datos a nivel de punto, cuando las unidades de análisis son individuales, y datos a nivel de área, cuando las unidades de análisis son las áreas geográficas en las que se divide una región de estudio. En este caso, las áreas geográficas se establecen habitualmente a partir de divisiones político-administrativas, tales como secciones censales, municipios o provincias. Por el contrario, los datos a nivel de punto corresponden a ubicaciones espaciales exactas en las que ocurrió el evento de salud. Cuando los datos corresponden a áreas geográficas, la información está disponible de forma agregada como conteos de eventos para cada una de las unidades geográficas en las que se divide la región de estudio. Las herramientas de estadística espacial utilizadas para calcular los indicadores que serán representados en los mapas de enfermedades dependerán del tipo de datos disponibles. En esta tesis, nos centramos en las técnicas de *disease mapping* diseñadas para estudiar la distribución geográfica de las enfermedades a través de regiones de estudio divididas en unidades geográficas pequeñas.

La ventaja de los datos de salud agregados geográficamente es que son fácilmente accesibles, ya que preservan la confidencialidad de los individuos y son recopilados rutinariamente por un gran número de instituciones estadísticas y de salud (Botella Rocamora et al., 2017). Por el contrario, el acceso a datos de salud a nivel individual está mucho más limitado y éstos rara vez suelen estar disponibles. Sin

embargo, la construcción de mapas de enfermedades a partir de datos agregados geográficamente implica una pérdida de información que sería conveniente limitar. Si el tamaño de las unidades geográficas es grande, se podrían enmascarar variaciones del riesgo que pudieran ocurrir dentro de ellas, cuando éstas son de evidente interés. Para evitar esto, es conveniente considerar la región de interés dividida en áreas geográficas del menor tamaño posible. Además, el estudio de la variabilidad espacial de los riesgos en áreas geográficas pequeñas permite análisis más similares al nivel individual. En ellas, la población es más homogénea, en cuanto a hábitos de vida y condiciones socioeconómicas, y también, el entorno presenta características similares, resultando más difícil que se produzcan variaciones importantes del riesgo dentro de ellas.

Cuando se trabaja con áreas pequeñas o enfermedades poco comunes, el mapeo directo de indicadores epidemiológicos crudos presenta algunos problemas. El indicador epidemiológico más habitual para evaluar el riesgo de enfermedad en las áreas de una región de estudio es la Razón de Morbilidad/Mortalidad Estandarizada (SMR). Esta medida se calcula mediante el cociente entre el número de eventos observados en cada unidad geográfica y el número de eventos esperados, en relación a sus habitantes y las edades de los mismos, si los riesgos para cada grupo de edad fueran los mismos que en cierta población de referencia (habitualmente el total de la región de estudio). Así, valores de la SMR superiores a 1 indicarían que los casos observados en la unidad geográfica son superiores a los casos esperados, según la estructura de su población, reflejando un exceso de riesgo en la unidad geográfica. Por el contrario, valores de la SMR inferiores a 1 indicarían un riesgo en la unidad geográfica inferior al de la población de referencia, habitualmente el de toda la región de estudio. Cuando la región de estudio se considera dividida en unidades geográficas pequeñas, el número de casos

observados y/o esperados por unidad suele ser bajo, como consecuencia de la existencia de poca población en las áreas. Esto hace que la SMR muestre gran variabilidad, dando lugar a mapas que alternan áreas contiguas con riesgos opuestos, que generalmente carecen de sentido epidemiológico. Con el fin de solucionar este problema y obtener mapas que reflejen variaciones del riesgo más sensatas, los indicadores epidemiológicos deben estimarse utilizando modelos estadísticos que tengan en cuenta la dependencia espacial que los datos podrían mostrar. La incorporación de la dependencia espacial en los modelos significa que los riesgos en cada unidad geográfica podrían ser estimados teniendo en cuenta también los riesgos de sus unidades cercanas. Esta información adicional permite obtener estimaciones más fiables que las SMR crudas originales, las cuales se consideraron geográficamente independientes cuando en realidad no lo son.

Un gran número de modelos estadísticos para el mapeo de enfermedades han sido propuestos en la literatura, la mayoría de ellos siguiendo una aproximación Bayesiana (Besag et al., 1991; Leroux et al., 1999; Lawson et al., 2000; Lawson and Clark, 2002; Assunçao, 2003; Best et al., 2005; Ugarte et al., 2006; MacNab, 2007; Lee, 2011; Bauer et al., 2016; Goicoa et al., 2016). Entre ellos, Besag et al. (1991) (BYM) y Leroux et al. (1999) son dos de los modelos más frecuentemente utilizados en estudios aplicados. Estas propuestas han supuesto un punto de referencia en el estudio de la distribución espacial del riesgo de enfermedades y han servido como base para la formulación y el desarrollo de nuevas propuestas de modelización (MacNab et al., 2006a; Congdon, 2008; Martinez-Beneito et al., 2008; Song et al., 2011). En esta tesis, las propuestas de BYM y Leroux et al. han sido evaluadas en diferentes escenarios y, también, han sido el punto de partida para el desarrollo de nuevos modelos de mapeo de enfermedades que mejoran estas propuestas en los escenarios considerados.

Esta tesis tiene cuatro objetivos principales, todos ellos relacionados con la aplicación, la evaluación y el desarrollo de modelos de mapeo de enfermedades en diferentes contextos.

El primero de los objetivos de esta tesis surge tras la aplicación de modelos estándar de suavización espacial para el mapeo de enfermedades, específicamente el ya mencionado modelo de referencia de BYM. En el estudio de enfermedades poco comunes, es posible que un gran número de unidades geográficas de la región de estudio no presenten casos observados de la enfermedad. En tales ocasiones, cuando se utilizan modelos estándar para llevar a cabo las estimaciones de los riesgos, es común encontrarnos con un problema de "exceso de ceros" en los datos: el número de áreas que no presentan casos de la enfermedad excede considerablemente el número de ceros que los modelos estándar podrían razonablemente explicar. En dicho caso, se produce un desajuste de los datos originales, en términos del número de ceros observados, y es necesario incorporar estrategias de modelización específicas para ajustar ese exceso de ceros, que mejoren, en este sentido, el ajuste de los modelos estándar de mapeo de enfermedades. En la literatura del mapeo de enfermedades, algunas propuestas para modelizar datos con exceso de ceros han sido sugeridas (Mullahy, 1986; Lambert, 1992; Gschlößl and Czado, 2008; Song et al., 2011; Musenge et al., 2013; Neelon et al., 2013). Las propuestas más populares para modelizar datos con exceso de ceros son el modelo de Poisson cero-inflado (ZIP) y el modelo de Poisson hurdle. Tras la aplicación y evaluación de estas propuestas, a diferencia de lo que esperaríamos, encontramos que dichas propuestas no ajustan suficientemente bien muchos conjuntos de datos. Por tanto, en este contexto, nos planteamos el objetivo de desarrollar estrategias de modelización alternativas que aseguren un buen ajuste del exceso de ceros en aquellos datos que lo requieran.

El segundo objetivo de esta tesis se centra en el estudio de modelos

multivariantes para el análisis conjunto de la distribución geográfica del riesgo de varias enfermedades. Habitualmente, los modelos de mapeo de enfermedades son univariantes, centrándose en la modelización de los riesgos de una sóla enfermedad. Sin embargo, es posible que varias enfermedades compartan factores de riesgo comunes y, por tanto, podrían beneficiarse de un análisis conjunto. Recientemente, la modelización multivariante ha recibido especial atención por parte de investigadores interesados en la modelización espacial conjunta de los riesgos de varias enfermedades simultáneamente. La modelización multivariante en disease mapping tiene como objetivo estimar la distribución geográfica de los riesgos de varias enfermedades, teniendo en cuenta la dependencia espacial para cada enfermedad y la dependencia entre las enfermedades. De esta forma, las estimaciones de los riesgos estarían basadas en una mayor cantidad de información, lo que permitiría obtener estimaciones más precisas en comparación con el mapeo univariante de enfermedades.

Con el fin de evidenciar las ventajas de la modelización multivariante con respecto a la modelización univariante, exploramos algunos modelos específicos de modelización espacial multivariante propuestos en la literatura. Específicamente, exploramos las propuestas de modelización multivariante de Botella-Rocamora et al. (2015). Tras evaluar los resultados obtenidos con estas propuestas, encontramos algunas limitaciones cuando éstas son aplicadas en regiones de estudio pequeñas. Las limitaciones encontradas se traducen en estimaciones poco precisas de los riesgos, como consecuencia de imponer una variabilidad común a todos los patrones de riesgo de las enfermedades consideradas. Por esta razón, nos planteamos en este caso el objetivo de extender los modelos multivariantes originales de Botella et al., permitiendo heterocedasticidad entre las enfermedades, que solucionen los problemas evidenciados y proporcionen estimaciones precisas de los riesgos.

Los estudios típicos de mapeo de enfermedades consideran dependencia espacial entre las observaciones a través de efectos aleatorios que siguen alguna distribución previa espacial. Probablemente las distribuciones previas espaciales más populares son la familia de distribuciones Condicionales Autorregresivas (CAR) (Besag, 1974; Besag et al., 1991). Estas distribuciones inducen dependencia espacial por medio de una matriz de pesos espaciales que reflejan la fuerza de dependencia entre cualquier par de unidades geográficas. El procedimiento más común para definir los pesos espaciales en las distribuciones CAR es utilizar un criterio de adyacencia. En ese caso, a todos los pares de unidades geográficas con bordes adyacentes se les asigna el mismo peso, típicamente 1, y al resto de unidades no adyacentes se les asigna un peso de 0, reflejando independencia (condicional) entre ellas. Sin embargo, asumir el mismo peso para todas las áreas vecinas puede ser demasiado rígido o inapropiado en algunos escenarios. Por esta razón, nuestro tercer objetivo es explorar y desarrollar procedimientos para estimar matrices de pesos espaciales en estudios de mapeo de enfermedades que resuelvan este problema. Específicamente, nos centramos en el desarrollo de distribuciones CAR adaptativas. Estas distribuciones consideran los pesos espaciales como variables aleatorias en los modelos, lo que les permite ser diferentes, de modo que puedan estimarse a partir de la información proporcionada por los datos.

El último objetivo de esta tesis es desarrollar y publicar en formato de aplicación web un atlas de mortalidad nacional avanzado a nivel municipal. Esta aplicación estudiaría un gran número de causas de muerte para el conjunto de toda España y utilizaría modelos apropiados de estimación en áreas pequeñas para estimar las SMRs suavizadas. Específicamente, las SMRs serían estimadas mediante modelos de suavización espacial y espacio-temporal. Por un lado, el

modelo BYM sería usado para obtener las SMRs suavizadas durante todo el período de estudio. Por otro lado, también es interesante estudiar la mortalidad en periodos de tiempo más cortos y analizar su evolución a lo largo del tiempo. La modelización espacio-temporal de los datos sería llevada a cabo mediante el modelo propuesto por Martinez-Beneito et al. (2008). Dicha propuesta de modelización consiste en la integración de modelos de suavización espacial y de series temporales, específicamente de procesos autorregresivos, para modelizar simultáneamente la dependencia espacial y temporal que las observaciones pueden mostrar. Este objetivo está alineado con la aplicación y evaluación de modelos de mapeo de enfermedades en grandes regiones de estudio, el objetivo inicial de esta tesis. El atlas de mortalidad nacional desarrollado será una herramienta muy útil tanto para la población general, que podrá conocer el estado de salud de su municipio y entorno, como para los profesionales de la salud, para los que el atlas puede proporcionar información epidemiológica de gran valor sobre la población a la que atienden.

Para llevar a cabo cada uno de los objetivos planteados en esta tesis se ha hecho uso de extensos conjuntos de datos reales. Específicamente, los conjuntos de datos analizados y utilizados en esta tesis han sido los siguientes:

1. Datos de mortalidad a nivel municipal para un total de 27 causas de muerte en hombres y mujeres en la Comunidad Valenciana durante el periodo 1987-2006.

   En este conjunto de datos, examinamos la necesidad de incorporar estrategias de modelización de exceso de ceros en los modelos estándar de mapeo de enfermedades y evaluamos nuestras propuestas, con respecto al ajuste de ceros, en los conjuntos de datos que así lo requieren (objetivo 1).

2. Datos de mortalidad a nivel de sección censal para un total de 20 causas de muerte en hombres y mujeres en las ciudades de Alicante, Castellón y Valencia durante el periodo 1996-2015.

   Este conjunto de datos ha sido utilizado, por un lado, en la evaluación y el desarrollo de modelos multivariantes de mapeo de enfermedades (objetivo 2) y, por otro lado, en el desarrollo de modelos espaciales con matrices de pesos adaptativas (objetivo 3).

3. Datos de mortalidad a nivel municipal para un total de 102 causas de muerte en hombres y mujeres en el conjunto de toda España durante el periodo 1989-2014.

   Este conjunto de datos corresponde al utilizado en el desarrollo del Atlas Nacional de Mortalidad en España (ANDEES) (objetivo 4).

La tesis que aquí presentamos es un compendio de tres artículos y un trabajo adicional. Los trabajos correspondientes a los tres primeros objetivos de esta tesis han sido publicados en forma de artículos de investigación en revistas indexadas en el índice de Estadística y Probabilidad del Journal Citation Reports (JCR). Específicamente, los artículos de investigación publicados han sido los siguientes:

- "Some findings on zero-inflated and hurdle Poisson models for disease mapping" publicado en *Statistics in Medicine* por F. Corpas-Burgos, G. García-Donato y M.A. Martinez-Beneito (2018).

- "On the convenience of heteroscedasticity in highly multivariate disease mapping" publicado en *Test* por F. Corpas-Burgos, P. Botella-Rocamora y M.A. Martinez-Beneito (2019).

- "On the use of adaptive spatial weight matrices from disease mapping multivariate analyses" publicado en *Stochastic Environmental Research and Risk Assessment* por F. Corpas-Burgos y M.A. Martinez-Beneito (2020).

El Atlas Nacional de Mortalidad en España (ANDEES) desarrollado, marcado como el cuarto y último objetivo de esta tesis, ha sido publicado online y puede visualizarse en el siguiente enlace web:

http://atlasnacional.fisabio.es

Las principales herramientas utilizadas en el desarrollo de esta tesis han sido el paquete estadístico `R` y el software para análisis Bayesiano utilizando métodos Markov chain Monte Carlo (MCMC) `WinBUGS`. Por un lado, `WinBUGS` ha sido usado para ajustar cada uno de los modelos a los datos y obtener las estimaciones de interés. Por otro lado, `R` y algunos de sus paquetes han sido usados para el manejo de los datos y los resultados de cada trabajo. Los principales paquetes de `R` utilizados han sido las librerías `Pbugs` y `Shiny`. `Pbugs` ha hecho posible automatizar las llamadas a `WinBUGS` desde `R`, ejecutando en paralelo (en diferentes procesadores) las diferentes cadenas que han sido necesarias en cada uno de los modelos ajustados y, por tanto, acelarando el tiempo de computación para obtener los resultados. `Shiny` ha permitido el desarrollo de la web que aloja los resultados de ANDEES.

Para cada uno de los artículos publicados, el código `R`, con la implementación de todos los modelos ajustados y para todos los análisis realizados, es proporcionado como material suplementario y puede ser encontrado en el Apéndice de esta tesis.

Esta tesis se estructura de la siguiente forma. En el Capítulo 1, presentamos una introducción general, incluyendo una descripción de los objetivos, de los datos analizados en cada trabajo y del software utilizado.

En el Capítulo 2, describimos el marco general de modelización en los estudios de mapeo de enfermedades, así como las propuestas de modelización originales sobre las que la presente tesis pretende realizar alguna aportación. Resumimos también las limitaciones encontradas en dichas propuestas y las nuevas propuestas de modelización desarrolladas en esta tesis para resolverlas. El Capítulo 3 resume los principales resultados obtenidos en cada trabajo. Los Capítulos 4, 5 y 6 contienen los tres artículos de investigación publicados que componen este compendio. En el Capítulo 7, describimos la metodología utilizada en el desarrollo de ANDEES y sus principales características y resultados. Por último, en el Capítulo 8, presentamos algunas conclusiones y posibles líneas de trabajo futuro.

A continuación, resumimos los principales resultados y conclusiones obtenidas en los trabajos desarrollados en esta tesis.

En primer lugar, con respecto a la modelización de datos con exceso de ceros, mostramos cómo el exceso de ceros puede encontrarse frecuentemente en la práctica cuando los datos son modelizados usando modelos estándar de mapeo de enfermedades. En el conjunto de datos de mortalidad de la Comunidad Valenciana, analizado en el Capítulo 4, encontramos que una proporción relevante de las enfermedades estudiadas muestran exceso de ceros. Por tanto, el exceso de ceros requiere atención en los estudios geográficos de la mortalidad y se necesitan modelos específicos para tratar con este problema, ya que de lo contrario, mapas con riesgos sobresuavizados son obtenidos. En este sentido, encontramos que los modelos ZIP y hurdle, propuestos para tratar con esta falta de ajuste, sin una modelización explícita de las probabilidades de ceros, no ajustan problemas de exceso de ceros suficientemente bien y son claramente insatisfactorios. Los resultados sugieren la necesidad de una modelización explícita de las probabilidades de ceros que deberían variar entre las unidades

geográficas. Desafortunadamente, demostramos en varios resultados teóricos que estas estrategias de modelización más flexibles pueden conducir fácilmente a distribuciones a posteriori impropias o arbitrarias. Esto hace que la modelización sea bastante complicada y se debe tener precaución para evitar propuestas de modelización erróneas. Nuestros resultados determinan algunas propuestas específicas de modelos ZIP y hurdle, frecuentemente propuestas en la literatura, que deberían de evitarse en general. Finalmente, proponemos varias alternativas de modelización válidas que no presentan los problemas anteriores y que son adecuadas para ajustar excesos de ceros. Mostramos que dichas propuestas solucionan los problemas de exceso de ceros y corrigen la mecionada sobresuavización de los riesgos en las unidades poco pobladas, representando patrones geográficos más adecuados a los datos.

En nuestro trabajo sobre mapeo de enfermedades multivariante, encontramos que la propuesta de Botella-Rocamora et al. (2015) para la modelización espacial conjunta de varias enfermedades muestra algunas limitaciones cuando los datos son más débiles. Específicamente, en tales situaciones, la estructura previa de esta propuesta puede influir significativamente en los patrones de riesgo estimados para todas las enfermedades consideradas. Este hecho es causado por el único parámetro de varianza común en la matriz M de este modelo, la cual controla la variabilidad general de todos los patrones de riesgo ajustados. Si la variabilidad de los patrones de riesgo considerados fuese diferente, estas asunciones previas pueden producir evidentes desajustes en los patrones de riesgos que son estimados. Una de las principales contribuciones de este trabajo ha sido evidenciar estas limitaciones, que son particularmente preocupantes cuando la propuesta original de Botella-Rocamora et al. (2015) se aplica a regiones de estudio pequeñas. En esta tesis, proponemos dos modificaciones del modelo multivariante anterior que incorporan diferentes parámetros para modelizar la

variabilidad de los riesgos de cada enfermedad y permiten resolver los problemas evidenciados. Estas nuevas propuestas heterocedásticas permiten que los patrones espaciales para cada enfermedad tengan mayor o menor variabilidad cuando sea necesario, haciendo posible obtener estimaciones de los riesgos más flexibles y precisas.

En nuestro trabajo sobre dependencia espacial adaptativa, proponemos un procedimiento para estimar la matriz de pesos espaciales en las distribuciones CAR de acuerdo a datos retrospectivos multivariantes. Nuestro procedimiento adaptativo hace que los modelos CAR sean más flexibles y mejoren el ajuste de posteriores análisis, adoptando la matriz de pesos espaciales estimada, que debería haber capturado las particularidades que los datos de mortalidad podrían mostrar en esa región. Además, el carácter multivariante de nuestra propuesta ha demostrado ser una herramienta indispensable para estimar adecuadamente la estructura espacial de los datos.

La metodología introducida podría tener diferentes usos. En primer lugar, el modelo adaptativo multivariante introducido podría usarse en estudios multivariantes, considerando también dependencia entre las causas de mortalidad, con estructuras espaciales adaptativas. Estos modelos proporcionarían estimaciones de los riesgos más precisas, aprovechando el carácter adaptativo de la dependencia espacial considerada. Un segundo uso de modelos CAR adaptativos sería el destacado en nuestro trabajo, es decir, hacer inferencia en la matriz de pesos espaciales de una región de estudio. Como consecuencia, esa matriz de pesos adaptativos podría usarse más tarde en posteriores estudios de mapeo de enfermedades con una estructura espacial no arbitraria, basada en datos y conocimientos previos. También hemos encontrado un tercer uso práctico de nuestro modelo adaptativo. Este uso sería el control de calidad de problemas sistemáticos que podrían estar presentes en los conjuntos de datos de salud. Específicamente, los

datos de mortalidad de la ciudad de Valencia, analizados en el Capítulo 6, pertenecen a un gran proyecto español que estudia la mortalidad en grandes ciudades, el proyecto MEDEA. Todas las muertes en ese conjunto de datos han sido geocodificadas mediante el uso de varias herramientas de geocodificación, en particular la API de geocodificación de Google y una segunda herramienta de geocodificación (Cartociudad) del Instituto Geográfico Nacional de España. Estas herramientas, como cualquier otra herramienta de geocodificación, no son perfectas y tienen errores, para algunas calles en particular, grupos de casos que son geocodificados en el centro de la ciudad, etc., que podrían distorsionar los análisis espaciales de esa base de datos. Hemos encontrado que el modelo adaptativo multivariante en esas bases de datos otorga bajos pesos espaciales a aquellas secciones censales con errores sistemáticos de geocodificación, ya que sus datos de mortalidad son algo diferentes de sus áreas circundantes. Esto nos ha permitido detectar esos errores (y corregirlos) al enfocarnos en aquellas secciones censales con bajos pesos espaciales y sin una posible explicación alternativa para ellos (sin geriátricos, sin áreas socialmente marginales, sin áreas de nueva construcción, etc.).

Finalmente, el Atlas Nacional de Mortalidad en España (ANDEES) desarrollado permite conocer a nivel municipal la distribución geográfica y la evolución temporal de la mortalidad debida a un gran conjunto de causas de muerte en toda España. Los resultados mostrados en ANDEES muestran la existencia de patrones geográficos de mortalidad muy diferentes según la causa, el sexo y el período de estudio analizado. Esta herramienta permitirá a los investigadores y expertos en Salud Pública examinar los patrones geográficos de las enfermedades y detectar áreas de alto riesgo, que no son evidentes a través de otros tipos de análisis. Los resultados presentados pueden desempeñar un papel crucial en la búsqueda de factores de riesgo, así como en el establecimiento de

prioridades, y orientar las políticas sociales y de salud.

ANDEES deja abiertas muchas posibles líneas de trabajo futuro. Por un lado, nos gustaría actualizar periódicamente los resultados del atlas, incorporando los datos de mortalidad posteriores a 2014. Por otro lado, también nos gustaría implementar otros modelos más complejos y flexibles para profundizar en el entendimiento de la distribución geográfica de las enfermedades. Específicamente, estaríamos interesados en implementar cada uno de los modelos desarrollados en esta tesis a nivel nacional. Así, evaluaríamos (y arreglaríamos) la posible existencia de problemas de exceso de ceros en cada uno de los conjuntos de datos analizados. Además, la modelización multivariante considerando grupos de enfermedades que pudieran tener factores de riesgo comunes, mejoraría en gran medida la estimación geográfica de los riesgos, al hacer uso de fuentes de información alternativas. Del mismo modo, la modelización espacial adaptativa también permitiría obtener mapas de riesgo con mayor variabilidad, permitiendo a los municipios con características especiales mostrar el comportamiento separado que requieren. Finalmente, la combinación de modelos espacio-temporales con estas propuestas, aquellas que muestren una mejora más evidente en el análisis espacial, permitiría obtener una visión actualizada y más precisa de los riesgos. La implementación de algunos de estos modelos para el análisis de la mortalidad en toda España podría dar lugar a problemas computacionales desafiantes, dado el gran tamaño de la región de estudio considerada y la gran cantidad de patrones geográficos que se estimarían en un sólo modelo. Como consecuencia, otra línea de trabajo futura sería resolver tales problemas computacionales, mediante la exploración de diferentes herramientas computacionales y la optimización de la implementación de cada uno de los modelos propuestos.

SOME CONTRIBUTIONS IN DISEASE MAPPING MODELING

UNIVERSITAT DE VALÈNCIA
Doctoral Program in Statistics and Optimization

Abstract

Disease mapping has received great interest for the past three decades. This research area pursues the study of the geographical distribution of health-related events, such as mortality or the incidence of diseases, aggregated in geographic units, in order to mainly identify those locations that show a higher risk. The application of advanced statistical methods to carry risk estimates out is essential to obtain accurate estimates and to improve the understanding of the geographical distribution of diseases.

In this thesis, we focus on the application and evaluation of several relevant modeling proposals, emerged in the disease mapping literature, to estimate the mortality geographic distribution, considering different scenarios with real data. Specifically, we study the distribution of mortality at the census tract level in the main cities of the Valencian Region and at the municipal level in the Valencian Region and in the whole of Spain. The evaluation of these previously published proposals reveals some statistical problems in their implementations. Therefore, our main goal with this thesis is the development of new methodological proposals that allow solving the problems of these previously published proposals. Likewise, we also pursue the development of an advanced national mortality atlas that allows to interactively visualize the

geographical distribution, and the temporal evolution, of mortality for a large number of causes and throughout a long period of study in the whole of Spain.

This thesis is a compendium of three articles and an additional work, which are structured as follows. In Chapter 1, we present a general introduction, including a description of the objectives, the data analyzed in each work and the software used. In Chapter 2, we introduce the general problem of disease mapping, as well as the modeling proposals that have been improved in each of our works. We also summarize the limitations found in these proposals and the new modeling proposals developed in this thesis. Chapter 3 summarizes the main results obtained in each of the subsequent works. Chapters 4, 5 and 6 contain the three published research articles that make up this compendium. In Chapter 7, we describe the methodology used in the development of the Spanish National Atlas of Mortality and its main characteristics and results. Finally, in Chapter 8, we present some conclusions and possible lines of future work.

Contents

List of Figures

List of Tables

1. Introduction

1.1. Motivation

Spatial epidemiology is the scientific discipline that pursues the study of the geographical distribution of events related to health, such as the incidence or deaths for some disease, as well as its determining factors in the population (Last, 2001). Its main objectives are to describe, quantify and explain the geographical variations of the diseases, to evaluate the association between the incidence of diseases and possible risk factors and to identify geographic groupings of the diseases (Elliott et al., 2000). The widespread access to geographically referenced health and population data, advances in computing and the development of adequate statistical methodologies have made the growth of this discipline possible.

Disease mapping has a long tradition as a branch of spatial epidemiology. Disease maps provide a visual summary of complex geographic information and allow to identify geographic patterns of diseases that, by simply looking at data organized in tables, could go unnoticed. In fact, these tools are used for descriptive purposes for public health surveillance, in order to:

- Detect those locations that show a higher risk.
- Generate etiological hypotheses to identify risk factors ruling the frequency of appearance of diseases.

- Help in the implementation of health policies and resource allocation to alleviate the geographic inequalities found.

Disease maps represent epidemiological indicators calculated from the available health data. Theses data are arranged in different formats depending on the unit of analysis studied. In general, we can distinguish between data at the point level, when the units of analysis are individuals, and data at the area level, when the units of analysis are the geographical areas into which a study region is divided. In that case, geographical areas are usually defined by political-administrative divisions, such as census tracts, municipalities or provinces. Conversely, point-level data correspond to exact spatial locations in which the health event occurred. When data correspond to geographical areas, the information is available in aggregate form as event counts for each of the geographic units into which the study region is divided. The spatial statistics tools used to calculate the indicators that will be represented in the disease maps will depend on the type of data available. In this thesis, we focus on disease mapping techniques devised to study the geographical distribution of diseases through regions of studied divided into small geographical units.

The advantage of geographically aggregated health data is that they are easily accessible, since they preserve the confidentiality of individuals and are routinely collected by a large number of statistical and health institutions (Botella Rocamora et al., 2017). In contrast, access to individual level health data is much more limited and these are seldom available. Nevertheless, the construction of disease maps from geographically aggregated data implies a loss of information that would be convenient to limit. If the size of the geographic units is large, risk variations that might occur within them could be masked, when these are of obvious interest. To avoid this, it is convenient to consider the region of interest divided into geographic areas of the smallest possible size. Furthermore, the study of spatial variability of risks in small geographic areas allows analyses more similar to the

individual level. In these, the population is more homogeneous in terms of lifestyle habits and socioeconomic conditions and, also, the environment presents similar characteristics, making it more unlikely important risk variations to occur within them.

When working with small areas or uncommon diseases, direct mapping of raw epidemiological indicators presents some problems. The most common epidemiological indicator for assessing disease risk in the areas of a study region is the Standardized Morbidity/Mortality Ratio (SMR). This measure is calculated by the quotient between the number of events observed in each geographic unit and the number of expected events, in relation to its inhabitants and their ages, if the risks for each age group were the same as in a certain reference population (usually the total of the study region). Thus, SMR values greater than 1 would indicate that the observed cases in that geographic unit are higher than the expected cases, according to its population structure, reflecting a risk excess in that geographic unit. Conversely, SMR values lower than 1 would indicate a risk in that geographic unit lower than that of the reference population, usually that of the whole study region. When the study region is divided into small geographic units, the number of observed and/or expected cases per unit is usually low as a consequence of the low population in those units. This makes the SMR show great variability, giving rise to maps that alternate contiguous areas with opposite risks, which usually lack of any epidemiological sense. In order to solve this problem and obtain maps that reflect more sensible risks variations, epidemiological indicators must be estimated by using statistical models that take into account the hypthetical spatial dependence that the data could show. The incorporation of spatial dependence into the models means that the risks in each geographic unit could be estimated taking also the risks of nearby units into account. This additional information allows obtaining more reliable estimates than the original raw SMRs, which were considered to be geographically independent when actually they are not.

In order to visually illustrate this phenomenon, we represent in Figure 1.1 the SMRs for mortality from malignant tumor of the oral cavity in men in the municipalities of the Valencian Region for the period 1987-2006. On the left map the raw SMRs are represented and on the right map those estimated by the statistical model proposed by Besag et al. (1991), which takes into account the spatial dependence between events that occur in nearby locations.

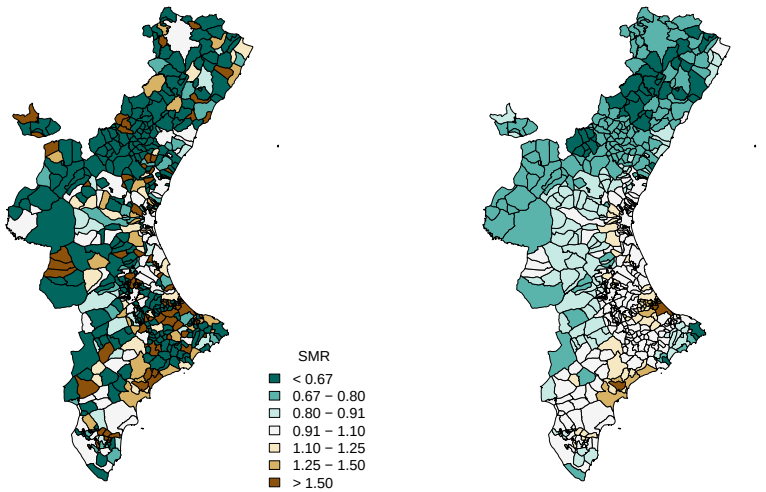


Figure 1.1.: Raw SMRs (left) vs. estimated SMRs using the statistical model proposed by Besag et al. (1991) for oral cavity tumor mortality in men in the Valencian Region during the period 1987-2006.

A large number of statistical models for disease mapping have been proposed in the literature, most of them following a Bayesian approach (Besag et al., 1991; Leroux et al., 1999; Lawson et al., 2000; Lawson and Clark, 2002; Assunçao, 2003; Best et al., 2005; Ugarte et al., 2006; MacNab, 2007; Lee, 2011; Bauer et al., 2016; Goicoa et al., 2016). Between them, Besag et al. (1991) (BYM) and Leroux et al. (1999) outstand of two of the most frequently used models in applied studies. These proposals have been a benchmark in the study of the

spatial distribution of disease risks and have served as a basis for the formulation and development of new modeling proposals (MacNab et al., 2006a; Congdon, 2008; Martinez-Beneito et al., 2008; Song et al., 2011). In this thesis, BYM and Leroux et al.'s proposals have been assessed in different scenarios and have been the starting point for the development of new disease mapping models improving these basic modeling proposals in those particular settings. Given the relevance of the BYM and Leroux et al.'s proposals, their formulation will be described in detail in Chapter 2 (Methodology).

Below we describe the main objectives of this thesis.

1.2. Objectives

This thesis has four main goals, all of them related to the application, evaluation and development of spatial statistical models in different contexts. These four objectives are:

- **Objective 1**. The first objective of this thesis arises after the application of standard spatial smoothing models for disease mapping, specifically the aforementioned BYM reference model. In the study of uncommon diseases, a large number of geographic units in the study region may not have observed cases of the disease. On such occasions, when standard models are used to carry out risk estimates, it is common to find a "zero excesses" problem in the data, that is the number of areas that do not present cases of the disease considerably exceeds the number of zeroes that standard models could reasonably explain. In this case, there is a poor fit of the original data, in terms of the number of observed zeroes, and it is necessary to incorporate specific modeling strategies for fitting those zero excesses, which improve in this sense, the fit of standard disease mapping models. In the disease mapping literature, some proposals for modeling zero excesses have been suggested (Mullahy, 1986; Lambert, 1992;

Gschlößl and Czado, 2008; Song et al., 2011; Musenge et al., 2013; Neelon et al., 2013). After the application and evaluation of some zero excess modeling proposals, unlike what we would expect, we find that they do not fit many data sets well enough. Therefore, in this context, we set ourselves the objective of developing alternative modeling strategies that ensure a good fit of the "zero excesses" in those data sets that require it.

- **Objective 2**. The second objective of this thesis focuses on the study of multivariate models for the joint analysis of the geographical distribution of several diseases. Typically, diseases mapping models are univariate, dealing with just one disease or process. However, several diseases may share common risk factors and may therefore benefit of a joint analysis. Recently, multivariate modeling has received special attention from researchers interested in the joint spatial modeling of the risks of several diseases simultaneously. Multivariate modeling in disease mapping aims to estimate the geographical distribution of risks for several diseases taking into account the spatial dependence for each disease and the dependence between diseases. In this way, risk estimates would be based in a greater amount of information, which would allow obtaining more accurate estimates in comparison to the univariate mapping of diseases.

  In order to show the advantages of multivariate modeling with respect to univariate models, we explore some specific multivariate proposals in the literature. Specifically, we explore the multivariate modeling proposals of Botella-Rocamora et al. (2015). After assessing the results obtained with these proposals, we found some limitations when they are applied in small study regions. The limitations found yield unreliable risks estimates as a consequence of imposing a common variability to all the risk patterns of the diseases considered. For this reason, we set ourselves the objective of extending the original Botella et al.'s

multivariate models, allowing heterocedasticity between diseases, which solves the evidenced problems and provides more accurate risk estimates.

- **Objective 3**. Typical disease mapping studies consider spatial dependence between observations through random effects that follow some spatial prior distribution. Probably the most popular spatial prior distributions are the family of Autoregressive Conditional (CAR) distributions (Besag, 1974; Besag et al., 1991). These distributions induce spatial dependence by means of a matrix of spatial weights which reflect the strength of dependency between any pair of geographic units. The most common procedure to define spatial weights in CAR distributions is to use an adjacency criterion. In that case, all pairs of geographic units with adjacent edges are given the same weight, typically 1, and the rest of the non-adjacent units are assigned a weight of 0, reflecting (conditional) independence between them. However, assuming the same weight to all neighboring areas may be too rigid or inappropriate in some scenarios. For this reason, we aim to explore and develop procedures to estimate spatial weight matrices in disease mapping studies that solve this problem. Specifically, we focus on the development of adaptive CAR distributions. These distributions consider the spatial weights as random variables in the models, allowing them to be different, so that they can be estimated from the information provided by the data.

- **Objective 4**. The last objective of this thesis is the development of an advanced national mortality atlas at the municipal level, as a web application. This application would study a large number of causes of death for the whole of Spain and would use appropriate small areas estimation models for estimating the smoothed SMRs. This objective is aligned with the application and evaluation of disease mapping models in very large study regions, the initial main goal of this thesis. The developed national mortality atlas

will be a very useful tool for both the general population, who will be able to know the state of health of their municipality and environment, and health professionals for whom the atlas can provide epidemiological information of great value on the population to which they attend. The development of the atlas will be carried out in collaboration with the *Bayensians* research group of the FISABIO Foundation.

1.3. Data

Extensive real data sets have been used to carry out each of the objectives of this thesis. Specifically, the data sets analyzed and used in this thesis have been the following:

1. **Municipal mortality data for a total of 27 causes of death in men and women in the Valencian Region during the period 1987-2006.**

   In this data set, we examine the need to model zero excesses in standard disease mapping models and evaluate our proposals in regards to the adjustment of zeroes in those data sets that require it (Objective 1).

2. **Mortality data at the census tract level for 20 causes of death in men and women in the cities of Alicante, Castellón and Valencia during the period 1996-2015.**

   This data set has been used, on the one hand, in the assessment and development of multivariate disease mapping models (Objective 2) and, on the other hand, in the development of spatial models with adaptive weight matrices (Objective 3).

3. **Spanish mortality data at the municipal level for a total of 102 causes of death in men and women during the period 1989-2014.**

This data set corresponds to that used in the development of the Spanish National Atlas of Mortality (ANDEES) (Objective 4).

1.4. Publications

The works corresponding to the first three objectives of this thesis have been presented in different talks at national and international conferences and published in the form of research articles in journals indexed in the Statistics and Probability index of the Journal Citation Reports (JCR). Specifically, the published research articles have been the following:

- **"Some findings on zero-inflated and hurdle Poisson models for disease mapping"** published in ***Statistics in Medicine*** by F. Corpas-Burgos, G. García-Donato and M.A. Martinez-Beneito (2018). This paper corresponds to Chapter 4 of this thesis.

- **"On the convenience of heteroscedasticity in highly multivariate disease mapping"** published in ***Test*** by F. Corpas-Burgos, P. Botella-Rocamora and M.A. Martinez-Beneito (2019). This paper corresponds to Chapter 5 of this thesis.

- **"On the use of adaptive spatial weight matrices from disease mapping multivariate analyses"** published in ***Stochastic Environmental Research and Risk Assessment*** by F. Corpas-Burgos and M.A. Martinez-Beneito (2020). This paper corresponds to Chapter 6 of this thesis.

The Spanish National Atlas of Mortality (ANDEES) developed, stated as the fourth and last objective of this thesis, has been published online and can be accessed at the following web link:

http://atlasnacional.fisabio.es

The methodology used in the development of such atlas together with some of its characteristics and results will be described in the Chapter 7 of this thesis. An enhanced version of Chapter 7 will be submitted for review and publication.

1.5. Software

The main tools used in the development of this thesis have been the statistical package `R` and the software for Bayesian analysis using Markov chain Monte Carlo (MCMC) methods `WinBUGS`. On the one hand, `WinBUGS` has been used to fit each of the models to the data and obtain the estimates of interest. On the other hand, `R` and some of its packages have been used to manage the data and results of each work. The main `R` packages used have been the `Pbugs` and `Shiny` libraries. `Pbugs` has made it possible to automate the call to `WinBUGS` from `R`, running the different chains that have been necessary in each of the adjusted models in parallel (different core processors), and therefore speeding up the computation time for obtaining the results. `Shiny` has allowed the development of the website that hosts the results of the Spanish National Atlas of Mortality (ANDEES).

For each of the published papers, the `R` code with the implementation of all the models fitted and for all the analyses performed is provided as supplementary material and can be found in the Appendix of this thesis and in https://github.com/pcorpas.

2. Methodology

In this chapter, we describe the general modeling framework for disease mapping studies and some of the most relevant models proposed to carry out the estimation of the geographical distribution of risks. Next, we show some limitations found in these models, after being used to analyze the distribution of mortality in different scenarios, and we summarize the modeling proposals developed in this thesis to solve these issues. These proposals will be explained in detail in Chapters 4, 5 and 6 in which each of the different published research articles are presented. Finally, we summarize the methodology used in the development of the Spanish National Atlas of Mortality (ANDEES) which will be described in detail in Chapter 7.

2.1. The disease mapping general modeling framework

In this section, we begin by describing the general modeling framework of disease mapping studies. First, we describe the univariate case in which the focus is on modeling the risks for a single disease. Next, we describe the multivariate case that will allow us to perform the joint spatial modeling of the risks of various diseases.

2.1.1. The univariate case

In disease mapping studies, the region of interest is considered divided into I contiguous geographic units, usually of small size, such as census tracts or municipalities. As already mentioned, the main objective of these studies is to determine the geographical distribution of the risks for some disease throughout the geographic region of interest. For this, the observed disease counts in each geographic unit $\{O_i : i = 1, \ldots, I\}$ are modeled as:

$$O_i \sim Poisson(E_i R_i), \quad i = 1, \ldots, I,$$

where E_i are the expected counts for each geographic unit, typically calculated by means of some age standardization (Martinez-Beneito and Botella Rocamora, 2019), and R_i are the corresponding risks that we would like to estimate. Regarding the modeling of this last term, the log-risks can be defined as:

$$log(R_i) = \mu + \eta_i, \tag{2.1}$$

where μ is an intercept, modeling the mean of the log-risks, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_I)'$ is a random effects vector. Random effects η_i are introduced into the model to allow the risks to vary for the different spatial units and are typically assumed to be spatially correlated as such variability is expected to exhibit this characteristic.

According to the previous expression, the risk in the i-th geographic unit would be equal to $exp(\mu+\eta_i)$. Therefore, this expression is what we call the *smoothed SMR* of such geographic unit under the spatial model considered. That is, the SMR is an epidemiological concept which is just equal to O/E for some population group, while the smoothed SMR is the result of the modeling of the risks throughout the region of study. That is, smoothed SMRs are just outputs of particular models but both

SMRs and smoothed SMRs try to estimate the same concept, the risk of the disease. Smoothed SMRs can be therefore considered enhanced (model-based) estimates of the risks in comparison to naive SMRs.

The random effects vector $\boldsymbol{\eta}$ is usually modeled using CAR prior distributions in order to induce spatial dependence on the SMRs and thus increasing the amount of information used to estimate them. A particularly popular case of CAR prior distributions is the Intrinsic CAR (ICAR) distribution (Besag et al., 1991) which can be defined as the following set of I univariate conditional distributions:

$$\phi_i|\boldsymbol{\phi}_{-i}, \sigma_\phi^2 \sim N\left(\frac{1}{w_{i+}}\sum_{k=1}^{I} w_{ik}\phi_k, \frac{\sigma_\phi^2}{w_{i+}}\right), \quad i = 1, \ldots, I. \tag{2.2}$$

In this expression, the subindex in $\boldsymbol{\phi}_{-i}$ denotes all the terms in $\boldsymbol{\phi}$ excepting its i-th component, w_{ik} weighs the contribution of the k-th random effect to the mean of ϕ_i, $w_{i+} = \sum_{k=1}^{I} w_{ik}$ and σ_ϕ^2 is a variance parameter. The dependence between elements of $\boldsymbol{\phi}$ is determined by the spatial weights w_{ik}, which are typically non-zero if areas i and k are considered neighbors and zero otherwise. Therefore, if two areas are considered neighbors, their random effects are conditionally dependent, while random effects of non-neighboring areas are conditionally independent.

A common assumption is to assume that the pair of areas (i, k) are neighbors if they share a common border (adjacency) and in that case set $w_{ik} = 1$ for all neighboring pairs of units (i, k). In that case, the conditional distributions above reduce to simply:

$$\phi_i|\boldsymbol{\phi}_{-i}, \sigma_\phi^2 \sim N\left(\frac{1}{n_i}\sum_{k\sim i} \phi_k, \frac{\sigma_\phi^2}{n_i}\right), \quad i = 1, \ldots, I, \tag{2.3}$$

where n_i stands for the number of neighboring areas of unit i and the subindex $k \sim i$ denotes all those units k which are neighbors of i. Now, the conditional mean of ϕ_i is equal to the raw (unweighed) mean of the random effects in the neighboring areas and its conditional variance is inversely proportional to the number of neighbors n_i.

Some relevant modeling proposals for $\boldsymbol{\eta}$

One of the most popular modeling proposals for $\boldsymbol{\eta}$ in disease mapping studies is that introduced in Besag et al. (1991) (BYM). In this proposal, the random effects vector $\boldsymbol{\eta}$ is considered to be the sum of two vectors of random effects $\boldsymbol{\eta} = \boldsymbol{\phi} + \boldsymbol{\theta}$. The term $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_I)'$, which follows an ICAR distribution, will be responsible for inducing spatial dependence on $\boldsymbol{R} = (R_1, \ldots, R_I)'$ and accounts for those risk factors of regional scope which take an effect on several contiguous spatial units, making them in principle similar. The second term, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_I)'$, whose components follow independent Normal distributions of mean zero and common variance σ^2_θ, accounts for risk factors of very limited geographical scope that take an effect just on isolated areal units, making their risks different to those of their surrounding units. Thus, this second term induces additional unstructured variability in $\boldsymbol{\eta}$. The amount of spatial/unstructured variability in $\boldsymbol{R}$ depends on the balance between σ_ϕ and σ_θ, which is determined by the model/data itself. If the first has higher (respectively lower) values, in comparison with the second, the final pattern will show substantial spatial dependence (respectively independence).

A second popular CAR prior distribution for inducing spatial correlation on the random effects vector $\boldsymbol{\eta}$ in Expression (2.1) is that introduced in Leroux et al. (1999). In contrast to the BYM model, $\boldsymbol{\eta}$ in this alternative proposal is not the sum of two additional components. In this case, the determination of the amount of spatial/unstructured variability is controlled by a spatial correlation parameter $\rho \in [0, 1]$ so that the special case of $\rho = 0$ simplifies to a model with independent

random effects and $\rho = 1$ corresponds to the ICAR distribution above. All intermediate values of $\rho \in (0, 1)$ induce patterns mixing both sources of dependence. Specifically, for the Leroux et al.'s proposal, the prior conditional distributions corresponding to η_i are given by:

$$\eta_i|\boldsymbol{\eta}_{-i}, \rho, \sigma_\eta^2 \sim N\left(\frac{\rho}{\rho w_{i+} + 1 - \rho}\sum_{k=1}^{I} w_{ik}\eta_k, \frac{\sigma_\eta^2}{\rho w_{i+} + 1 - \rho}\right), \quad i = 1, ..., I.$$

For the usual assumption of $w_{ik} = 1$ for adjacent spatial units, and 0 otherwise, the Leroux et al.'s proposal reduces to:

$$\eta_i|\boldsymbol{\eta}_{-i}, \rho, \sigma_\eta^2 \sim N\left(\frac{\rho}{\rho n_i + 1 - \rho}\sum_{k\sim i} \eta_k, \frac{\sigma_\eta^2}{\rho n_i + 1 - \rho}\right), \; i = 1, \ldots, I.$$

This expression makes clear the equivalence of this distribution to either independent Normal random effects or an ICAR distribution for $\rho = 0$ and $\rho = 1$, respectively.

Some relevant modeling proposals for data with zero excesses

As already mentioned in the introduction to this thesis, it is possible that standard disease mapping models, such as BYM and Leroux et al.'s proposal, underestimate the number of geographic units without observed cases of the disease. This lack of fit for the number of zeroes is commonly known as a problem of "zero excesses" in the data and leads to obtaining risk maps that are oversmoothed. Therefore, in such situations, it is necessary to consider specific "zero excesses" modeling strategies in standard disease mapping models to improve their fit. In the disease mapping literature, the most popular proposals to model "zero excesses" in data are the zero-inflated Poisson (ZIP) model and

the hurdle Poisson model. In its simplest form, the ZIP model assumes that the events observed in each geographic unit O_i follow a mixture of a degenerate distribution with all its mass at zero and a Poisson distribution with weights $1 - \pi^Z$ y π^Z, respectively. This inflates the number of zeroes expected by the Poisson distribution as a function of π^Z. On the other hand, the hurdle Poisson model assumes that the data follow a mixture of a degenerate distribution with all its mass at zero and a Poisson distribution truncated to take values above 0. That is, in contrast to the ZIP model, all the zeroes observed in the hurdle model are assumed to come from the zero-degenerate distribution. Therefore, the parameter $1 - \pi^H$ in the hurdle model, represents the probability that the number of observed cases in one unit is zero instead of the percentage of extra-Poisson zeroes, the interpretation of $1 - \pi^Z$ in the ZIP model. ZIP and hurdle models are often combined with specific disease mapping proposals, such as BYM or Leroux et al., in order to yield flexible spatial models accounting for zero excesses.

2.1.2. The multivariate case

In multivariate disease mapping, the goal is the joint modeling of the risks of various diseases. In that case, the observed counts for each geographic unit and disease $\{O_{ij} : i = 1, \ldots, I; j = 1, \ldots, J\}$ are modeled as:

$$O_{ij} \sim Poisson(E_{ij}R_{ij}), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J,$$

where E_{ij} are the expected counts and R_{ij} the relative risk for the i-th geographical unit and j-th disease. As in univariate modeling, log-risks can be defined as:

$$log(R_{ij}) = \mu_j + \theta_{ij}.$$

In this case, the term μ_j is just an intercept for the j-th disease and $\boldsymbol{\Theta} = \{\theta_{ij} : i = 1, \ldots, I; j = 1, \ldots, J\}$ is a collection of random effects whose joint distribution specifies how dependence is defined within and between diseases. Specifically, dependence among the columns of $\boldsymbol{\Theta}$ induces dependence between diseases and, similarly, dependence among its rows induces spatial dependence within diseases (geographical units). We will focus now on the $\boldsymbol{M}$-modeling proposal, which will be the multivariate model that we will paying particular attention in Chapter 5.

The Botella-Rocamora et al.'s M-model

An interesting proposal for modeling spatial dependence between geographic units and dependence between diseases is introduced by Botella-Rocamora et al. (2015). In this proposal multivariate spatial dependence is induced by setting

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\mathbf{M} \tag{2.4}$$

where $\boldsymbol{\Phi}$ is an $I \times K$ matrix of random effects with independently distributed columns that typically follow some spatially correlated distribution, such as BYM or Leroux et al.'s proposal. Those spatial distributions induce dependence between geographical units and therefore between rows of $\boldsymbol{\Theta}$. Additionally, $\mathbf{M}$ is a $K \times J$ random matrix which induces dependence between the different columns in $\boldsymbol{\Theta}$, that is, between the different diseases considered in the analysis. Usually $K = J$, although they could be different, such as for the multivariate formulation of the BYM model, where two random effects are included per disease and therefore $K = 2J$. The variance parameter of the random effects in the columns of $\boldsymbol{\Phi}$ is usually set to 1, since $\mathbf{M}$ cells are responsible for controlling the variability of $\boldsymbol{\Theta}$. Otherwise, those variances and the cells of $\mathbf{M}$ would not be identifiable as they would cancel each other out. On the other hand, as proposed by Botella-Rocamora et al., the cells of $\mathbf{M}$ are independently defined as $M_{ij} \sim N(0, \sigma^2)$ $i = 1, ..., K,\ j = 1, ..., J$,

where σ could be either a fixed (typically large) value, and therefore the M_{ij}s would follow vague independent prior distributions, or an additional variable to be estimated in the model. In the first case, we call the corresponding modeling *fixed effects* $\boldsymbol{M}$*-modeling*, since $\mathbf{M}$ cells would be modeled in that manner and, alternatively, we call the second case *random effects* $\boldsymbol{M}$*-modeling*, once again because of the modeling of the cells carried out in $\mathbf{M}$. A theoretical property of this model is that, as shown in the Botella-Rocamora et al. (2015) paper, assigning $N(0, \sigma^2)$ prior distributions to the entries in $\mathbf{M}$ yields a $Wishart(K, \sigma^2\mathbf{I}_J)$ prior distribution for the covariance matrix between diseases $\mathbf{\Sigma}_b$ when all spatial models share the same spatial distribution, which can be computed as simply $\mathbf{M}'\mathbf{M}$. Hence, the independent modeling of the cells of $\mathbf{M}$ entails a prior mean for $\mathbf{\Sigma}_b$ proportional to an identity matrix or, alternatively, it assumes prior independence in the columns of $\mathbf{\Theta}$.

2.2. Some limitations (and some potential solutions) of the previous models

So far, we have described the general modeling framework for disease mapping studies and some relevant models for deriving risk estimates in different scenarios. In this section, we describe some limitations evidenced by the previous proposals, once applied to some real contexts. We also summarize the different methodological proposals that we have developed to solve the limitations found, although the details of these proposals are described later, in the chapters that contain the corresponding original articles that have been published.

2.2.1. Modeling data with zero excesses

In order to examine the need to incorporate zero excesses modeling strategies in the analysis of mortality data and to evaluate the behavior

of the most popular models in this context, the BYM model and its corresponding ZIP and hurdle versions have been implemented on an extensive real setting. Specifically, these models have been applied to the study of the geographical distribution of mortality, for a total of 46 geographical patterns, at the municipal level in the Valencian Region during the period 1987-2006. We pay particular attention to the fit of the models in terms of the number of predicted zeroes, in comparison to those actually observed for each cause of death, and to the estimated SMRs for each of the models.

In summary, we find that the BYM model may fit quite poorly the number of zeroes for certain data sets. Namely, in more than 30% of the data sets considered the 95% posterior predictive intervals for the number of zeroes in BYM were far from containing the real observed zero counts, which were always above these intervals. With respect to the approaches with a particular treatment of zeroes, the results are not satisfactory either. Surprisingly, ZIP does not help much in fitting more zeroes and showed 95% posterior predictive intervals which do not contain the real observed number of zeroes in almost as many data sets as the BYM model. The risk maps for the different diseases estimated with these models showed hardly any difference. Conversely, for hurdle model, all intervals contained the observed number of zeroes. Nevertheless, this better fit of the proportion of zeroes yielded unrealistic estimates of the risks, showing strange polarized patterns in almost all the diseases analyzed and completely different to those obtained with BYM and ZIP. Thus, naive ZIP and hurdle models are clearly unsatisfactory. We show that this is a consequence of the implementation of these models where the π^Z and π^H parameters, modeling the probabilities of non-zeroes, are common to all municipalities. These results suggest the need for an explicit modeling of the probabilities that should vary across geographical units.

An alternative to a common zero excess probability is modeling unit-specific π_is (we will denote π when we refer indistinctly to either

π^Z or π^H) by means of, for example, logistic regression. This has been repeatedly done in the disease mapping context for both ZIP (Dalrymple et al., 2003; Gschlößl and Czado, 2008; Neelon et al., 2010; Musenge et al., 2013; Nieto-Barajas and Bandyopadhyay, 2013) and hurdle models (Dalrymple et al., 2003; Neelon et al., 2010, 2013; Upfill-Brown et al., 2014; Neelon et al., 2014; Arab, 2015). That is, following several of the proposals in the literature, for both ZIP and hurdle models we could consider:

$$logit(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \varphi_i \tag{2.5}$$

where $\boldsymbol{\beta}$ model the effect of some set of covariates $\mathbf{X}$ and $\boldsymbol{\varphi}$ is a vector of (possibly spatial) Gaussian random effects modeling the effect of those factors that cannot be explained by $\mathbf{X}$.

However, after exploring these more flexible modeling strategies, we have found important posterior impropriety problems in hurdle and ZIP models when the vector of probabilities $\boldsymbol{\pi}$ is modelled with either fixed or random effects and with non-informative (improper) prior distributions for these parameters. These impropriety problems in the posterior distributions shows up as huge MCMC convergence problems in the model parameters. One could think of using vague proper prior distributions, instead of improper priors, as a posible strategy to avoid impropriety issues. This is a procedure frequently found in the literature, supposedly to avoid MCMC convergence problems. However, as we demonstrate, in these cases the posterior distribution of the parameters will depend on the (arbitrary) vagueness of its prior distribution, which makes unadvisable that option. Therefore, we discourage the use of those models particularly in a non-informative or objective setting and we propose some valid alternatives with different π_is. These proposals are summarized below.

Some valid proposals for modeling π

We propose some (non-informative) procedures for modeling π that avoid the conditions for posterior impropriety stated above. Specifically,

we formulate 3 separate modeling proposals.

Fixed effects modeling:

A potentially valid modeling proposal (we will refer to this as *FE* (Fixed Effects) henceforth) would be to consider a hurdle model with

$$logit(\boldsymbol{\pi}^H) = \boldsymbol{X}\boldsymbol{\beta}.$$

A suitable proposal that could be used in principle for any disease mapping model would be: $\boldsymbol{X} = [\mathbf{1}_I, log(\boldsymbol{E})]$, where $\boldsymbol{E}$ stands for the vector of expected values used in the Poisson likelihood of hurdle models. We have taken the logarithm of the expected values to avoid any potential effect of the usually skewed distribution of this variable caused by the presence of very few large cities.

This proposal models the logit of the probabilities of non-zeroes as a function of the expected observations at each areal unit. This seems quite reasonable since units with lower expected counts would show more easily zero observed counts meanwhile those larger units will show positive counts in general. For this proposal we will consider an improper Uniform prior distribution for each component of $\boldsymbol{\beta}$.

Nested fixed effects modeling:

The use of expected values as a surrogate of the (population) size of the areal units in the FE modeling seems quite reasonable. Nevertheless, this does not depend at all on the probabilities of non-zeroes resulting from the Poisson side of hurdle models: $\boldsymbol{\pi}^P = 1 - exp(-\boldsymbol{\mu})$. These probabilities could be also used as sensible covariates for modeling the probabilities of non-zeroes for hurdle models $\boldsymbol{\pi}^H$, instead of just $\boldsymbol{E}$. Thus, our second proposal for modeling $\boldsymbol{\pi}^H$ in hurdle models would be

$$logit(\pi_i^H) = logit(\pi_i^P) + \gamma.$$

This would be an alternative fixed effects logistic modelling of $\boldsymbol{\pi}^H$ using $logit(\boldsymbol{\pi}^P)$ as an offset. The values of that offset would be leveraged by γ so that if it takes values close to 0 this model would reproduce the non-inflated probabilities in the Poisson layer, even for zero-counts, meanwhile for $\gamma < 0$ the zero-specific probabilities would be inflated in regards to the Poisson model. We will refer to this model as *NFE*, Nested Fixed Effects model. Once again we will consider an improper Uniform prior distribution for γ so that any potential posterior impropriety problem in this model appears.

Geometric modeling:

Since resorting to logit regression has proved to bring lots of problems into ZIP and hurdle models, we could try to avoid those transformations in order to make sensible proposals. Thus, making

$$\pi_i = 1 - (1 - \pi^G)^{E_i}$$

seems a reasonable proposal for both ZIP and hurdle models. For this proposal we would have that the probability of observing a zero count for a unit with n expected cases is $(1-\pi^G)^n$, where $1-\pi^G$ is that same probability for a unit with 1 expected case. This geometric progression also holds for the Poisson process where the probability of observing zeroes with n expected cases $exp(-n\ \lambda) = exp(-\lambda)^n$ follows that same relationship. Thus, the probabilities of zero counts for this proposal are in agreement with the Poisson side of the model. For π^G, which can be interpreted as the probability of observing a positive count for units with one expected case, we set an Uniform prior distribution between 0 and 1. Since this prior is proper we avoid any posterior impropriety coming from this term. We will refer to the ZIP and hurdle versions of this model henceforth as *ZGeo* and *HGeo* respectively.

2.2.2. Heterocedastic multivariate modeling

The multivariate proposal presented by Botella-Rocamora et al. (2015) has been implemented to study the geographical distribution of mortality in the cities of Alicante, Castellón and Valencia composed of 215, 95 and 553 census tracts (the geographical unit for this analysis), respectively. We consider the multivariate joint spatial modeling of 20 different causes of mortality and both *fixed* and *random effects* $\boldsymbol{M}$*-models* with underlying BYM spatial patterns for all three cities separately. In order to evaluate the benefits of multivariate modeling, we compare the results obtained against those derived with independent BYM patterns for each disease.

In the case of the city of Castellón, markedly different risk maps are obtained with the multivariate fixed and random effects $\boldsymbol{M}$-models, as compared to the univariate BYM models. On the one hand, while univariate modeling generally provides maps with low variability for most of the diseases in the study, except in a few cases such as AIDS, fixed effects $\boldsymbol{M}$-modeling provides maps with great variability in all diseases, with hardly any smoothing, which resemble the corresponding maps of unsmoothed SMRs. The lack of smoothing of fixed effects $\boldsymbol{M}$-model is noticed, but to a much lesser extent, in the results drawn from Alicante and Valencia. On the other hand, we observed that random effects $\boldsymbol{M}$-modeling estimates in Castellón yield flat risk maps for all 20 diseases studied, which in a few cases, such as AIDS, are very different from those estimated with univariate modeling. Strikingly, this performance was only seen in Castellón, but not (or not so evidently) in Alicante or Valencia.

In the case of the cities of Alicante and Valencia, the differences found between the fixed and random effects risk patterns were much milder. For these two cities, both multivariate models take advantage of the additional information provided by the set of diseases considered, depicting more detailed spatial patterns in general than their univariate alternatives. This suggests that the results found for Castellón could

be due to the smaller size of this city, where the prior structure that the $\boldsymbol{M}$-model induces could be more influential than in Alicante and Valencia. Thus, the prior covariance structure of the $\boldsymbol{M}$-model could be having an undesirable effect on the final fit that, when available data are weaker, might be influencing the spatial patterns determined.

Next, we try to give a statistical explanation of these results. Regarding the fixed effects $\boldsymbol{M}$-model, we have mentioned that it was equivalent to assuming a $Wishart(K, \sigma^2\mathbf{I}_J)$ prior distribution on the covariance matrix between diseases $\boldsymbol{\Sigma}_b$. Since σ is usually set to a large value for the fixed effects approach, this entails that the prior mean of $\boldsymbol{\Sigma}_b$ is equal to $K\sigma^2 I_J$, for a high value of σ. Therefore, this model assumes the prior covariances between diseases to be centered at 0 and the prior variances of the log-risks for each spatial pattern to be high. These prior assumptions could explain the results found in Castellón for the fixed effects model, where the prior information in $\mathbf{M}$ could overwhelm the information provided by the data. For this city, the cells of $\boldsymbol{\Theta}$ do not produce any smoothing in the risks fitted, as a consequence of their large prior variances (subsumed in matrix $\mathbf{M}$), which does not produce any shrinkage. As a consequence, the smoothed SMRs estimated for this model reproduce the unsmoothed original SMRs that disease mapping models typically try to avoid.

The random effects $\boldsymbol{M}$-model also leads to a prior mean of $K\sigma^2 I_J$ for $\boldsymbol{\Sigma}_b$ but with σ now being a parameter to be estimated within the model. In this case the prior mean will just be proportional to the identity matrix but the proportionality constant will be estimated by the model itself, which will be set to a common consensus value for all the diseases. Univariate BYM models for each of the diseases in Castellón yielded posterior standard deviations for the log-SMRs ranging from 0.05 to 0.42, depending on the disease. AIDS was the disease with a higher standard deviation, far larger than the median standard deviation for the set of diseases considered (0.13). Thus, the distribution of the standard deviations of the log-SMRs for the different diseases has a

pronounced asymmetrical right-tailed distribution. In consequence, the consensus scale parameter σ for the random effects model takes a value that is much lower than that required to appropriately describe the spatial variability of AIDS mortality. This could explain perfectly why the initial pattern highlighted by the univariate BYM model for AIDS vanishes when the random effects $\boldsymbol{M}$-model is fitted.

In sum, the Castellón multivariate mortality study above has shown important prior sensitivity problems for the $\boldsymbol{M}$-model, mainly for smaller data sets. Specifically, the fixed effects $\boldsymbol{M}$-model has a tendency to yield unsmoothed risk estimates. Furthermore, the random effects version has an inclination toward the shrinkage of all diseases to a common point in terms of variability. Our proposal for fixing the prior sensitivity problems of the $\boldsymbol{M}$-model consists in a modification of its random effects version. Specifically, we relax the assumption of a common scale parameter for the cells of $\mathbf{M}$. In particular we propose two different ways to do this. The first proposal considers $M_{ij} \sim N(0, \sigma_i^2)$ for $i = 1, ..., K$, while our second alternative proposal considers $M_{ij} \sim N(0, \sigma_j^2)$ for $j = 1, ..., J$. We will refer to these two proposals as the *row variance-adaptive* random effect $\boldsymbol{M}$-model (or simply RVA $\boldsymbol{M}$-model) and the *column variance-adaptive* random effect $\boldsymbol{M}$-model (or simply CVA $\boldsymbol{M}$-model), respectively. Obviously these two proposals will be more adaptive in terms of variability than the original random effects $\boldsymbol{M}$-model, which will allow us to solve the shrinkage problems toward a common variability evidenced in the random effects $\boldsymbol{M}$-model. We will refer to this latest model as the *non variance-adaptive* model (NVA model).

2.2.3. Adaptive spatial modeling

CAR distributions are commonly used to model spatial dependence between nearby geographic units in disease mapping studies. These distributions induce spatial dependence by means of a spatial weights matrix that quantifies the strength of dependence between any two

neighboring spatial units. As previously described, the most common procedure for defining that spatial weights matrix is using an adjacency criterion. In that case, all pairs of spatial units with adjacent borders are given the same weight (typically 1) and the remaining non-adjacent units are assigned a weight of 0. However, assuming all spatial neighbors in a model to be equally influential could be possibly a too rigid or inappropriate assumption. This imposes all neighboring regions to be equally influential on any particular risk, which may not correspond to reality. In order to overcome this limitation, we propose a procedure for estimating the spatial weights matrix in disease mapping studies. Specifically, we propose an adaptive extension for both ICAR and Leroux et al. spatial distributions in which the spatial weights for adjacent areas are additional random variables in the model, allowing variability between them, and estimates are based on the information provided by the data.

We start first by introducing the estimation of spatial weights matrices for ICAR distributions. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_I)'$ be a random effects vector with ICAR distribution, that is:

$$\boldsymbol{\phi}|\sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2(\boldsymbol{D} - \boldsymbol{W})^-).$$

For this expression, we will assume that $\boldsymbol{D}$ and $\boldsymbol{W}$ are defined according to adjacency between spatial units, i. e. $\boldsymbol{D} = diag(n_1, \cdots, n_I)$ for n_i the number of neighbors of unit i and $\boldsymbol{W} = (w_{ik})$ where $w_{ik} = 1$ if (i, k) are adjacent units and 0 otherwise.

Let us now consider a random vector $\boldsymbol{c} = (c_1, \cdots, c_I)'$ of positive values, a new spatial weights matrix $\boldsymbol{W}^*(\boldsymbol{c}) = diag(\boldsymbol{c})^{1/2}\boldsymbol{W}diag(\boldsymbol{c})^{1/2}$ and $\boldsymbol{D}^* = diag(w_{1+}^*, ..., w_{I+}^*)$. With this, we propose the following adaptive CAR prior distribution:

$$\boldsymbol{\phi}|\boldsymbol{c}, \sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2(\boldsymbol{D}^* - \boldsymbol{W}^*(\boldsymbol{c}))^-)$$

$$c_i \sim Gamma(\alpha, \alpha).$$

The elements of the vector $\boldsymbol{c}$ are assumed to be positive since the non-zero weights of the new spatial weights matrix $\boldsymbol{W}^*$ are $w^*_{ij} = (c_i c_j)^{1/2}$ so, in this manner they all will be well defined and positive. Accordingly, we have used a Gamma prior distribution for its elements, which seems a natural choice. The Gamma distribution considered has mean 1, in accordance with the value of the non-zero cells of $\boldsymbol{W}$ when an adjacency criterion is considered. Thus, $\boldsymbol{W}^*(\boldsymbol{c})$ will be on average equal to $\boldsymbol{W}$, although its non-zero weights will not necessarily have to be equal to 1. Hence the new adaptive distribution will be more flexible than the regular ICAR distribution. Note that, as defined, the (prior) standard deviation of any element of $\boldsymbol{c}$ is equal to $\alpha^{-0.5}$, which could guide us to set a prior distribution for this parameter. In fact, we have considered a prior Uniform distribution on $\alpha^{-0.5}$, with lower and upper limits intended to make it vague, in order to complete the hierarchical structure above.

Alternatively, the definition of the adaptive ICAR distribution above could be restated as a set of conditional distributions $\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{c}, \sigma^2_\phi,\ i = 1, ..., I$, of mean

$$E[\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{c}, \sigma^2_\phi] = \frac{1}{w^*_{i+}} \sum_{k=1}^{I} w^*_{ik}\phi_k = \frac{c_i^{1/2} \sum_{k\sim i} c_k^{1/2}\phi_k}{c_i^{1/2} \sum_{k\sim i} c_k^{1/2}} = \frac{\sum_{k\sim i} c_k^{1/2}\phi_k}{\sum_{k\sim i} c_k^{1/2}} \tag{2.6}$$

and variance

$$Var[\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{c}, \sigma^2_\phi] = \frac{\sigma^2_\phi}{w^*_{i+}} = \frac{\sigma^2_\phi}{c_i^{1/2} \sum_{k\sim i} c_k^{1/2}}. \tag{2.7}$$

These two expressions provide some quite valuable insights on the model proposed. The expected value in Expression (2.6) is just a weighted mean of the random effects for the corresponding neighbors. The weights in that expression are given by the vector $\boldsymbol{c}$, thus if c_i had

a low value for some i, that region will have a low contribution to the means of its surroundings units. Additionally, Expression (2.7) suggests that if c_i is low, then the conditional variance of ϕ_i will be in contrast high. Thus, if c_i was low, these two expressions suggest that it is as if spatial unit i was "disconnected" from its spatial neighbors, since ϕ_i will be less influential on them and will have higher variance, allowing it to move independently from the rest of the units. Conversely, if c_i was high, unit i will become more influential on its neighbors and will take a value in close agreement with them. Therefore, in some manner, the adaptive ICAR distribution would impose a tighter dependence between this unit and its neighbors.

In the case of the Leroux et al. model, $\boldsymbol{\phi}$ is distributed as:

$$\boldsymbol{\phi}|\rho, \sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2((1-\rho)\boldsymbol{I}_I + \rho(\boldsymbol{D} - \boldsymbol{W}))^-).$$

Following the development above, let us assume

$$\boldsymbol{\phi}|\rho, \boldsymbol{c}, \sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2((1-\rho)diag(\boldsymbol{c}^{1/2}) + \rho(\boldsymbol{D}^* - \boldsymbol{W}^*(\boldsymbol{c})))^-),$$

where $\boldsymbol{D}^*$ and $\boldsymbol{W}^*(\boldsymbol{c})$ are defined as for the adaptive ICAR distribution. In this manner, for $\rho = 1$ this distribution would be equivalent to an adaptive ICAR distribution, while for $\rho = 0$ it would yield a collection of independent Normal random effects with adaptive (heteroscedastic) variance. For that proposal, the conditional mean and variance of the

random effect ϕ_i can be expressed as:

$$\begin{aligned} E[\phi_i|\boldsymbol{\phi}_{-i},\rho,\boldsymbol{c},\sigma_\phi^2] &= \frac{\rho}{(1-\rho)c_i^{1/2}+\rho w_{i+}^*}\sum_{k=1}^{I} w_{ik}^*\phi_k \\ &= \frac{\rho c_i^{1/2}}{(1-\rho)c_i^{1/2}+\rho c_i^{1/2}\sum_{k\sim i} c_k^{1/2}}\sum_{k\sim i} c_k^{1/2}\phi_k \\ &= \frac{\rho}{1-\rho+\rho\sum_{k\sim i} c_k^{1/2}}\sum_{k\sim i} c_k^{1/2}\phi_k \end{aligned}$$

and

$$Var[\phi_i|\boldsymbol{\phi}_{-i},\rho,\boldsymbol{c},\sigma_\phi^2] = \frac{\sigma_\phi^2}{(1-\rho)c_i^{1/2}+\rho w_{i+}^*} = \frac{\sigma_\phi^2}{c_i^{1/2}(1-\rho+\rho\sum_{k\sim i} c_k^{1/2})}.$$

Considering I new parameters in the model to allow different strength of spatial dependence between neighboring geographic units leads to a considerable increase in the number of parameters to be estimated. As a consequence, data in univariate disease mapping models may be not strong enough as to make inference on vector $\boldsymbol{c}$ possible. For this reason, we propose the use of adaptive spatial weight matrices in a multivariate disease mapping context so that the spatial dependence structure between spatial units is shared and estimated from a sufficiently large set of mortality causes. This spatial weight matrix, which should capture the geometric/demographic/geographic features of the region of study, will be common to all the diseases involved in the study. This formulation would allow an appropriate estimation of the vector $\boldsymbol{c}$ and therefore an appropriate estimation of the weights matrix $\boldsymbol{W}^*(\boldsymbol{c})$ corresponding to the set of diseases and region of study considered.

2.3. Development of the Spanish National Atlas of Mortality (ANDEES)

The Spanish National Atlas of Mortality (ANDEES) is an interactive web application that allows for the visualization of the spatial and spatio-temporal distribution of mortality throughout the whole of Spain during the period 1989-2014. ANDEES considers the municipality (8,063 for the whole of Spain) as unit of analysis and studies by separate 102 causes of death for both men and women.

Two data sets have been used for the development of this atlas. The first of them contains all the deaths occurred in Spain during the period 1989-2014. These data have been provided by the Spanish National Statistics Institute (INE) and tabulated according to sex, five-year age group (considering a final age group of 85 years or more), municipality and cause of death, for the total period 1989-2014 and considering eight triennial periods, from 1991-1993 to 2012-2014. The second data set contains information on the population (number of people at risk) in the region and study period. This information has been obtained from the municipal population registers and tabulated according to sex, five-year age group, municipality and cause of death, for the total period 1989-2014 and for the eight mentioned triennial periods. Population data for each age group, sex and municipality have been used to calculate the number of deaths that would be expected in each municipality, if the risks for each age group were the same as in the reference population, the whole country.

The mortality indicators represented in the atlas maps have been estimated from the observed and expected deaths in each municipality, for all the period and the study subperiods. The main mortality indicators represented in the maps have been the smoothed SMRs. Smoothed SMRs have been estimated using spatial and spatio-temporal smoothing models. On the one hand, the BYM model has been used to obtain smoothed SMRs for the whole period of study. On the other

hand, due to the great length of the study period, it is also interesting to study mortality over shorter periods of time and analyze its evolution over time. To do so, the observed and expected events in each of the municipalities of Spain, disaggregated in the eight triennial periods already mentioned, have been considered. The use of spatio-temporal models allows the study of these disaggreagetd periods, providing several estimates for each municipality, instead of a single risk estimate corresponding to the entire period. In this way, the bias that occurs when considering risks as static amounts over time is avoided, since it is probable that some temporary change may have occurred in them (Ocaña Riola, 2007).

Within the spatio-temporal modeling literature, an interesting proposal to model spatial and spatio-temporal dependence between geographic units and study periods is that suggested by Martinez-Beneito et al. (2008). The spatio-temporal modeling of the data in this atlas has been carried out using this proposal. The spatio-temporal model proposed by Martinez-Beneito et al. (2008) consists of the integration of spatial smoothing and time series models, specifically auto-regressive processes, to simultaneously model the spatial and temporal dependence that observations may show. In this model, a spatio-temporal structure is defined in which the risks are spatially and temporally dependent at the same time, allowing nearby places to have similar spatial and temporal evolutions.

Spatial and spatio-temporal models have been run in `WinBUGS` using the statistical software `R` and some of its packages. The enormous volume of data to analyze and the number of estimates to be obtained in each analysis have made this project a challenge from a computational point of view. Some tools have been necessary to reduce the computing time in obtaining the results. Specifically, we have made use of the `Pbugs` package of `R`, which has allowed us to automate the calls to `WinBUGS` from `R` and to run them in parallel, thus speeding up computations.

Once the posterior distributions of the smoothed SMRs were

estimated from the BYM and the Martinez-Beneito et al.'s proposal, we calculated their posterior means for each municipality and period (in the case of spatio-temporal modeling). Additionally, the mass of each of these distributions above 1 was also calculated as a confidence measure for the risk excesses shown by our study. Both the smoothed SMR and the P(SMR>1) for each municipality, sex and cause of death are shown in the maps drawn in ANDEES.

The web application that contains the results and allows their visualization in ANDEES has been developed using the `Shiny` package of R (Chang et al., 2020). Nowadays, this package is becoming very popular and several applications for spatial and spatio-temporal data analysis and visualization have already been developed (Moraga, 2017; Adin et al., 2019a; Moraga, 2019). The application enables user interaction through several control widgets (mainly selectors to specify sex, cause of death and study period and configure the representation of the results) and creates interactive visualizations of the data and results.

The main results that can be visualized in ANDEES are:

- Maps with the estimated smoothed SMRs and probabilities of risk excess for the different causes of death, sex and study period. The created maps support interactive panning and zooming which is very convenient for exploring in detail particular areas. In addition, when clicking on a municipality, information appears with its name, the province in which it is located and the estimates obtained.

- Line plots showing the temporal evolution of risks, for the selected sex and cause of death, at the national and provincial levels throughout the different subperiods that make up the whole study period. These plots include interactive features such as panning, zooming and series highlighting.

– Data tables containing the estimates of interest. These tables support filtering, pagination and sorting which is very helpful in situations where we wish to locate the information from one particular municipality or show the municipalities with highest or lowest values.

All these results can be downloaded by users.

The maps with the risk patterns for each analysis are shown interactively thanks to the use of some functions available in the `Leaflet` package (Cheng et al., 2019). Some of these functions have had to be optimized in order to speed up the rendering time of the maps when modifying the different selectors. Line plots with the evolution of the risks in the different subperiods of the study have been built with the `Plotly` package (Sievert et al., 2020) and data tables for displaying the estimates of interest are shown by making use of the `DT` package (Xie et al., 2020).

3. Main results

In this chapter, we summarize the main results of the different modeling proposals developed in this thesis. Such results will be described in more detail in the following chapters in which the published research articles are presented. Finally, we briefly describe some geographical mortality patterns estimated and shown in ANDEES.

3.1. Some findings on zero-inflated and hurdle Poisson models for disease mapping

We fitted the 4 proposed models for dealing with zero excesses: FE, NFE, HGeo and ZGeo, to the previously mentioned mortality data set (46 geographical patterns) at the municipal level in the Valencian Region. First, we evaluate the models fit in terms of the number of predicted zeroes by any of those models in comparison to those actually observed for each disease. As a summary, we observe that the posterior predictive distribution for the number of zeroes for all 4 models agree with the real observed zeroes for each disease. Namely, all 3 hurdle models yield similar results to the naive hurdle model, being the posterior predictive median for the number of zeroes always very close to the real observed zeroes. The zeroes modeling in ZGeo yields a great improvement over naive ZIP models since for ZGeo the posterior predictive median for the number of zeroes is always very close to the

real observed zeroes. All 95% prediction intervals for the number of zeroes for all diseases and models contain the real observed zeroes, as would be expected in models which perform an explicit modeling of that particular feature in the data.

Second, we also compare the fit of the proposed models in general terms by using the Deviance Information Criterion (DIC). Regarding the FE model, its DIC is higher than that of the BYM model for 43 out of 46 data sets, so its performance in general does not seem very satisfactory. This suggests that the modeling proposed in FE is worse than that of the BYM model. As a consequence we will not pay further attention to this model. The NFE model attains better DICs for 11 out of the 15 data sets identified as having zero excesses for naive BYM models. Meanwhile, NFE attained lowed DICs than BYM for just 5 out of the remaining 31 data sets with no evidence of zero excess, as could be expected since BYM is less complex than NFE and for these data sets NFE should not yield any improvement. Thus, NFE attains in general lower DICs in those settings where it would be expected. HGeo attained lower DICs than BYM for 6 out of 15 data sets needing a particular treatment for zeroes and 2 out of 31 times when that treatment was not required in principle. Finally, ZGeo obtained similar results to HGeo, improving BYM in 5 out 15 times where zero excesses were evidenced and 8 out of 31 times when these were not so evident. Thus, the results of Geometric models are overall satisfactory although not as good, in terms of DIC, as those of NFE.

Finally, we compared the smoothed SMRs choropleth maps for BYM, NFE, HGeo and ZGeo. Both hurdle maps (NFE and HGeo) modify the risks mainly in those regions less populated and more prone to show zeroes, decreasing their risks in order to get those extra zeroes required. In contrast, regions having high SMRs hardly show any change. ZGeo introduces more differences with regard to BYM. New regions with both high and low risks emerge for this model. In view of these results, we recommend to use NFE as benchmark proposal among all those

introduced in this thesis to address "zero excesses" problems. We have found particularly satisfactory that NFE shows a better performance in terms of DIC than the rest of models. Moreover, this model seems to yield the risk patterns more similar to those of BYM, but with the zeroes issue fixed, what makes it seem the safest option.

3.2. On the convenience of heteroscedasticity in highly multivariate disease mapping

We implement the new RVA and CVA variance-adaptive proposals to study the geographic distribution of mortality at the census tract level in the cities of Alicante, Castellón and Valencia. We consider the multivariate joint spatial modeling of 20 different causes of death independently for all three cities. In order to evaluate those proposals, we compare the new estimated risks with those obtained with the original NVA $\boldsymbol{M}$-model proposed in Botella-Rocamora et al. (2015) and the univariate BYM models.

In the case of AIDS mortality in Castellón, the new modeling proposals provide risk maps with greater variability than that obtained with the NVA model and closely similar to that estimated with the univariate BYM model. For the rest of the diseases, the risk maps estimated with the new modeling proposals present a considerable lower variability than the risk map for AIDS. This shows that both RVA and CVA have solved the problems evidenced by the original multivariate NVA model, which provided risk maps with a similar variability for all the diseases in the study. Nevertheless, the original patterns estimated by the univariate BYM models seem to be reinforced for both the RVA and the CVA models, almost certainly as a consequence of sharing information between diseases. RVA and CVA estimates for Valencia and Alicante confirm the visual conclusions also drawn for Castellón,

although maybe to a lesser extent, since data for these cities are stronger than for Castellón. Thus, the results for these larger cities are far more robust to the multivariate model used to smooth the risks.

Besides the visual comparison of the estimated risk maps with the different modeling proposals, we have also compared the fit of these models in general terms by using the Deviance Information Criterion (DIC). We observe that the model that provides a better fit in terms of DIC for all three cities is the RVA $\boldsymbol{M}$-model, followed by the CVA $\boldsymbol{M}$-model in two out of the three cities in the study. This seems to confirm that, besides the evident visual differences found, the heteroscedastic nature of the RVA and CVA models yields an important enhancement also in terms of the predictive fit of the models.

3.3. On the use of adaptive spatial weight matrices from disease mapping multivariate analyses

We evaluate the performance of the BYM and Leroux et al. models with adaptive spatial weights. The main data set for this analysis corresponds to the observed deaths in the Valencia city census tracts, for a total of 15 different mortality causes in men for the period from 1996 to 2015. First, we estimate the weights matrix for the Valencia census tracts, which reflect the dependence structure of the mortality causes considered over the whole city. Subsequently, we use the estimated spatial structure matrix in posterior univariate analyses in order to assess the improvement that its use could bring, in comparison to the traditional adjacency criterion that assumes fixed weigths, equal to 1, for each adjacent pair of units.

For the BYM model, the values of the estimated spatial weights (their posterior means) range from 0.098 to 2.042, with a mean value

of 1.240, while for the Leroux et al. model these values range from 0.027 to 1.886, with a mean value of 1.264. Both adaptive proposals of the BYM and Leroux et al. models estimate a closely similar spatial dependence structure for the region of study. The correlation between the estimated spatial weights for the adaptive BYM and Leroux et al. models is 0.956. We observe that the census tracts with lowest spatial weights have certain peculiarities that make them special with respect to their adjacent units. On the one hand, residential homes for elderly or socially excluded people are frequently located in some of those "special" census tracts. As a consequence, these units often show higher observed deaths than expected for most of the mortality causes considered, which makes them exhibit a different behavior from those of their neighbors. On the other hand, new building areas and socially marginal regions of the city also frequently show lowest spatial weights. The use of a broad set of mortality causes, with 15 diseases, has allowed the models to identify those census tracts with these particularities that lead them to exhibit a very particular behavior in terms of mortality. That behavior requires an adaptation of the spatial weights matrix, otherwise their risks would be oversmoothed and estimated more similarly to their neighbors than they should. The high values of the spatial weights vector seem to be used to connect more tightly those regions of the map in the outskirts that would otherwise have an excessively independent behavior, preventing them from being isolated. Thus, the adaptive proposal run seems to change some geometric properties of the graph that could make some census tracts less connected to the rest of the graph than would be desirable.

Once the spatial weights matrix of the spatial random effects has been estimated for a region, it could be used for subsequent univariate disease mapping analyses on that same region. We assess that procedure on our data set comparing it with the use of the spatial weights matrices estimated with the most traditional procedure which uses the adjacency criterion. Specifically, for all 15 diseases in our data set we have fitted univariate BYM and Leroux et al. models assuming either the spatial

dependence structure estimated from the multivariate analysis above or the traditional adjacency-based weights matrix. Next, we compare the results of both analyses for each mortality cause according to the smoothed SMRs of both alternatives and also according to the Conditional Predictive Ordinate (CPO) and the Deviance Information Criterion (DIC) of each model.

In order to make a fair comparison, avoiding the use of the data twice (once for estimating spatial weights and once for estimating the smoothed SMRs with the corresponding univariate models), we have used different spatial weights in our comparisons. For each mortality cause, we have estimated spatial weights with a multivariate study of 14 diseases, all excepting the corresponding disease, which avoids using the data twice for the univariate (preestimated) adaptive analyses.

Both models provide risk maps with similar spatial patterns. However, the model using the adaptive spatial weights reproduces higher variability than its adjacency-based alternative allowing some of Valencia's neighborhoods to be reproduced more clearly and making it possible for some census tracts to reproduce more extreme risks. Thus, the adaptive weights avoid the excessive smoothing of the smoothed SMRs by allowing additional flexibility. Afterwards, we have compared the fit of the adaptive vs. the non-adaptive weights models according to the CPO and DIC criterion. According to DIC (CPO), we observe that BYM and Leroux et al. models with adaptive weight matrices provide a better fit than the corresponding adjacency based model in 14 (13) and 13 (9), respectively, out of the 15 mortality causes considered. This confirms that the greater flexibility of the adaptive models really improves the smoothed SMRs estimates in comparison to the traditional adjacency-based analyses.

3.4. Some interesting mortality geographic patterns found in ANDEES

In this section, we summarize some of the main results shown in ANDEES. Specifically, we describe the estimated mortality geographic patterns for all causes of death joinly and for some specific causes of particular interest, separately for men and women.

3.4.1. All-causes mortality

All-causes mortality risk maps for the whole period 1989-2014 show a territorial distribution marked by a north-south pattern. In the case of men, the highest mortality occurs especially in the southwestern half of the peninsula (Extremadura and western of Andalucía). In contrast, the areas with the lowest mortality are found in the northern Meseta and below the Pyrenees. In the case of women, the highest mortality occurs in the southern half of the peninsula, especially in the southern and western part of Andalucía (Huelva, Sevilla, Cádiz and Málaga). On the other hand, the areas with the lowest mortality are once again located in the northern half of the Meseta and south of the Pyrenees. This higher mortality from all causes in men and women in southern areas of the country compared to mortality in the northern zone could be reflecting, in part, the existing socioeconomic inequalities between these regions (Benach and Yasui, 1999). The temporal evolution of mortality at the provincial and national levels from all-causes in men and in women shows a clear downward trend throughout the study period. This overall downtrend in general mortality throughout the whole country could be explained by the public health and sanitary improvements that have occurred over the years of the long period of study. The spatio-temporal results show how, in general, the north-south geographic pattern found is maintained in all the subperiods of the study, and areas with an evolution different from the general one are not found.

3.4.2. Malignant tumor of the trachea, bronchi and lung mortality

Regarding mortality risk maps for malignant tumor of the trachea, bronchi and lung in the period 1989-2014, the areas with the highest risks for this cause in men are mainly concentrated in the southwestern part of the peninsula (Extremadura, Huelva, Sevilla and Cádiz), also in some areas of the Mediterranean coast and in zones of Asturias. In contrast, the areas with the lowest mortality risks are located in the northern half of the peninsula, as well as the northeast of Andalucía, east of Castilla-La Mancha and the Canary Islands. On the contrary, for women the areas with the highest mortality due to this cause of death appear scattered, highlighting mainly the Canary Islands, some areas of Madrid, Bizkaia and Pontevedra, and coastal zones of Málaga, Alicante and western Mallorca. Many of these correspond to regions with a high presence of residential tourism. The evolution of malignant tumor of the trachea, bronchi and lung mortality at the provincial and national levels in men remains stable in most provinces in the periods between 1991 and 2008, exhibiting a slight downward trend from 2008. The decrease in risks in the provinces with the highest mortality (Huelva, Sevilla, Cádiz, Cáceres and Badajoz) is observed from 1996 and is more pronounced than in the other provinces. In the case of women, mortality shows a clear upeard trend from the year 2000. The spatio-temporal results show how, in general, the geographical mortality pattern found for each sex for the entire study period is maintained with slight variations in the different subperiods.

3.4.3. Malignant tumor of the stomach mortality

The areas with the highest mortality risks for malignant tumor of the stomach in men during the period 1989-2014 are mainly concentrated in Castilla y León and some neighboring areas. The Galician Atlantic coast, Cáceres, Ciudad Real and areas of the northern interior of Cataluña also show high mortality. The lowest relative risks are concentrated on the

Canary and Balearic Islands, and more dispersed over the Mediterranean coast. In the case of women, the geographical distribution shows a similar pattern to that of men, although it extends to a greater extent in areas of southern Galicia, northern Cataluña and Ciudad Real. The evolution of mortality risks at the provincial and national levels in men and in women shows a downward trend. The spatio-temporal results show how the geographic pattern found in both sexes is maintained in the different subperiods of the study, and areas with a different evolution of the general one are not found.

3.4.4. Diabetes mellitus mortality

Diabetes mellitus mortality in men during the period 1989-2014 shows a lower mortality in the northern half of the peninsula. The areas with the highest mortality are especially concentrated in the Canary Islands, Sevilla, Cádiz and zones of Jaén, Ciudad Real and Valencia, while those with the lowest mortality are located in the eastern provinces of Castilla y León, Galicia and Madrid. In the case of women, the territorial distribution also follows a north-south pattern, where the areas with the highest mortality are in the south: Extremadura, Andalucía, southern Castilla-La Mancha, Murcia, Valencia and the Canary Islands. Low risk areas are observed in the northeast of the Meseta (Soria, Segovia, Burgos) and some areas of Teruel, León and Galicia. In short, we find that mortality risks from diabetes in both sexes increase considerably from north to south. The municipalities of the Canary Islands are those that show the highest relative risks in Spain.

The risks at the provincial and national levels in men remain practically stable throughout the period between 1991 and 2005. Starting in 2005, we observed a slight downtrend, except in Las Palmas and Santa Cruz de Tenerife where the trend is upwards. In the case of women, the evolution of the risks shows a downward trend throughout the study period except in Las Palmas and Santa Cruz de Tenerife where the risks keep stable. The spatio-temporal results show the

existence of zones with a different temporal evolution in the different subperiods of the study compared to the results for the entire period.

3.4.5. Leukemia mortality

Leukemia mortality in men in the period 1989-2014 shows very little variability. This flat pattern, far flatter than that of most diseases, is the most striking feature of this mortality cause. We find that the areas with the lowest relative risks are found in zones of the Canary and Balearic Islands, Galicia and eastern of Castilla y León. On the contrary, the areas with higher risks are found in Cáceres, Barcelona and Córdoba. In the case of women, the geographic pattern of mortality from leukemia is completely flat, with no areas with particularly higher/lower risks as compared to the whole country. In view of these results, we conclude that leukemia mortality is distributed homogeneously throughout the national territory. The temporal evolution of leukemia mortality in men and in women shows a downward trend throughout the study period. The spatio-temporal results show how the risk maps for both sexes keep constant in time with hardly any variation in the different subperiods of the study.

3.4.6. AIDS mortality

Finally, we describe the AIDS mortality geographic pattern in men for the period 1989-2014. The mortality risk map from AIDS in women is not available since this cause of death does not have a sufficient number of cases to be considered in the study. The areas with the highest mortality risks for this cause in men are mainly concentrated in zones of Sevilla and the Andalusian coast, Valencia and the eastern coast, Asturias, Madrid and Barcelona. Thus, this disease is mostly typical of urban and costal municipalities. The AIDS mortality temporal evolution in men shows a significant increase during the periods 1991-1993 and 1994-1996, but for all the subsequent periods the mortality trend is

decreasing. The spatio-temporal results show the existence of zones with a different temporal evolution in the different subperiods of the study. We found that the municipalities with a steeper decrease are located in Cataluña, Valencia, Madrid, Vizkaia, Guipuzkoa and Navarra. On the contrary, we find that the municipalities with a milder risk decrease are located in Andalucía, Extremadura, western Castilla y León and Galicia. Thus, AIDS mortality seems to move during the period of study from more urban areas to areas with lower socioeconomic status.

4. Some findings on zero-inflated and hurdle Poisson models for disease mapping

In this chapter, we present our paper "Some findings on zero-inflated and hurdle Poisson models for disease mapping" by Francisca Corpas-Burgos (Foundation for the Promotion of Health and Biomedical Research of Valencia Region), Gonzalo García-Donato (University of Castilla-La Mancha) and Miguel A. Martinez-Beneito (Foundation for the Promotion of Health and Biomedical Research of Valencia Region) published in *Statistics in Medicine* (2018), 37(23):3325-3337.

Abstract

Zero excess in the study of geographically referenced mortality data sets has been the focus of considerable attention in the literature, with zero-inflation being the most common procedure to handle this lack of fit. Although hurdle models have also been used in disease mapping studies, their use is more rare. We show in this paper that models using particular treatments of zero excesses are often required for achieving appropriate fits in regular mortality studies since, otherwise, geographical units with low expected counts are oversmoothed. However,

as also shown, an indiscriminate treatment of zero excess may be unnecessary and has a problematic implementation. In this regard, we find that naive zero-inflation and hurdle models, without an explicit modeling of the probabilities of zeroes do not fix zero excesses problems well enough and are clearly unsatisfactory. Results sharply suggest the need for an explicit modeling of the probabilities that should vary across areal units. Unfortunately, these more flexible modeling strategies can easily lead to improper posterior distributions as we prove in several theoretical results. Those procedures have been repeatedly used in the disease mapping literature and one should bear these issues in mind in order to propose valid models. We finally propose several valid modeling alternatives according to the results mentioned that are suitable for fitting zero excesses. We show that those proposals fix zero excesses problems and correct the mentioned oversmoothing of risks in low populated units depicting geographic patterns more suited to the data.

Keywords

Disease mapping, hurdle Poisson model, posterior impropriety, zero excess, ZIP

4.1. Introduction

Zero excesses have been frequently addressed within the disease mapping literature, see for example Ugarte et al. (2004); Song et al. (2011); Nieto-Barajas and Bandyopadhyay (2013); Musenge et al. (2013); Arab (2015). We consider this problem from a Bayesian perspective, a paradigm frequently adopted in this context (the last four references above are Bayesian). This topic has received considerable attention in recent years. For example, popular Bayesian software such as `INLA` (Rue et al., 2009) has included up to 5 different functions that implement

specific models to handle situations of zero excesses. This issue is not exclusive to disease mapping problems but, on the contrary, is related to any type of data taking in general positive integer values (including 0), such as for example Poisson, binomial or negative-binomial distributed data. Zero excesses are a source of overdispersion caused by a disagreement between the data and the distribution assumed: we have more zeroes in our data set than the proposed distribution could reasonably explain. As a consequence, zero excesses are features inherent to particular combinations of distributions (or models in general) and data sets, but not intrinsic to particular data sets. The presence of a large number of zeroes is symptomatic of a zero excess situation, but not necessarily indicative of one since observing many zeroes could be perfectly compatible with a Poisson distribution with a low expected value. Therefore an indiscriminate use of models dealing with zero excesses is, in principle, not necessary. In this sense several procedures have been developed for assessing zero excesses in specific problems like Van Der Broek (1995) or Bayarri et al. (2008) which deal with this issue on Poisson data with constant or covariate-dependent expected cases.

Many disease mapping studies have incorporated zero excesses modeling strategies in the analysis of mortality spatial data. Nevertheless, to our knowledge, it has not been extensively tested whether zero-specific treatments should be routinely used in this context or if, on the contrary, the standard Poisson assumption (with spatially varying random effects) fits regular mortality data well enough. Moreover, it is rarely the case that the pursued positive effect of such treatments is checked with the unexpected possible consequence that the original data misfit, in terms of zero counts, still remains. A motivating aspect of this research is to shed some light on these two relevant questions using a real extensive setting with 540 areal units and 46 geographical patterns corresponding to roughly 27 different causes of mortality. In particular we consider the zero-inflated and hurdle Poisson models, the most popular models in the related literature.

With respect to the first question, in roughly 15 patterns out of the 46 considered (barely 32% of the cases) we have observed a serious departure from the number of zeroes predicted with traditional disease mapping models, while the need for specific zero excess treatment for the rest is questionable. Our findings for the second question are more worrisome from a practical point of view. As we report, a preventive extra zero modeling may be totally innocuous for the zero-inflated approach without a particular modeling of the zero-specific component. For hurdle models the situation is even worse, since the estimations of the underlying risks can be dramatically influenced by spurious circumstances like the spatial distribution of the population along the region of study. The consequence is relevant since, for many cases, we could be reporting nonsense estimations based on an unneeded zero excess treatment.

The results observed in the real application indicate that for regular zero-inflated and hurdle Poisson proposals a specific modeling of the probability of zero-excess is needed in order to construct satisfactory methods. This is admittedly the path followed by many applied works in the literature (references will be given). Nevertheless, as we prove, such modeling has an unforeseen important difficulty, namely that conditions for impropriety of the posterior distribution (an invalidating fact for many not so formal related approaches) are very soft. These theoretical results make the assignment of the prior distributions a very delicate issue, preventing the use of highly popular "casual" non-informative priors frequently implemented by-default in specialized Bayesian software. Our result is quite general and affects several components of the model (like fixed effects or variances of the random effects) and many of the link functions (e.g. logit or probit). Additionally, we propose alternative modeling strategies that, as we argue, are safer in terms of validity of the results.

This paper is divided into 6 sections. Section 4.2 introduces the BYM model (Besag et al., 1991), the most popular proposal for disease

mapping and two specific refinements, zero-inflation and hurdle Poisson modeling, in order to cope with zero excesses. Section 4.3 shows the performance of these proposals in the analysis of the Valencian Mortality Dataset. Section 4.4 contains the main theoretical results about conditions for impropriety of posterior and presents some valid proposals to overcome the problems encountered. Section 4.5 illustrates the dangers of using vague prior distributions on some particular variables of models treating zero excesses and reassesses the behavior of the proposals made in Section 4.4 on the previous Valencian Mortality Dataset. Finally, Section 4.6 draws some conclusions from the results derived in this paper.

4.2. Some proposals for treating zero excesses in disease mapping

The goal of disease mapping is dealing with the sparse information in the observed counts of some health outcome over a set of areal units. In general these units are small in statistical terms, with frequent low observed counts, that makes them noisy and weakly informative of the underlying risk of the disease for many of them. Thus, statistical modeling is needed for drawing acceptable risk estimates in those units. The models used for this task mainly rely on spatial conditional autoregressive random effects to induce geographical dependence on the risk estimates and therefore to increase the amount of information used to estimate them. Among the models using these random effects we highlight one that is particularly popular, the Besag, York and Mollié's model (Besag et al., 1991), BYM henceforth. For this model, data $\{O_i : i = 1, ..., I\}$ representing observed counts on the areal units are modeled as

$$O_i|R_i \sim Poisson(E_i R_i),\ i = 1, ..., I,$$

where E_i are the expected counts for each unit, typically calculated by means of some age standardization, and R_i are the corresponding risks

that we would like to estimate. Regarding the modeling of this last term, BYM defines the log-risks as:

$$log(R_i) = \mu + \phi_i + \theta_i, \tag{4.1}$$

where μ stands for an intercept modeling the mean of the log-risks and the two subsequent terms are Gaussian random effects. The term $\boldsymbol{\phi}$ follows an intrinsic conditional autoregressive (ICAR) distribution, i.e. their components are assumed to have the following prior conditional distributions:

$$\phi_i|\boldsymbol{\phi}_{-i}, \sigma_\phi \sim N\left(n_i^{-1}\sum_{j\sim i}\phi_j, n_i^{-1}\sigma_\phi^2\right),\ i = 1, ..., I,$$

where n_i stands for the number of neighboring areas of unit i, the subindex in $\boldsymbol{\phi}_{-i}$ indicates all terms in $\boldsymbol{\phi}$ excepting its i-th component and the subindex $j \sim i$ denotes all those units j which are neighbors of i. This definition can be further elaborated introducing some parameters in order to weight the contribution of some units with respect to others, although we will not use that option. This term induces spatial dependence on $\boldsymbol{R}$ and accounts for those factors of regional scope which take effect on several contiguous units, making them similar. In contrast, the term $\boldsymbol{\theta}$ in Expression (4.1) accounts for risk factors of very limited geographical scope that take an effect just on isolated areal units and make their risks different to those of their surrounding units. The terms introducing independent variability on the risks are modeled as independent Gaussian random effects, i.e

$$\theta_i|\sigma_\theta \sim N(0, \sigma_\theta^2),\ i = 1, ..., I.$$

The amount of spatial dependence in $\boldsymbol{R}$ depends on the balance between σ_ϕ and σ_θ. If the first has higher (respectively lower) values, in comparison to the second, the final pattern will show substantial spatial dependence (independence).

Besides the spatial modeling that could be done with the BYM model the data available may require a specific treatment of the observed zero counts if the model fitted could not explain the amount of observed zeroes in the data set. The most used tool for dealing with zero excesses is zero-inflation. Specifically, in case of modeling observed counts with a Poisson likelihood, the resulting model is known as zero-inflated Poisson (ZIP) (Lambert, 1992). In its simplest form, ZIP models assume the observed counts to follow a mixture of a degenerate distribution with all its mass at zero and a $Poisson(\lambda)$ distribution, with weights $1 - \pi^Z$ and π^Z, respectively. This inflates the amount of zeroes expected by the Poisson distribution as a function of π^Z.

ZIP models for disease mapping fuse the simplest ZIP approach just introduced with spatial models (such as BYM). This yields flexible ZIP models with different (and dependent) λ_is, acknowledging that the studied data set may have more zeroes than those reproduced by BYM. Being more precise, a ZIP version of the BYM model could be formulated as follows: The observed data are assumed to follow a Poisson distribution of mean $E_i R_i Z_i$, where E_i stands for the expected cases, R_i for the spatially-varying risks in the Poisson distribution of the BYM model and Z_i for a binary variable modeling if the observed counts correspond to an extra-Poisson zero ($Z_i = 0$) or correspond to a value coming from the Poisson distribution ($Z_i = 1$). The risks R_i would be modeled as in Equation (4.1) and the Z_is would follow a $Bernoulli(\pi^Z)$ distribution, with unknown π^Z. For this model the smoothed Standardized Mortality Ratios (SMR) would be computed as $R_i Z_i$, i.e. a mixture of the BYM-based risks and 0. Examples of applications that adopt this modeling approach include Gschlößl and Czado (2008); Song et al. (2011); Musenge et al. (2013).

As an alternative to ZIP, data sets showing zero excesses are sometimes modeled as hurdle Poisson models (Mullahy, 1986), simply hurdle models henceforth. This proposal assumes the data to follow a mixture of a degenerate distribution with all its mass at zero and

a zero-truncated Poisson distribution. That is, in contrast to ZIP models, all observed zeroes in hurdle models are assumed to come from the zero-degenerate distribution. Thus, the parameter $1-\pi^H$ in hurdle models represents the probability that a given areal unit has zero observed cases instead of the percentage of extra-Poisson zeroes, the interpretation of $1-\pi^Z$ in ZIP. As for ZIP, hurdle models are combined with specific disease mapping proposals, such as BYM, in order to yield flexible spatial models accounting for zero excesses.

More specifically, for a hurdle version of the BYM model the observed counts O_i are assumed

$$P(O_i|\pi^H,\boldsymbol{\mu}) = (1-\pi^H)^{I_{\{0\}}(O_i)}\left(\pi^H\left(exp(-\mu_i)\frac{\mu_i^{O_i}}{O_i!}(1-exp(-\mu_i))^{-1}\right)\right)^{I_{(0,\infty)}(O_i)},$$

where $\mu_i = E_i R_i$ and $I_\Omega(x)$ is the indicator function for the condition $x \in \Omega$. The risks R_is in this model would follow Expression (4.1). For this proposal the smoothed SMR for the i-th unit should be computed as $\pi^H(\mu_i/(1-exp(-\mu_i)))/E_i$ (Neelon et al., 2013), where π^H is the probability of belonging to the truncated Poisson component and $\mu_i/(1-exp(-\mu_i))$ is the expected value given that the observation belongs to that component. This term is divided by the expected cases E_i since $\pi^H(\mu_i/(1-exp(-\mu_i)))$ would be the mean of O_i but we want to draw an estimate of O_i/E_i instead.

Both ZIP and hurdle versions of the BYM model, as introduced above, are posed under a Bayesian approach since BYM is also originally formulated from a Bayesian point of view. As a consequence all the parameters in BYM, ZIP and hurdle in this paper will have their own prior distribution. We will discuss prior distributions for these models more in depth in Section 4.4. Nevertheless, for now, we will not pay them further attention as they will be mostly irrelevant for the issues discussed in the next section. Anyway, the prior distributions used

in our analyses could be considered as regular prior choices for these models in the literature. Full details on the priors used can be found in the supplementary material of this paper (Annex A, Section A.3), which contains all the code used for its analyses.

In the next section, we implement these three different approaches in a real extensive setting in order to assess their practical utility. As we will see, the results are far from being as satisfactory as expected.

4.3. An initial analysis of the Valencian Mortality Data: A motivating application

Now that we have introduced the BYM model and two potential tools to cope with zero excesses, we are going to test their performance in an extensive real setting. We will pay particular attention to their fit in terms of the number of predicted zeroes in comparison to those actually observed. Our particular data set for this task is the mortality data used in the Spatio-temporal Mortality Atlas of the Valencian Region (1987-2006) (Zurriaga et al., 2010) in which we have ignored the temporal component. This atlas studies 46 geographical patterns corresponding to the distribution of 27 causes of mortality for each sex, excepting some particular combinations without enough deaths or without biological sense (e.g. prostate cancer in women). Mortality is disaggregated at the municipal level in a total of 540 municipalities of very different sizes, ranging from 22 to about 750,000 inhabitants (year 2000). Thus, observed deaths are expected to show substantial variability between municipalities, with some locations showing systematically 0 deaths for most of the causes.

The number of observed zeroes for the 46 geographical patterns analyzed ranges from 4 to 243. As we mentioned, such numbers,

although sometimes high do not necessarily mean zero excesses. They can simply represent low mortality for any of those causes or low population for some municipalities. Thus, for assessing zero excesses with regard to the models introduced in Section 4.2 we have run each of them on the available data. For each model and cause of death we have sampled values from the posterior predictive distribution of the observed deaths for each municipality and we have compared those samples against the observed values. Specifically, we have compared the number of zeroes observed for each cause of mortality and those predicted by the models from the MCMC.

Table 4.1 shows the results obtained for some causes of deaths, specifically the first 10 causes. The full table with all 46 analyses made is annexed as supplementary material to this paper (Annex A, Section A.2). The second column of Table 4.1 contains the number of zeroes observed for each data set meanwhile the next 3 columns correspond to that same number as predicted by each model run. Namely, we have run the BYM model without any particular treatment of zeroes as well as ZIP and hurdle versions of that same model. Bold fonts in Table 4.1 denote those combinations of models and data sets evidencing zero excesses according to their predictive intervals.

All models in this paper were run in `WinBUGS` and the code for each of them can be found as annex material at Annex A, Section A.3. A R-markdown document with all the analysis carried out can be found in that Annex. Three chains were run for each model and data set with 50,000 iterations, whose first 5,000 iterations were used as burn-in period. Of these, one of every 135 iterations was saved yielding a final sample size of 1,002 iterations. Convergence was assessed by means of the Brooks-Gelman-Rubin statistic (we required this to be lower than 1.1 for each variable in the model) and the effective sample size (required to be at least 100 for every variable in the model) implemented with the `R2WinBUGS` package of `R`.

Table 4.1.: Observed zeroes for each data set and posterior predicted zeroes for each model and for the first 10 mortality causes. Values in the *Obs. zeroes* column correspond to the real observed zeroes for each data set. For the next 3 columns, numbers correspond to the posterior predictive medians for this same quantity for each model run and the corresponding unilateral 95% posterior predictive intervals. Bold fonts denote those combinations of models and data sets evidencing zero excesses according to their predictive intervals.

Sex & Cause	Obs. zeroes	BYM	ZIP	Hurdle
(Men, All tumours)	4	2 [0,5]	3 [0,5]	5 [0,11]
(Women, All tumours)	7	6 [0,10]	6 [0,10]	8 [0,15]
(Men, Mouth)	216	**196 [0,211]**	**199 [0,215]**	216 [0,242]
(Men, Stomach)	105	**91 [0,103]**	**92 [0,104]**	105 [0,127]
(Women, Stomach)	150	137 [0,151]	138 [0,152]	150 [0,173]
(Men, Colorectal)	73	**58 [0,68]**	**59 [0,69]**	74 [0,93]
(Women, Colorectal)	74	72 [0,82]	73 [0,83]	74 [0,93]
(Men, Colon)	96	**79 [0,91]**	84 [0,96]	96 [0,119]
(Women, Colon)	98	91 [0,102]	92 [0,104]	99 [0,119]
(Men, Rectum)	201	**180 [0,196]**	**183 [0,199]**	202 [0,228]
...	...	...	...	...

Table 4.1 (and in more detail the full table in Annex A, Section A.2) shows how BYM may fit quite poorly the number of zeroes for certain data sets. Namely, for 15 out of the 46 data sets considered the 95% posterior predictive intervals for the number of zeroes in BYM did not contain the real observed zero counts and for 5 additional data sets the upper limit of that interval coincided with the observed zeroes –this seems excessive since we would expect a priori just 2 or 3 of the observed zeroes to lay outside of the predictive intervals–. The main conclusion is a substantial lack of fit for BYM in terms of the number of zeroes predicted and therefore a general advice for specific treatment

of those cases. On the contrary, BYM seems to accommodate well the number of zeroes in the rest of datasets (26), making it questionable the need for particular treatments of excess of zeroes in those settings.

With respect to the approaches with a particular treatment of zeroes, the results are not satisfactory for different reasons. Surprisingly, ZIP does not help much in fitting more zeroes and 11 out of the 46 original data sets showed 95% posterior predictive interval which do not contain the real observed number of zeroes and in 1 occasion the upper limit of the interval coincided with those zeroes. This performance, although better than that of BYM is also unacceptable since the number of predictive intervals that do not contain the corresponding observed value is far above of that corresponding to the nominal probability of the interval. On the contrary, for hurdle, all intervals contained the observed number of zeroes. Nevertheless, this better fit of the proportion of zeroes has a pernicious effect on the estimations of the SMRs that make them barely reliable. To understand this effect, we have represented in Figure 4.1 choropleth maps for the SMRs fitted for all three models in Table 4.1 for rectum cancer in males, one of the cases where the presence of a zero excess for BYM and ZIP is evident.

We first highlight that the maps for BYM and ZIP are quite similar for this data set and in general for all diseases fitted (maps not shown). This is not surprising according to the fit of the π^Z parameter in ZIP for all the causes. The posterior mean for this parameter, which measures the weight of the Poisson side of ZIP models, for all 46 data sets ranges from 0.973 to 0.998. Thus even though ZIP models should be able to fit zero excesses, they refuse to do it by minimising the weight of the zero-specific component. This may be a consequence of the implementation where the π^Z parameter is common to all municipalities. So, decreasing π^Z for making room to more zeroes in smaller municipalities also entails an increase in the probability of observing zeroes in large cities where that probability is virtually zero. Since the amount of information available in large municipalities is

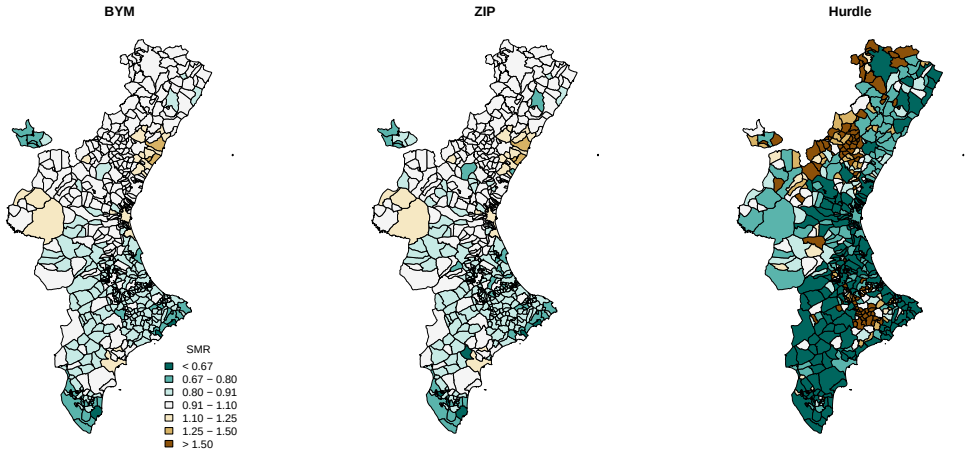


Figure 4.1.: Choropleth maps for the SMR estimates for all three models in Section 4.3, rectum cancer in men.

much higher than that in the smaller ones, ZIP decides to reject the zero-specific term as its contribution is more harmful, in likelihood terms, for the large municipalities than beneficial for the smaller ones (those with potential lack of zeroes).

The SMR map for hurdle shows a weird pattern completely different to BYM. This map shows a polarized pattern with high SMRs in the smaller municipalities and low SMRs for the rest. This pattern is systematically repeated for most of the data sets analyzed (maps not shown), being more evident for those data sets with more observed zeroes. In our opinion, this is also an effect of having a common π^H parameter for all the municipalities. In contrast to ZIP now π^H for the different data sets is not so close to 1, being its posterior mean always very close to the proportion of non-zero observed counts for each data set. Nevertheless, as mentioned in the previous section, π^H takes also

an effect on the calculation of the SMRs for this model, decreasing the mean of the Poisson component in that same proportion. For the small cities this makes the number of predicted zeroes to be increased but, alternatively, for the non-small cities this makes the SMRs to be underestimated as evidenced in Figure 4.1.

These results suggest that in our case both zero-specific treatments using these naive proposals which put the same zero-specific probabilities to all units do not seem a good choice. At least in our extensive analysis, ZIP does not seem to have a clear effect with regard to the baseline BYM model. For hurdle the particular (naive) treatment of zeroes makes misleading the corresponding SMRs map. Since considering a common probability for the zero-specific side seems to be the cause of these problems, we will explore from now on the opportunities and benefits that the modeling of those probabilities could bring.

4.4. Modeling of the probability of observing a zero

One of the most valuable advantages of Bayesian hierarchical models is the possibility of modeling particular features of the data that we could be interested in. Nevertheless, that ability is not always good as it can lead us to models which are not necessarily well formulated and therefore to misleading or plainly wrong results.

As introduced in the previous section, both ZIP and hurdle models require a particular treatment of the assignment of the observed counts to any of the two processes intervening in each of them. That assignment follows a binary process which, up to now, has depended on a single parameter π common to all areal units. We will denote π when we refer indistinctly to either π^Z or π^H. The obvious alternative to a common probability is modeling unit-specific π_is by means of, for example, logistic regression. This has been repeatedly done in the

disease mapping context for both ZIP (Dalrymple et al., 2003; Gschlößl and Czado, 2008; Neelon et al., 2010; Musenge et al., 2013; Nieto-Barajas and Bandyopadhyay, 2013) and hurdle models (Dalrymple et al., 2003; Neelon et al., 2010, 2013; Upfill-Brown et al., 2014; Neelon et al., 2014; Arab, 2015). That is, following several of the proposals in the literature, for both ZIP and hurdle models we will consider from now on

$$logit(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta} + \varphi_i \tag{4.2}$$

where $\boldsymbol{\beta}$ model the effect of some set of covariates $\mathbf{X}$ and $\boldsymbol{\varphi}$ is a vector of (possibly spatial) Gaussian random effects modeling the effect of those factors that cannot be explained by $\mathbf{X}$.

In the next subsection we introduce a series of results of great interest for the models that we want to explore now. Namely, we have found important posterior impropriety problems in hurdle and ZIP models when the vector of probabilities $\boldsymbol{\pi}$ is modelled with either fixed or random effects. This makes that modeling quite tricky and caution has to be taken in order to avoid flawed modeling proposals. These results will determine some ZIP and hurdle specific proposals that should be avoided in general. We will discourage the use of those models particularly in a non-informative or objective setting. Additionally, Subsection 4.4.1 will allow us to focus on some valid proposals with different π_is, that will be later developed at Subsection 4.4.2.

4.4.1. Some theoretical results warning against the use of certain popular casual non-informative priors

Once a suitable model is specified, when it comes the need to assign the prior distribution, the applied literature is flooded with casual possibilities that include, for example, a uniform prior on fixed effects parameters or its 'proper' counterpart of a normal density with an arbitrarily large variance. Obviously, these proposals are valid (in the

sense that results are covered by laws of probability) as far as the associated posterior distribution is proper (see below the comment on the 'vague' counterparts), a property that it is rarely checked in practice. We do so here and conclude that the conditions for propriety of the posterior are quite severe and are not fulfilled by many popular non-informative choices.

We start by introducing some results for hurdle models which consider $\boldsymbol{\pi}^H$ as proposed in (4.2). The proofs and full formulation of the results introduced in this subsection are provided in the supplementary material to this paper (Annex A, Section A.1).

First, we have shown that the hurdle model with $\boldsymbol{\pi}^H$ modeled as in (4.2) is problematic since some issues arise on the use of both fixed and random effects in that expression. As stated in Corollary 1 (Annex A, Section A.1) the use of random effects with improper prior distributions for σ, the standard deviation of the random effects, yields an improper posterior distribution regardless of the other elements in the model. This means that the use of random effects in (4.2) with many of the default prior choices in the literature for their variability should be avoided. Besides, if for the j^* column of $\boldsymbol{X}$, x_{ij^*} is positive for every i with $O_i > 0$ and negative otherwise (or vice versa) and the prior distribution of β_{j^*} is improper for large positive (respectively negative) values then the posterior distribution is also improper. So, we could also have posterior impropriety problems using fixed effects for modeling $\boldsymbol{\pi}^H$. Fortunately, this condition (although just a sufficient condition, not necessary, for impropriety) will not be fulfilled easily since it depends in a binary manner on all the (random) values of the outcome of the model. That binary condition should be fulfilled for all the observed outcomes which is not that easy, especially for regions with a large number of units. Additionally, Corollary 1 is very general since as stated there, these results above would hold equally for other common link functions in (4.2) such as probit or tobit; they would also hold for non-Poisson based likelihoods such as binomial or negative binomial and for other

different spatial structures (with positive-definite covariance matrices) besides BYM.

The situation for ZIP models is not better. As stated in Corollary 2 (Annex A, Section A.1) the use of random effects in (4.2) for ZIP models is as problematic as for hurdle models, since it yields an improper posterior distribution under the same premises. Moreover, the results for ZIP are equally general since they also apply for different link functions, likelihood families and spatial structures for the mean of the non-zero process. Nevertheless, the case of fixed effects is substantially different (worse) for ZIP since these yield posterior impropriety more easily than hurdle models. Thus, we have found that a sufficient condition for posterior impropriety in ZIP would be that for any column j^* of $\boldsymbol{X}$, $x_{ij^*} > 0$ (respectively $x_{ij^*} < 0$) for all i and β_{j^*} to diverge for large positive (respectively negative) values. This condition is much more general since this could be fulfilled by design of the covariates, regardless of the observed counts $\boldsymbol{O}$. In principle we could easily get rid of this issue by, for example, subtracting the mean of any of the covariates in the model but the problem would remain for the intercept. The intercept is positive for all the units in the model so any improper prior distribution on its corresponding term in $\boldsymbol{\beta}$ would yield an improper posterior distribution, independently of the additional problems that the rest of covariates in the model could also entail.

One could be tempted to use vague proper prior distributions, instead of improper priors, as a possible strategy to avoid impropriety issues. This is a procedure frequently found in the literature, supposedly to avoid MCMC convergence problems. Nevertheless, according to the results stated above, these "convergence problems" are a numerical manifestation of the more worrisome fact of having an improper limiting posterior distribution. Berger (2006) argues that the use of a vague prior mimicking an improper prior with an associated improper posterior can only hide but not solve the problem. In our context, this of course invalidates the use of standard approaches like a vague normal priors

on each component of $\boldsymbol{\beta}$, vague gamma priors on the precision of the random effects or uniform prior distributions with a large upper limit on their standard deviations. Interestingly though, a tentative use of vague proper priors could serve as a diagnostic test to detect possible underlying problems of posterior impropriety. For instance, results assuming a uniform prior on the standard deviation of $\boldsymbol{\varphi}$ with an arbitrary large upper limit show high sensitivity to such upper limit, warning clearly about the possible impropriety of the posterior distribution.

4.4.2. Some valid proposals for modeling π

The previous subsection has stated some procedures to be avoided when modeling the probabilities $\boldsymbol{\pi}$ in both ZIP and hurdle models. One option would be to use informative prior distributions for $\boldsymbol{\beta}$ and σ. In this sense Agarwal et al. (2002) have made one proposal of informative prior distributions for $\boldsymbol{\beta}$ for ZIP models. Nevertheless, we would rather avoid informative prior distributions. So, we will propose some (non-informative) procedures for modeling $\boldsymbol{\pi}$ that do not fulfill the conditions for posterior impropriety stated above. Regretfully, we do not have a proof for the posterior propriety of these proposals since the impropriety conditions formulated are just sufficient but not necessary. In any case, these new proposals do not fall into the premises of those results, in contrast to many of the proposals formulated in the literature. Moreover, in our experience, these new proposals do not seem to show at all any of the MCMC convergence problems appearing when one of the models yielding improper posterior (according to the conditions stated in the previous subsection) were used. We formulate now 3 separate modeling proposals.

Fixed effects modeling:

Although, as described above, the use of random effects for modeling $\boldsymbol{\pi}$ is quite problematic, the use of fixed effects for modeling $\boldsymbol{\pi}^H$ in hurdle seems a much less troublesome option. Thus, a potentially valid modeling proposal (we will refer to this as *FE* [Fixed Effects] henceforth) would be to consider a hurdle model as defined in Section 4.2 with

$$logit(\boldsymbol{\pi}^H) = \boldsymbol{X}\boldsymbol{\beta}.$$

A suitable proposal that could be used in principle for any disease mapping model would be: $\boldsymbol{X} = [\mathbf{1}_I, log(\boldsymbol{E})]$, where $\boldsymbol{E}$ stands for the vector of expected values used in the Poisson likelihood of hurdle models. We have taken the logarithm of the expected values to avoid any potential effect of the usually skewed distribution of this variable caused by the presence of very few large cities. According to the results above this could yield an improper posterior distribution if O_i is positive for each region with $E_i > 1$ and $O_i = 0$ otherwise (or vice versa). But, for a reasonably high number of areal units this condition seems very unlikely to be fulfilled.

This proposal models the logit of the probabilities of non-zeroes as a function of the expected observations at each areal unit. This seems quite reasonable since units with lower expected counts would show more easily zero observed counts meanwhile those larger units will show positive counts in general. This could be achieved for β_2 (the coefficient corresponding to the log-expected cases) taking positive values. For this proposal we will consider an improper uniform prior distribution for each component of $\boldsymbol{\beta}$. This is because we specifically want to avoid the use of vague prior distributions that could hide posterior impropriety problems into just MCMC convergence problems due to the almost impropriety of posterior distributions.

Interestingly, note the link between this proposal and the EZIP1 proposal in Song et al. (2011). In that paper a ZIP model with $\pi_i^Z =$

$\frac{E_i}{\delta+E_i}$ is proposed. A logit transformation of this expression yields $logit(\pi_i^Z) = log(E_i) - log(\delta)$ which would be a ZIP version of the FE model just proposed. However, note that this model is valid since δ is assumed to have a $Unif(0,1)$ prior distribution in the EZIP1 model which yields an exponential distribution of mean 1 on $-log(\delta)$. This proper prior obviously avoids any potential impropriety on the posterior distribution.

Nested fixed effects modeling:

The use of expected values as a surrogate of the (population) size of the areal units in the FE modeling seems quite reasonable. Nevertheless, this does not depend at all on the probabilities of non-zeroes resulting from the Poisson side of hurdle models: $\boldsymbol{\pi}^P = 1 - exp(-\boldsymbol{\mu})$. Although these probabilities have been evidenced to produce some misfit in the data in terms of zero excesses, they could be also used as sensible covariates for modeling the probabilities of zeroes $\boldsymbol{\pi}^H$, instead of just $\boldsymbol{E}$. This approach was already introduced in the zero-altered model of Heilbron (1994). These probabilities $\boldsymbol{\pi}^P$ would not just take into account the size of the areal units, through the expected counts $\boldsymbol{E}$, but also the risk attributed to any of them by the Poisson side of the model. These risks could be an additional source of information making considerable improvements as compared to the use of simple expected counts. Thus, our second proposal for modeling $\boldsymbol{\pi}^H$ in hurdle models would be

$$logit(\pi_i^H) = logit(\pi_i^P) + \gamma.$$

This would be an alternative fixed effects logistic modelling of $\boldsymbol{\pi}^H$ using $logit(\boldsymbol{\pi}^P)$ as an offset. The values of that offset would be leveraged by γ so that if it takes values close to 0 this model would reproduce the probabilities in the Poisson layer, even for zero-counts, meanwhile for $\gamma < 0$ the zero-specific probabilities would be inflated in regards to the Poisson model. Note that in case of adapting this modeling to the hurdle-BYM model in Section 4.2, the original (uninflated) BYM model

could be reproduced within this proposal by making $\gamma = 0$, thus we will henceforth refer to this model as *NFE*, Nested Fixed Effects model. Once again we will consider an improper uniform prior distribution for γ so that any potential posterior impropriety problem in this model appears.

Geometric modeling:

Since resorting to logit (or probit, tobit) regression has proved to bring lots of problems into ZIP and hurdle models, we could try to avoid those transformations in order to make sensible proposals. Thus, making

$$\pi_i = 1 - (1 - \pi^G)^{E_i}$$

seems a reasonable proposal for both ZIP and hurdle models. For this proposal we would have that the probability of observing a zero count for a unit with n expected cases is $(1 - \pi^G)^n$, where $1 - \pi^G$ is that same probability for a unit with 1 expected case. This geometric progression also holds for the Poisson process where the probability of observing zeroes with n expected cases $exp(-n\lambda) = exp(-\lambda)^n$ follows that same relationship. Thus, the probabilities of zero counts for this proposal are in agreement with the Poisson side of the model. For π^G, which can be interpreted as the probability of observing a positive count for units with one expected case, we set a uniform prior distribution between 0 and 1. Since this prior is proper we avoid any posterior impropriety coming from this term. One of the main advantages of this model is that since the modeling of $\boldsymbol{\pi}$ does not rely on any improper prior distribution this model could be also set up for ZIP modeling. This is contrast to the previous proposals whose ZIP counterparts would be discouraged since they rely on fixed effects logit modeling. We will refer to the ZIP and hurdle versions of this model henceforth as *ZGeo* and *HGeo*, respectively.

4.5. Empirical illustration of the modeling proposals introduced

We start this section by illustrating the problems induced on ZIP and hurdle models by arbitrary prior vagueness. With this section we seek to make clear how prior problems are not just present for improper prior distributions but also for vague proper priors, which are commonly used in ZIP and hurdle disease mapping models. Finally, we will show how the modelling proposals introduced in the previous section perform with the same datasets used in Section 4.3 where naive ZIP and Hurdle models showed a deficient performance.

4.5.1. An illustration of the prior vagueness problems in ZIP and hurdle models

We are going to illustrate the dangers of using vague proper priors, instead of improper priors, for modeling $\boldsymbol{\pi}^Z$ and $\boldsymbol{\pi}^H$ in ZIP and hurdle models. We have already proved that using improper priors for some variables in these models would yield improper posteriors but we want to evidence that using vague proper prior does not seem to be a safe option in any case. Thus, we have run two separate models in this study: a ZIP model with $logit(\pi_i^Z) = \alpha$ for $i = 1, ..., I$ and a hurdle model with $logit(\pi_i^H) = \alpha + \gamma_i$ and $\gamma_i \sim N(0, \sigma_\gamma)$ for $i = 1, ..., I$. These models are somewhat naive, indeed, as mentioned in the paper the ZIP model proposed will not fit in general any risk excess, and additional regressors could be used for modelling both $\boldsymbol{\pi}^Z$ and $\boldsymbol{\pi}^H$ in order to improve them. Nevertheless, we have preferred to keep these models as simple as possible in order to illustrate the prior specification problems that they show. We have run these two models on the rectum cancer data set that has also illustrated the results in Section 4.3. All models were run in `WinBUGS`.

Regarding the ZIP model mentioned, we have run it for several different prior choices for α: $\alpha \sim N(0, \sigma_\alpha^2)$ for σ_α^2 equal to $10, 100, 1000$ and 10000. For the first of these choices α has a posterior mean of 5.26 and a 95% posterior credible interval of [3.22,8.68]. For $\sigma_\alpha^2 = 100$ we obtain a posterior mean of 10.85 and a credible interval of [3.98,24.33]. For $\sigma_\alpha^2 = 1000$ we obtain a posterior mean of 18.87 and a credible interval of [5.18,35.15]. Finally, for $\sigma_\alpha^2 = 10000$ `WinBUGS` finds a numerical error (`TRAP 66`) and is not able to run this model. As we see, posterior inference on α completely depends on the prior distribution set for this parameter. None of the models run, excepting that with $\sigma_\alpha^2 = 10000$, show any evident convergence problem. Thus, someone fitting these models without an additional sensitivity analysis, such as ours, will accept as good the results for any of the models run, when these models are just hiding the impropriety problems of an hypothetical improper prior choice for α. Note that as we increase σ_α^2, α increases steadily, giving zero probability to the zero-specific component. This reinforces the idea that naive ZIP proposals with logit modeling of $\boldsymbol{\pi}^Z$ do not fit appropriately zero excesses.

Regarding the random effects hurdle model, we have run it also with different prior distributions for σ_γ: $\sigma_\gamma \sim Unif(0, U_\gamma)$ for U_γ equal to $2, 10$ and 100. For U_γ equal to 2 the posterior mean of σ_γ is equal to 1.1 with 95% posterior credible interval [0.1,2.0]. For U_γ equal to 10 the posterior mean of σ_γ is equal to 6.4 with 95% posterior credible interval [0.7,9.8]. Finally, for U_γ equal to 100 the posterior mean of σ_γ is equal to 69.2 with 95% posterior credible interval [17.0,99.0]. Note how the upper limits of the posterior credible intervals for σ_γ are always very close to U_γ, pointing out the informativeness of these supposedly uninformative choices. Thus, in summary, we see how the posterior distribution of σ_γ heavily depends on the (arbitrary) vagueness of its prior distribution, which makes unadvisable the use of arbitrary vague proper priors for σ_γ as a safe substitute of an improper prior distribution.

4.5.2. A re-analysis of the Valencian Mortality Dataset

We turn back once again to the analysis of the Valencian Mortality Dataset in Section 4.3. We have run all 4 models proposed in the previous subsection: FE, NFE, HGeo and ZGeo, on the diseases considered there. First, we have assessed their fit in terms of the number of zero counts reproduced, i.e. the equivalent of Table 4.1 but for these new models. Table A.2 of the supplementary material to this paper (Annex A, Section A.2) shows, for all of them, the posterior medians and 95% credible intervals for the number of predicted zeroes. As a summary, in contrast to the results shown in Table 4.1, the posterior predictive distribution for the number of zeroes in all 4 models agree with those numbers observed for the real data sets. Namely, all 3 hurdle models yield similar results to the hurdle model in Table 4.1 with the posterior predictive median for the number of zeroes in the data sets always very close to the real observed zeroes. The modeling of the probabilities of zeroes in ZGeo has made a great improvement over naive ZIP models since for ZGeo the predictive posterior median for the number of zeroes is always very close to the real observed zeroes. All 95% credible intervals for the number of predicted zeroes for all diseases and models contain the real observed zeroes as would be expected in models which are performing an explicit modeling of that particular feature in the data.

Second, we have also compared the fit of these models in general terms by using the Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002). The DICs for all models and 46 data sets, with their corresponding deviances and number of effective parameters, can be found at Table A.3 of Annex A, Section A.2. Regarding the FE model its DIC is higher than that of the BYM model for 43 out of 46 data sets so its performance in general does not seem very satisfactory. Although the FE model is more complex than BYM (has two additional parameters) the deviances obtained are in general substantially higher

than those of BYM models. This suggests that the modeling proposed in FE is worse than that of the BYM model, thus maybe a linear function of $log(\boldsymbol{E})$ is not as good as it could seem in principle. As a consequence we will not pay further attention to this model from now on. The NFE model attains better DICs for 11 out of the 15 data sets identified as having zero excesses. Meanwhile, for just 5 out of the remaining 31 data sets with no evidence of zero excess NFE was better in terms of DIC, as could be expected since BYM is less complex than NFE and for these data sets NFE should not yield any improvement. Thus, NFE attains in general lower DICs in those settings where it would be expected. Regarding HGeo, it attained 6 out of 15 DICs lower than BYM for those data sets needing a particular treatment for zeroes and 2 out of 31 times was lower for those data sets that did not need that treatment in principle. Finally, ZGeo also obtained similar results to HGeo, improving BYM in 5 out 15 times where zero excesses were evidenced and 8 out of 31 times when these were not so evident. Thus the results of Geometric models are overall satisfactory although not as good, in terms of DIC, as those of NFE.

Regarding the estimates of the parameters in the models proposed, those of NFE showed a particularly coherent performance. Thus, for all data sets needing zero treatment the parameter γ in the model attained a 95% posterior credible interval completely below 0 (we mentioned that $\gamma < 0$ should be a sign of zero correction with respect to BYM). On the contrary, for only 1 out of the 31 data sets not showing zero excesses the 95% credible interval for γ was completely below zero. Posterior means and 95% credible intervals for γ for all 46 data sets can be found at Table A.4 in Annex A to the paper (Section A.2). We do not find anything particularly interesting in the π^G estimates obtained in the Geometric models. These parameters have a cumbersome interpretation since they are referred as the probability of the zero-specific term for units having $E = 1$, but each data set and spatial unit have different expected values. Thus, no particularly intuitive result is drawn from their estimates.

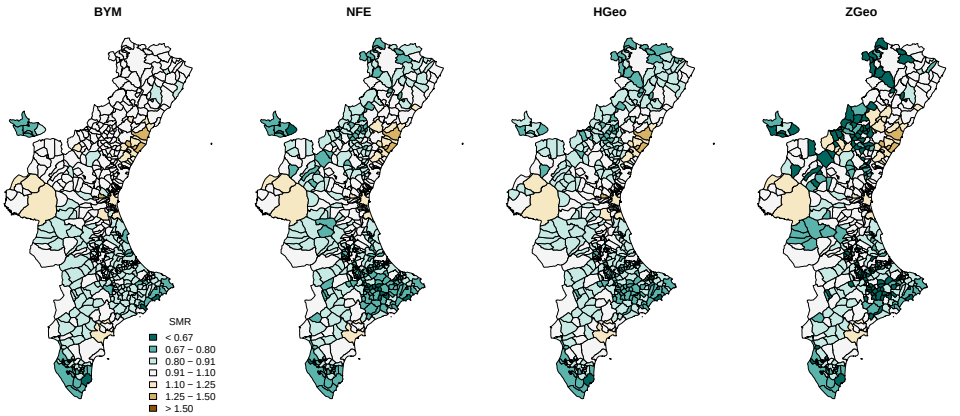


Figure 4.2.: Choropleth maps for the SMR estimates of NFE, HGeo and ZGeo models, rectum cancer in men.

Figure 4.2 shows choropleth maps with the smoothed SMRs for rectal cancer in men for BYM, NFE, HGeo and ZGeo. Recall that this pattern was one of those needing some zero treatment. Both hurdle maps (NFE and HGeo) are similar as their modeling of the probabilities of zeroes is also similar, as mentioned in Section 4.4. They mainly modify the risks in those regions less populated and more prone to zeroes (upper-left side of the maps) decreasing their risks in order to get those extra zeroes needed. As seen in Section 4.3 this differential performance of the low populated areas could not be achieved with the naive models introduced in Section 4.2. In contrast, regions having high SMRs hardly show any change. Thus hurdle models mainly modify the left tail of the distribution of the SMRs in order to fit the zero excess, but leaves the right tail of the distribution mostly unchanged. ZGeo introduces more differences with regard to BYM in both tails

of the distribution. New regions with both high and low risks have emerged in this map. Several regions of very low risk have emerged in the upper-left side of the map. This result of ZGeo is very common and can also be seen in many of the diseases studied (see Annex A, Section A.2 for seeing all 4 maps for the whole set of diseases).

Although the choice of a particular model for treating zero excesses is not a goal of this work, we would recommend to use NFE as benchmark proposal between all those introduced in this paper. We have found particularly satisfactory that NFE shows a better performance in terms of DIC than the rest of models and the estimates of its γ parameter seems very coherent. Moreover, this model seems to yield conservative results in that the change in their geographic patterns compared to BYM is milder than that for the rest of models, yet enough to correct the original zero excesses in the BYM models. Finally, the logit formulation of NFE makes it particularly well suited for further modelling $\boldsymbol{\pi}^H$ if needed in contrast to the geometric proposals. Thus we overall find NFE a convenient proposal for modeling data sets showing zero excesses.

4.6. Conclusions

Disease mapping models with zero-specific treatment can be considered as enhanced disease mapping models controlling overdispersion in the observed counts inducing also dependence on the underlying risks. Nevertheless, overdispersion fitting procedures in general may not be enough for solving zero excesses problems, which are a unique kind of overdispersion. Thus, specific models are needed to deal with this problem. As shown in this paper zero excesses are present in certain data sets concerning mortality data, at least for the Valencian Mortality Dataset. A relevant proportion of the diseases studied have been found to show zero excesses, even after accounting for overdispersion with disease mapping models. Thus, as evidenced, zero excesses require attention for mortality geographic studies in general.

The Valencian Mortality Dataset is somewhat particular in some senses due to the high demographic variability of the units of study. This could have made naive ZIP and hurdle models (without modeling of the probabilities of the zero-specific component) seem particularly bad as zero-specific components also put substantial probability to zero counts in large cities. As a consequence the zero-specific components are discarded. Nevertheless, we expect substantial differences in the expected cases for regular disease mapping studies since otherwise those expected values would not be omnipresent in so many studies. In any case the Valencian Mortality Dataset is comprehensive enough and representative of real mortality data so that the need of zero-treatment evidenced in this dataset could be a signal of a general fact in mortality data sets of other regions.

Maybe one reason why mortality data may show frequent zero excesses when smoothing the SMRs is inherent to the smoothing process. Smoothing procedures usually combine information on the observed data and the prior structure defined by the model. When that observed information is low (small units) the shrinkage towards the prior structure is stronger. As a consequence the risks in the smaller units may be easily oversmoothed towards the mean, or a local mean, yielding conservative risk estimates. Models treating zero excesses with a different probability of the zero-specific component solve this problem by decreasing the risk in the smaller units (those which are more likely to show zeroes) and therefore increasing the number of zeroes predicted. Nevertheless, a similar oversmoothing could exist in small units showing high risks. In that case their SMRs should be higher but they are oversmoothed towards the mean because of the small information in each of them. Proposals modeling zero excesses in no way would fix this issue which only alleviates the oversmoothing of small units showing low risks.

In our opinion the theoretical results in Subsection 4.4.1 are also of high importance from an applied point of view. They show that proposals leading to wrong (improper) results have been

frequently proposed in the literature. These problems can have different consequences such as plain improper posterior distributions or, if arbitrary vague prior distributions are used, arbitrary posterior distributions which are extremely sensitive (possibly unnoticedly) to prior parameters. These problems are often interpreted in the literature as simple MCMC convergence problems. Although this may seem obvious, we would advice modelers to pay further attention to those convergence problems. In our experience those problems have been an excellent guidance for formulating the theoretical results in Subsection 4.4.1 since they clearly warn that something suspicious could be happening. In our opinion this is an additional advantage of MCMC inference since convergence problems can be treated, at least in this context, as a trace of problems in model formulations instead of simple drawbacks inherent to MCMC as an inferential tool.

The main purpose of this paper has not been to propose a particularly suitable model for dealing with zero excesses. Besides showing the high prevalence of zero excesses problems in regular mortality data, which would deserve further epidemiological research, the purpose of this paper is double. On one hand, we pretend to show some theoretical pointing out wrong procedures in this area. In our opinion this is quite important in order to avoid works proposing flawed models. The main value of this side of the paper is warning modelers on what procedures not to do instead of setting what to do with zero excesses. Our results are just sufficient, not necessary, conditions for posterior impropriety in these models. On the other hand, this paper illustrates several "valid" proposals for modeling zero excesses, i.e. we wanted to illustrate suitable proposals for handling this particular issue that were admissible in light of the results shown in Section 4.4. It would be desirable to have a proof of the posterior propriety of these proposals, or even better necessary conditions for the posterior impropriety of zero-specific models in general. Regretfully we do not have that proof but anyway the value of the results proved still remain since they guide us on what procedures not to follow which is a

valuable guidance according to many models already proposed in the literature.

A more thorough comparison of the models in Subsection 4.4 and possibly some further models would be greatly advisable although that comparison is beyond the scope of this paper. We have found more interesting to illustrate several modeling proposals instead of exposing just one of them. We also find it convenient to conclude pointing out that, although the conditions in our results are fairly general, there are some settings that are not covered by them. For example, our results do not shed light on the use of multivariate random effects for modeling the zeroes and Poisson processes by means of multivariate spatial distributions (Neelon et al., 2013, 2014). Moreover, our results apply only to the case of having a single observed count per spatial unit, if more counts were available (Neelon et al., 2010, 2013, 2014) this would be beyond the scope of this paper. Anyway, the current results suggest the need for further research on these settings but also suggest a high dose of caution when formulating proposals in this area.

Finally, we would like to point out that according to Natarajan and McCulloch (1995) the conditions stated there for posterior impropriety in the modeling of binary data are similar to those formulated in Albert and Anderson (1984) for non-existence of MLE in logit frequentist modeling. Indeed, the conditions for posterior impropriety in the Bayesian approach are more restrictive than those for non-existence of the MLEs in the frequentist context. The conditions set at Natarajan and McCulloch (1995) have been those also set as conditions for posterior impropriety for the modeling of the probability of the zero-specific component with random effects in our work. Thus, the frequentist formulation of ZIP and hurdle models from a frequentist setting could be in principle as problematic as that same formulation from a Bayesian point of view. The conditions under which frequentist ZIP and hurdle models yield valid (or unvalid) MLEs should be further explored but the results of this paper and the work of Albert and Anderson (1984)

shed some doubts on those formulations from a frequentist point of view.

Acknowledgments

The authors acknowledge the support of the research grants MTM2016-77501-P from the Spanish Ministry of Economy and Competitiveness, BA15/00003 of Instituto de Salud Carlos III and predoctoral contract UGP-15-156 of FISABIO.

5. On the convenience of heteroscedasticity in highly multivariate disease mapping

In this chapter, we present our paper "On the convenience of heteroscedasticity in highly multivariate disease mapping" by Francisca Corpas-Burgos (Foundation for the Promotion of Health and Biomedical Research of Valencia Region), Paloma Botella-Rocamora (Conselleria de Sanitat Universal i Salut Pública) and Miguel A. Martinez-Beneito (Foundation for the Promotion of Health and Biomedical Research of Valencia Region) published in *Test* (2019), 28(4):1229-1250.

Abstract

Highly multivariate disease mapping has recently been proposed as an enhancement of traditional multivariate studies, making it possible to perform the joint analysis of a large number of diseases. This line of research has an important potential since it integrates the information of many diseases into a single model yielding richer and more accurate risk maps. In this paper we show how some of the proposals already put forward in this area display some particular problems when applied to small regions of study. Specifically, the

homoscedasticity of these proposals may produce evident misfits and distorted risk maps. In this paper we propose two new models to deal with the variance-adaptivity problem in multivariate disease mapping studies and give some theoretical insights on their interpretation.

Keywords

Gaussian Markov random fields, Multivariate disease mapping, Bayesian statistics, Spatial statistics, Mortality studies

5.1. Introduction

The analysis of geographical variations in rates of diseases has a long tradition in epidemiology and statistics. This area of research, known as disease mapping, has generated substantial interest from a methodological point of view. In the beginning, disease mapping studies focused mainly on the modeling of a single disease. However, there may be several diseases with common shared risk factors. Recently, multivariate disease mapping has received considerable attention by researchers interested in the simultaneous joint spatial modeling of several diseases (MacNab, 2016b,a; Martinez-Beneito et al., 2017). Multivariate disease mapping models attempt to estimate the risk of a disease in specific locations by using its spatial dependence as well as the geographical distribution of the risks for other related diseases. By so doing, a greater amount of information is used in the estimation of the risks than in univariate models, which allows more precise estimates to be obtained.

Martinez-Beneito (2013) recently developed a general framework for multivariate disease mapping capable of reproducing many of the Bayesian models in this area previously proposed in the literature. The problem with that approach, and that of most of the multivariate

disease mapping models in the literature, is its complexity and computing requirements. Consequently, most of the existing literature is restricted to multivariate modeling of two or three diseases at most. However, Botella-Rocamora et al. (2015) extended the previous work by developing a simpler and computationally more convenient proposal capable of handling a considerably large number (tens) of diseases. A second important advantage of this proposal is that it can be implemented in regular Bayesian simulation packages such as `WinBUGS` (Lunn et al., 2000).

In this work, we present an application of the methodology proposed in Botella-Rocamora et al. (2015) for the spatial modeling of several diseases in the cities of Alicante, Castellón and Valencia, which belong to the Valencian region, one of the 17 administrative regions that Spain is divided into. After observing the results obtained, some limitations of the previous methodology are evidenced when it is applied to smaller cities, as is the case of Castellón. For this reason, we propose an enhancement, variance adaptivity, of Botella-Rocamora et al.'s methodology, which allows the problems evidenced to be solved and thereby improving multivariate risk estimates. Moreover, we also focus particular attention on the multivariate implementation of the Besag et al. (1991) model (BYM henceforth), which has not previously been developed within the context of $\boldsymbol{M}$-models, the multivariate disease mapping proposal used in this paper.

This paper is organized as follows. Section 5.2 describes the modeling proposal in Botella-Rocamora et al. (2015) for multivariate disease mapping and introduces the particular implementation of that model with BYM spatial structures. Section 5.3 shows an application of the previous methodology to real data in the Spanish cities of Alicante, Castellón and Valencia. In Section 5.4 we propose a modification of the previous model that makes it possible to solve some of the problems found in the estimation of the risk maps for the city of Castellón. Section 5.5 presents and compares the results obtained with the new

modeling proposal in a simulated study and on the mortality data for Alicante, Castellón and Valencia, previously analyzed. Finally, Section 6.5 contains some conclusions about the models and the results obtained in the previous sections.

5.2. The M-model for multivariate disease mapping

A general statistical framework for the multivariate disease mapping problem can be described as follows. Let O_{ij} and E_{ij} denote, respectively, the number of observed and expected cases for the i-th geographical unit of study and the j-th disease. The data likelihood assumes that

$$O_{ij} \sim Poisson(E_{ij}RR_{ij}) \quad i = 1, ..., I, \quad j = 1, ..., J$$

where RR_{ij} is the relative risk for the i-th geographical unit and j-th disease, and is modeled as $log(RR_{ij}) = \mu_j + \theta_{ij}$. The term μ_j is just an intercept for the j-th disease and $\boldsymbol{\Theta} = \{\theta_{ij} : i = 1, ..., I; j = 1, ..., J\}$ is a collection of random effects whose joint distribution specifies how dependence is defined within and between diseases. Specifically, dependence among the columns of $\boldsymbol{\Theta}$ induces dependence between diseases and, similarly, dependence among its rows induces spatial dependence within diseases (geographical units).

The original modeling proposal in Botella-Rocamora et al. (2015) induces spatial multivariate dependence by setting

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\mathbf{M} \tag{5.1}$$

where $\boldsymbol{\Phi}$ is an $I \times K$ matrix of random effects with independently distributed columns that typically follow some spatially correlated distribution, such as a proper CAR (Conditional Auto-Regressive)

distribution. In that case different spatial correlation parameters could be set for the columns of $\mathbf{\Phi}$, to reproduce a non-separable covariance structure (Martinez-Beneito, 2013). Additionally, several columns with different dependence structures could also be used to reproduce more complex spatial dependence structures, such as Besag, York and Mollié's model (Besag et al., 1991). Those spatial distributions induce dependence between spatial units and therefore between rows of $\mathbf{\Theta}$. Additionally, $\mathbf{M}$ is a $K \times J$ random matrix which induces dependence between the different columns in $\mathbf{\Theta}$, that is, between the different diseases considered in the analysis. Usually $K = J$, although they could be different, such as for the multivariate formulation of the BYM model, where two random effects are included per disease and therefore $K = 2J$. The variance parameter of the random effects in the columns of $\mathbf{\Phi}$ is usually set to 1, since $\mathbf{M}$ cells are responsible for controlling the variability of $\mathbf{\Theta}$. Otherwise, those variances and the cells of $\mathbf{M}$ would not be identifiable as they would cancel each other out. On the other hand, as proposed by Botella-Rocamora et al., the cells of $\mathbf{M}$ are independently defined as $M_{ij} \sim N(0, \sigma^2)$ $i = 1, ..., K,\ j = 1, ..., J$, where σ could be either a fixed (typically large) value, and therefore the M_{ij}s would follow vague independent prior distributions, or an additional variable to be estimated in the model. In the first case, we call the corresponding modeling *fixed effects $\boldsymbol{M}$-modeling*, since $\mathbf{M}$ cells would be modeled in that manner and, alternatively, we call the second case *random effects $\boldsymbol{M}$-modeling*, once again because of the modeling of the cells carried out in $\mathbf{M}$. Botella-Rocamora et al. chose a uniform vague prior distribution on σ, which is the one we will also use throughout this paper.

To conclude this brief introduction to Botella-Rocamora et al.'s proposal, we believe it is also worth mentioning a theoretical property of this model that will be used later in this work. Thus, as shown in the original paper, assigning $N(0, \sigma^2)$ prior distributions to the entries in $\mathbf{M}$ yields a $Wishart(K, \sigma^2\mathbf{I}_j)$ prior distribution for the covariance matrix between diseases $\mathbf{\Sigma}_b$ when all spatial models share the same

spatial distribution, which can be computed as simply $\mathbf{M}'\mathbf{M}$. Hence, the independent modeling of the cells of $\mathbf{M}$ entails a prior mean for $\boldsymbol{\Sigma}_b$ proportional to an identity matrix or, alternatively, it assumes prior independence in the columns of $\boldsymbol{\Theta}$.

5.2.1. BYM M-models

The multivariate generalization of the BYM model by means of $\boldsymbol{M}$-models has not been described in the literature to date. Since BYM models are one of the most popular modeling options in univariate disease mapping, we will develop that extension in detail in this section. The BYM models spatial patterns as the sum of two random effects, one spatially correlated ($\boldsymbol{\psi}$) following an Intrinsic CAR (ICAR) distribution and the other independent between the units of study ($\boldsymbol{\gamma}$). The first random effect induces dependence between nearby spatial units, while the second allows them to have markedly different risks to those of their neighbors, if appropriate. To formulate an $\boldsymbol{M}$-model depending on BYM spatial structures, we should consider $\boldsymbol{\Phi} = [\boldsymbol{\Psi} : \boldsymbol{\Gamma}]$, where $\boldsymbol{\Psi} = [\boldsymbol{\psi}_1 : ... : \boldsymbol{\psi}_K]$ and $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1 : ... : \boldsymbol{\gamma}_K]$ are $I \times K$ matrices of spatial (ICAR) and heterogeneous terms. In the following we will consider K as equal to J so that we will have as many spatial and heterogeneous terms as spatial patterns to be modeled. Therefore, $\boldsymbol{\Phi}$ will be an $I \times (2J)$ matrix. In a similar manner we will consider $\boldsymbol{M} = [\boldsymbol{M}'_s : \boldsymbol{M}'_h]'$, where $\boldsymbol{M}_s$ and $\boldsymbol{M}_h$ are $J \times J$ matrices in charge of modeling the covariance between diseases for the spatial and heterogeneous terms.

For these definitions of $\boldsymbol{\Phi}$ and $\boldsymbol{M}$, the original matrix product in Expression (5.1) is

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{M} = \boldsymbol{\Psi}\boldsymbol{M}_s + \boldsymbol{\Gamma}\boldsymbol{M}_h.$$

If for any $I \times J$ matrix $\boldsymbol{X}$ we denote $vec(\boldsymbol{X}) = (\boldsymbol{x}'_{\cdot 1}, ..., \boldsymbol{x}'_{\cdot J})'$, then

$$vec(\boldsymbol{\Theta}) = vec(\boldsymbol{\Psi}\boldsymbol{M}_s) + vec(\boldsymbol{\Gamma}\boldsymbol{M}_h)$$

and therefore the covariance matrix of this vector can be decomposed as

$$\boldsymbol{\Sigma}_{vec(\boldsymbol{\Theta})} = \boldsymbol{\Sigma}_{vec(\boldsymbol{\Psi}\boldsymbol{M}_s)} + \boldsymbol{\Sigma}_{vec(\boldsymbol{\Gamma}\boldsymbol{M}_h)}.$$

Since $vec(\boldsymbol{\Psi}\boldsymbol{M}_s) = (\boldsymbol{M}'_s \otimes \boldsymbol{I}_I)vec(\boldsymbol{\Psi})$ (see, for example, Expression (3.76) in Gentle (2007)), then

$$\begin{aligned}\boldsymbol{\Sigma}_{vec(\boldsymbol{\Psi}\boldsymbol{M}_s)} &= (\boldsymbol{M}'_s \otimes \boldsymbol{I}_I)\boldsymbol{\Sigma}_{vec(\boldsymbol{\Psi})}(\boldsymbol{M}_s \otimes \boldsymbol{I}_I) \\ &= (\boldsymbol{M}'_s \otimes \boldsymbol{I}_I)(\boldsymbol{I}_J \otimes \boldsymbol{\Sigma}_s)(\boldsymbol{M}_s \otimes \boldsymbol{I}_I) = (\boldsymbol{M}'_s\boldsymbol{M}_s \otimes \boldsymbol{\Sigma}_s),\end{aligned}$$

where $\boldsymbol{\Sigma}_s = (\boldsymbol{D} - \boldsymbol{W})^-$ denotes the Moore-Penrose generalized inverse of the precision matrix of the ICAR distribution of the columns of $\boldsymbol{\Psi}$. In a similar manner

$$\boldsymbol{\Sigma}_{vec(\boldsymbol{\Gamma}\boldsymbol{M}_h)} = (\boldsymbol{M}'_h\boldsymbol{M}_h \otimes \boldsymbol{I}_I).$$

Therefore, for the BYM $\boldsymbol{M}$-model defined above

$$\boldsymbol{\Sigma}_{vec(\boldsymbol{\Theta})} = (\boldsymbol{M}'_s\boldsymbol{M}_s \otimes \boldsymbol{\Sigma}_s) + (\boldsymbol{M}'_h\boldsymbol{M}_h \otimes \boldsymbol{I}_I), \tag{5.2}$$

where all the blocks in that matrix are of the form $(\sigma^2_s)_{ij}(\boldsymbol{D} - \boldsymbol{W})^- + (\sigma^2_h)_{ij}\boldsymbol{I}_I$, so that all the variances and cross-covariances in this model have BYM spatial structures of different spatial and heterogeneous variances.

One particularity of the implementation of $\boldsymbol{M}$-models with BYM spatial structures is the following. Let us assume all the cells in $\boldsymbol{M}$ follow an $N(0, \sigma^2)$ as for the original implementation of $\boldsymbol{M}$-models with proper CAR distributions. In that case, the covariance matrix of the corresponding model will take the form of Expression (5.2), where $\boldsymbol{\Sigma}^s_b = \boldsymbol{M}'_s\boldsymbol{M}_s$ and $\boldsymbol{\Sigma}^h_b = \boldsymbol{M}'_h\boldsymbol{M}_h$ follow $Wishart(J, \sigma^2\boldsymbol{I}_J)$ distributions. As a consequence, both $\boldsymbol{\Sigma}^s_b$ and $\boldsymbol{\Sigma}^h_b$ have the same prior distribution and if data are not strong enough, the two matrices will tend to be similar. This may be a problem as it could interfere in the balance between the spatial and heterogeneous terms of the BYM model that

controls the amount of spatiality and heterogeneity in the geographical patterns that are fitted. By considering all the cells in $\boldsymbol{M}$ as having the same prior distribution, that balance would be set in advance, regardless of the spatial dependence that data could show. A possible solution to this problem could be to set different variances for the elements of $\boldsymbol{M}_s$ and $\boldsymbol{M}_h$, that is $(\boldsymbol{M}_s)_{ij} \sim N(0, \sigma_s^2)$ and $(\boldsymbol{M}_h)_{ij} \sim N(0, \sigma_h^2)$. In this case the covariance matrix of the BYM $\boldsymbol{M}$-model would also take the form of Expression (5.2), but in this case $\boldsymbol{\Sigma}_b^s \sim Wishart(J, \sigma_s^2 \boldsymbol{I}_J)$ and $\boldsymbol{\Sigma}_b^h \sim Wishart(J, \sigma_h^2 \boldsymbol{I}_J)$, which will allow the balance between the spatial and heterogeneous terms to be determined within the model. This is the implementation of the BYM $\boldsymbol{M}$-model that we have run for the next example.

A final issue that deserves some attention when implementing BYM $\boldsymbol{M}$-models is the estimation of the covariance matrix between diseases. For multivariate models with a separable structure, the covariance matrix can be calculated as $\boldsymbol{M}'\boldsymbol{M}$ (Botella-Rocamora et al., 2015). Nevertheless, separability would require the columns of $\boldsymbol{\Phi}$ to share a common distribution with common parameters. This is a restrictive assumption, but even for pCAR $\boldsymbol{M}$-models with different but similar correlation parameters, $\boldsymbol{M}'\boldsymbol{M}$ could be a reasonable estimate of the covariance matrix between diseases. For BYM $\boldsymbol{M}$-models, however, one half of the underlying patterns in $\boldsymbol{\Phi}$ have ICAR priors while the other half have independent Normal priors, which are very different. Moreover, the (marginal) scale of these two prior distributions may be very different (Bernardinelli et al., 1995; Schrödle and Held, 2011), which is something that should be borne in mind in order to estimate any sensible covariance matrix between diseases. Thus, for BYM $\boldsymbol{M}$-models, we find it far more sensible to summarize the covariance matrix between diseases as the covariance matrix of the columns of $\log(\boldsymbol{\Theta})$ instead of making $\boldsymbol{M}'\boldsymbol{M}$, as for $\boldsymbol{M}$-models based on a single distribution for the columns of $\boldsymbol{\Phi}$.

5.3. A motivating analysis

5.3.1. A multivariate mortality study in Castellón

The multivariate proposal put forward by Botella-Rocamora et al. (2015) has been implemented to study the geographical distribution of mortality in the cities of Alicante, Castellón and Valencia. In this section, we present some results obtained in the city of Castellón, which was composed of 95 census tracts (the geographical unit for this analysis) and had around 170,000 inhabitants in 2016. Parallel results for the analyses performed in Alicante and Valencia, composed of 215 and 553 census tracts respectively, are included as supplementary material to this paper due to lack of space (see Annex B, Section B.1). We consider the multivariate joint spatial modeling of 20 different causes of mortality and both *fixed* and *random effects* $\boldsymbol{M}$*-models* for all three cities separately. In order to evaluate the benefits of multivariate modeling, we compare the results obtained with Botella-Rocamora et al.'s $\boldsymbol{M}$-models with underlying BYM spatial patterns against those obtained with independent BYM patterns for each disease.

All models were run in `WinBUGS` and the `R` code for calling each of them can be found as annex material in supplementary material (Annex B, Section B.2). Three chains were run for each model with 30,000 iterations, the first 5,000 of which were discarded as burn-in period. Of these, one out of every 75 iterations was saved, thereby yielding a final sample size of 1,002 iterations. We reran the model for Castellón with 300,000 iterations but we did not find any differences in the results of the two runs. We have therefore preferred to keep the results with just 30,000 iterations for the rest of the paper since we felt that these were enough. Convergence was assessed by means of the Brooks-Gelman-Rubin statistic (we required this to be lower than 1.1 for each variable) and the effective sample size (required to be at least 100 for each variable). Convergence was explored for the intercepts of the diseases, risk estimates, cells of the covariance matrices between diseases ($\boldsymbol{\Sigma}_b^s$ and $\boldsymbol{\Sigma}_b^h$) and deviance, i.e., for the variables in the model

which are identifiable. We have not assessed their convergence of the rest of the parameters in the model that are not identifiable (cells of $\boldsymbol{M}$ and $\boldsymbol{\Phi}$), since they could yield a false sensation of convergence by simply exchanging their values for each step of the MCMC. Convergence was assessed with the `R2WinBUGS` package of `R`.

Figure 5.1 shows the results obtained with univariate BYM models (upper row), fixed effects $\boldsymbol{M}$-modeling (middle row), and random effects $\boldsymbol{M}$-modeling (lower row) for 3 out of the 20 causes of death under study in Castellón: AIDS, Cerebrovascular disease, and Suicides in men. Results shown for the fixed effects $\boldsymbol{M}$-model assume improper $M_{ij} \propto 1$ distributions, that is, we implicitly assume σ to be set to ∞ in this case. Nevertheless, we have also run the same model with σ set to high fixed values, such as 100 or 1000, obtaining results that are barely distinguishable. Green colors correspond to census tracts with estimated low risks (Smoothed Standardized Mortality Ratio (sSMRs)$< 0.67 = (1.5)^{-1}$ for darker greens), while brown colors correspond to units of high risk (sSMRs$>$ 1.5 for darker browns). Light-colored units denote milder deviations from the overall risk for the city.

As can be appreciated, markedly different risk maps are obtained with the multivariate fixed effects $\boldsymbol{M}$-model, as compared to the univariate BYM models. Although the risk maps for AIDS for both models do not present such marked differences (a map with great variability is obtained for both models), in the case of Cerebrovascular disease and Suicides, quite distinct risk maps are obtained. While univariate modeling generally provides maps with low variability, fixed effects $\boldsymbol{M}$-modeling provides maps with great variability, with hardly any smoothing, which resemble the corresponding maps of unsmoothed SMRs (not shown in the paper). This performance of the fixed effects $\boldsymbol{M}$-modeling in Castellón has also been observed for most of the diseases in the study. Interestingly, this lack of smoothing is noticed, but to a much lesser extent, in the results drawn from Alicante and Valencia (see

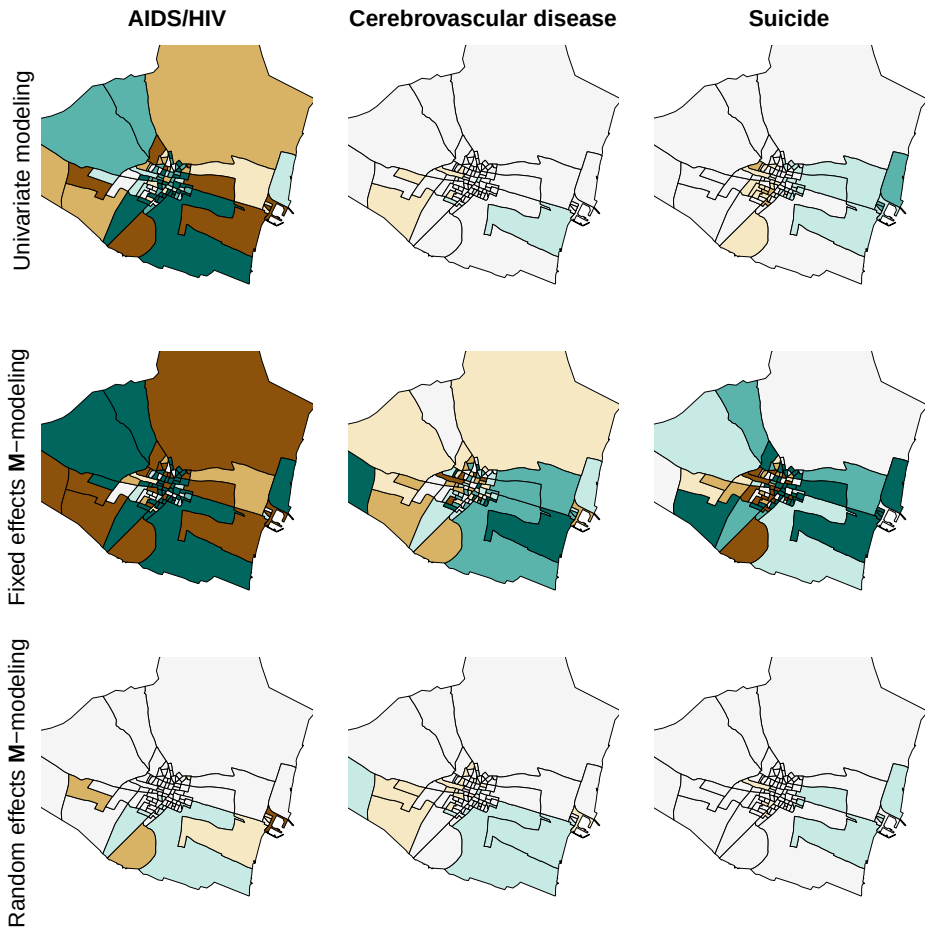


Figure 5.1.: Choropleth maps for the estimated risk patterns using traditional univariate modeling (BYM), above, fixed effects $\boldsymbol{M}$-modeling, center row, and random effects $\boldsymbol{M}$-modeling, below.

the previously mentioned supplementary material Annex B, Section B.1) where the results for the fixed and random effects $\boldsymbol{M}$-models are very similar, in accordance with the original paper by Botella-Rocamora et al. (2015).

The lower row in Figure 5.1 shows the results for these same three diseases for the random effects $\boldsymbol{M}$-model. As can be seen, in

this case, there are no major differences between the risk maps of Cerebrovascular disease and Suicides for the independent BYM models and the random effects $\boldsymbol{M}$-model. Both models show low variability and similar geographical patterns. However, the two risk maps obtained for AIDS mortality are dramatically different. The univariate model points out several census tracts with extreme risk in some specific locations in the city of Castellón that are known to be quite deprived. In contrast, a much flatter map (more similar in this sense to those of Cerebrovascular disease and Suicides) is obtained with the random effects $\boldsymbol{M}$-modeling. On this map, no census tract shows high risk, as is also the case for Cerebrovascular disease and Suicides. In general, we observed that random effects $\boldsymbol{M}$-modeling estimates in Castellón yield flat risk maps for all 20 diseases studied, which in a few cases, such as AIDS, are very different from those estimated with univariate modeling. Strikingly, this performance was only seen in Castellón, but not (or not so evident) in Alicante or Valencia.

In the next subsection we will attempt to explain why these strange results are obtained for the fixed and random effects $\boldsymbol{M}$-models in the city of Castellón.

5.3.2. A statistical interpretation of the results in the motivating analysis

First, we find it interesting to emphasize that far milder differences were found between the fixed and random effects risk patterns in Alicante and Valencia. For these two cities, both multivariate models take advantage of the additional information provided by the set of diseases considered, depicting more detailed spatial patterns in general than their univariate alternatives. This suggests that the results found for Castellón could be due to the smaller size of this city, where the prior structure that the $\boldsymbol{M}$-model induces could be more influential than in Alicante and Valencia. Thus, the prior covariance structure of the $\boldsymbol{M}$-model could

be having an undesirable effect on the final fit that, when available data are weaker, might be influencing the spatial patterns determined.

Regarding the fixed effects $\boldsymbol{M}$-model, we have mentioned that it was equivalent to assuming a $Wishart(K, \sigma^2 \mathbf{I}_J)$ prior distribution on the covariance matrix between diseases $\boldsymbol{\Sigma}_b$. Similarly, for a BYM $\boldsymbol{M}$-model, the prior $Wishart(K, \sigma^2 \mathbf{I}_J)$ distribution would be assigned to $\boldsymbol{\Sigma}_b^s$ and $\boldsymbol{\Sigma}_b^h$. Since σ is usually set to a large value for the fixed effects approach, this entails that the prior mean of $\boldsymbol{\Sigma}_b$ is equal to $K\sigma^2 I_J$, for a high value of σ. Therefore, this model assumes the prior covariances between diseases to be centered at 0 and the prior variances of the log-risks for each spatial pattern to be high. These prior assumptions could explain the results found in Castellón for the fixed effects model, where the prior information in $\mathbf{M}$ could overwhelm the information provided by the data. For this city, the cells of $\boldsymbol{\Theta}$ do not produce any smoothing in the risks fitted, as a consequence of their large prior variances (subsumed in matrix $\mathbf{M}$), which does not produce any shrinkage. As a consequence, the smoothed SMRs estimated for this model reproduce the unsmoothed original SMRs that disease mapping models typically try to avoid.

The random effects $\boldsymbol{M}$-model also leads to a prior mean of $K\sigma^2 I_J$ for $\boldsymbol{\Sigma}_b$ but with σ now being a parameter to be estimated within the model. This would potentially avoid the undesirable non-shrinking effect of the fixed effects $\boldsymbol{M}$-model when applied to smaller data sets. In this case the prior mean will just be proportional to the identity matrix but the proportionality constant will be estimated by the model itself, which will be set to a common consensus value for all the diseases (two common values for BYM). Univariate BYM models for each of the diseases in Castellón yielded posterior standard deviations for the log-SMRs ranging from 0.05 to 0.42, depending on the disease. AIDS was the disease with a higher standard deviation, far larger than the median standard deviation for the set of diseases considered (0.13). Thus, the distribution of the standard deviations of the log-SMRs for the different diseases has a pronounced asymmetrical right-tailed distribution. In consequence,

the consensus scale parameter σ for the random effects model takes a value that is much lower than that required to appropriately describe the spatial variability of AIDS mortality. This could explain perfectly why the initial pattern highlighted by the univariate BYM model for AIDS vanishes when the random effects $\boldsymbol{M}$-model is fitted.

In sum, the Castellón multivariate mortality study above has shown important prior sensitivity for the $\boldsymbol{M}$-model, mainly for smaller data sets. Specifically, the fixed effects $\boldsymbol{M}$-model has a tendency to yield unsmoothed risk estimates. Furthermore, the random effects version has an inclination toward the shrinkage of all diseases to a common point in terms of variability. Although this could be fine for some particular data sets, in general it will be a restrictive performance of this model which would be advisable to improve by seeking more adaptive models, at least in terms of the variance of the diseases. This is the goal that we pursue from now on.

5.4. An heteroscedastic modification of the M-model

Our proposal for fixing the prior sensitivity problems of the $\boldsymbol{M}$-model consists in a modification of its random effects version. Specifically, we relax the assumption of a common scale parameter for the cells of $\mathbf{M}$. In particular we propose two different ways to do this. The first proposal considers $M_{ij} \sim N(0, \sigma_i^2)$ for $i = 1, ..., K$, while our second alternative proposal considers $M_{ij} \sim N(0, \sigma_j^2)$ for $j = 1, ..., J$. From now on we will refer to these two proposals as the *row variance-adaptive* random effect $\boldsymbol{M}$-model (or simply RVA $\boldsymbol{M}$-model) and the *column variance-adaptive* random effect $\boldsymbol{M}$-model (or simply CVA $\boldsymbol{M}$-model), respectively. Obviously these two proposals will be more adaptive in terms of variability than the original random effects $\boldsymbol{M}$-model, which will, hopefully, allow us to solve the shrinkage problems toward a common variability evidenced in the previous section. Henceforth, we will refer

to the random effects $\boldsymbol{M}$-model introduced in Section 5.2 as simply the *non variance-adaptive* model (NVA model) in order to emphasize its main feature as compared to the two new variance-adaptive models that we have just introduced. Note that the concepts of both variance adaptivity and, alternatively, heteroscedasticity have already been used in univariate disease mapping studies (MacNab et al., 2006a; Congdon, 2008) to refer to adaptive models in terms of reproducing different variances for the different spatial units of study. We will use both these terms in this paper, as mentioned, to consider different variances for the different diseases in multivariate disease mapping problems.

Implementing the two $\boldsymbol{M}$-model modifications proposed for proper CAR versions of these models will be straightforward. Nevertheless, as seen in Section 5.2, the implementation of $\boldsymbol{M}$-models with BYM spatial structures requires some care. In particular, the RVA implementation for BYM spatial models is also straightforward, since for this model all the rows of $\mathbf{M}$ have Normal prior distributions of different variances. Consequently, the scale of the elements in $\boldsymbol{M}_s$ and $\boldsymbol{M}_h$ will be different, so there will be nothing that we need to be careful about. The balance of the spatial and heterogeneous terms in this model will be determined by the data instead of the prior structure of matrix $\mathbf{M}$. On the contrary, for the CVA $\boldsymbol{M}$-model, the scale of the cells of $\mathbf{M}$ vary only between columns, therefore fixing the balance between the spatial and heterogeneous terms in this model. Thus, the CVA $\boldsymbol{M}$-model will show similar problems to those in the original implementation of the NVA $\boldsymbol{M}$-model. These problems can now be solved in a similar manner as for the NVA $\boldsymbol{M}$-model. We will consider $\boldsymbol{M}_s$ and $\boldsymbol{M}_h$ to have different variances per column but these variances will also be different for those two matrices, that is, we assume $(M_s)_{ij} \sim N(0, (\sigma_s^2)_j)$ and $(M_h)_{ij} \sim N(0, (\sigma_h^2)_j)$. In this way the balance between the spatial and heterogeneous terms will vary freely for each pair of spatial and heterogeneous random effects.

For the rest of this section we will interpret the CVA and RVA

for $\boldsymbol{M}$-models in general. We will not assume underlying BYM spatial models for those models since the comments made are valid for $\boldsymbol{M}$-models in general. For BYM $\boldsymbol{M}$-models, all considerations made below on $\boldsymbol{\Sigma}_b$ may also be applied to $\boldsymbol{\Sigma}_b^s$ and $\boldsymbol{\Sigma}_b^h$.

5.4.1. An insight on the log-risks separation strategies for the RVA and CVA proposals

For all three RVA, CVA and NVA models, $\boldsymbol{M}$ can be stated as either $\boldsymbol{D}\boldsymbol{M}^*$ or $\boldsymbol{M}^*\boldsymbol{D}$ for $\boldsymbol{D} = diag(\boldsymbol{\Sigma})$ and $M_{ij}^* \sim N(0,1)$, for $\boldsymbol{\Sigma}$ a vector of the appropriate length. Specifically, $\boldsymbol{M} = \boldsymbol{D}\boldsymbol{M}^*$ for the RVA model, $\boldsymbol{M} = \boldsymbol{M}^*\boldsymbol{D}$ for the CVA model, and $\boldsymbol{M}$ can be stated as either $\boldsymbol{M}^*\boldsymbol{D}$ or $\boldsymbol{D}\boldsymbol{M}^*$ for $\boldsymbol{D} = \sigma\boldsymbol{I}_J$ or simply $\boldsymbol{M} = \sigma\boldsymbol{M}^*$ for the NVA model. This allows us to formulate the RVA model as

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{M}^*, \tag{5.3}$$

or the CVA model as

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{M}^*\boldsymbol{D} \tag{5.4}$$

in terms of Expression (5.1). In a similar manner, $\boldsymbol{\Theta}$ in the original NVA model could also now be expressed as

$$\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{M}^*\sigma, \tag{5.5}$$

or as both (5.3) and (5.4) for $\boldsymbol{D} = \sigma\boldsymbol{I}_J$, instead of a general diagonal matrix as for RVA or CVA. We will use these expressions to further study the theoretical properties of these proposals instead of the RVA and CVA formulations in the first paragraph of this section. Although the formulation in that paragraph is more convenient in computational terms (indeed it has been the one used to implement these models in `WinBUGS`), the matrix formulations above are more convenient for studying the statistical properties of the corresponding models. So we will use them extensively from now on.

Expression (5.3) evidences an interesting interpretation of the RVA model, that is, the matrix decomposition there can also be viewed as $(\boldsymbol{\Phi D})\boldsymbol{M}^*$ and thus the standard deviations $\boldsymbol{\Sigma}$ in that model may be interpreted as those corresponding to the underlying spatial patterns in $\boldsymbol{\Phi}$. Hence, this model can be viewed as a set of underlying spatial patterns of different variability (in contrast to NVA) that are later made dependent by their postmultiplication by $\boldsymbol{M}^*$. On the other hand, the CVA model first makes the spatial patterns in $\boldsymbol{\Phi}$ dependent (which originally had the same variability) and later those unscaled dependent patterns are scaled by means of the postmultiplication by $\boldsymbol{D}$. Therefore, the standard deviations $\boldsymbol{\Sigma}$ in both the RVA and CVA models have very different interpretations. First, for the RVA model, these standard deviations correspond to the underlying spatial patterns, whereas for the CVA model they scale the (spatial and multivariate dependent $\boldsymbol{\Phi M}^*$) patterns available according to the variability needed for each particular disease.

Expressions (5.3) to (5.5) separate the different sources involved in the multivariate covariance structure into different terms. Similar separation strategies are also advocated by Barnard et al. (2000) in multivariate (non-spatial) problems and by MacNab (2018) in multivariate disease mapping studies. Our proposal runs in that same direction, with some advantages that we will describe below.

By this separation of $\boldsymbol{\Theta}$ into several components, $\boldsymbol{\Phi}$ is in charge of modeling the spatial dependence of the data, $\boldsymbol{M}^*$ is in charge of modeling the multivariate dependence between diseases, and $\boldsymbol{D}$ models the scale of $\boldsymbol{\Theta}$. In any case, note that some confounding will remain between $\boldsymbol{M}^*$ and $\boldsymbol{D}$ since, ideally, $\boldsymbol{M}^*$ would be in charge of modeling the correlation matrix between diseases, but it does not do exactly that. To model the correlation matrix between diseases $\boldsymbol{C}_b$, $\boldsymbol{M}^*$ should be defined so that $\boldsymbol{C}_b = (\boldsymbol{M}^*)'\boldsymbol{M}^*$. This would entail J column restrictions on $\boldsymbol{M}^*$, specifically $\{\boldsymbol{M}^*_{\cdot j}\boldsymbol{M}^*_{\cdot j} = 1 : \ j = 1, ..., J\}$, which are generally detrimental for MCMC algorithms (in our experience

neither `WinBUGS` nor `Stan` tolerate restrictions of this kind very well). In contrast, we propose modeling $M^*_{ij} \sim N(0,1)$. With this choice we have that $\boldsymbol{M}^*_{\cdot j}\boldsymbol{M}^*_{\cdot j} \sim \chi^2_J$, which will give J as expected value. Thus, our definition of $\boldsymbol{M}^*$ does not allow us to set $\boldsymbol{M}^*_{\cdot j}\boldsymbol{M}^*_{\cdot j}$ to some specific value and therefore it will not model any correlation matrix. Nevertheless, the feature $E(\boldsymbol{M}^*_{j\cdot}\boldsymbol{M}^*_{\cdot j}) = J$, which makes $J^{-1}(\boldsymbol{M}^*)'\boldsymbol{M}^*$ on average a correlation matrix, fixes the scale of $\boldsymbol{M}^*\boldsymbol{M}^*$. This allows the modeling of the scale of the multivariate patterns to be separated into the separate matrix $\boldsymbol{D}$, since that scale cannot be controlled by $\boldsymbol{M}^*$.

Hence, we now have two alternative separation strategies that could fix the non-adaptability, in terms of variability, of the NVA proposal in Botella-Rocamora et al. (2015). We are now going to explore their differences for modeling $\boldsymbol{\Theta}$ through $\boldsymbol{\Sigma}_b$.

5.4.2. An insight on the RVA and CVA proposals in terms of the modeling of $\boldsymbol{\Sigma}_b$

The main difference between the RVA and CVA proposals lies in their inherently different ways of modeling $\boldsymbol{\Sigma}_b$. Thus, for RVA $\boldsymbol{\Sigma}_b = \boldsymbol{M}'\boldsymbol{M} = (\boldsymbol{D}\boldsymbol{M}^*)'(\boldsymbol{D}\boldsymbol{M}^*) = (\boldsymbol{M}^*)'\boldsymbol{D}^2\boldsymbol{M}^*$, whereas for CVA $\boldsymbol{\Sigma}_b = (\boldsymbol{M}^*\boldsymbol{D})'(\boldsymbol{M}^*\boldsymbol{D}) = \boldsymbol{D}(\boldsymbol{M}^*)'\boldsymbol{M}^*\boldsymbol{D}$. According to these decompositions of $\boldsymbol{\Sigma}_b$ and expressions (5.3) and (5.4), RVA and CVA have markedly different interpretations. We start by analyzing RVA. Note that $\boldsymbol{M}^*$ in Expression (5.3) could be QR-decomposed as $\boldsymbol{M}^* = \boldsymbol{Q}\boldsymbol{R}$ for suitable orthogonal ($\boldsymbol{Q}$) and upper triangular ($\boldsymbol{R}$) matrices. Therefore, Expression (5.3) could be alternatively stated as $\boldsymbol{\Theta} = \boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{Q}\boldsymbol{R}$. If $\boldsymbol{R} = \boldsymbol{I}_j$, then, for RVA, we would have $\boldsymbol{\Sigma}_b$ being equal to $(\boldsymbol{D}\boldsymbol{M}^*)'(\boldsymbol{D}\boldsymbol{M}^*) = (\boldsymbol{Q}\boldsymbol{R})'\boldsymbol{D}^2\boldsymbol{Q}\boldsymbol{R} = \boldsymbol{Q}'\boldsymbol{D}^2\boldsymbol{Q}$, that is, $\boldsymbol{Q}$ and $\boldsymbol{D}^2$ would contain the eigenvectors and eigenvalues, respectively, of $\boldsymbol{\Sigma}_b$. Hence, in this case, we could interpret the RVA model as a PCA decomposition of $\boldsymbol{\Theta}$, where $\boldsymbol{\Phi}$ would be the (spatially correlated) individual scores corresponding to each geographical unit, $\mathbf{D}$ would weight the contribution of each axis to the multivariate dependence

structure in $\boldsymbol{\Theta}$, and $\mathbf{Q}$ contains the orthogonal axis defining the PCA. In the most general case in which $\mathbf{R}$ was not necessarily equal to $\mathbf{I}_j$, the axis in the PCA would not be just $\mathbf{Q}$, but $\mathbf{QR} = \boldsymbol{M}^*$ and therefore, in that case, RVA can be understood as a PCA of $\boldsymbol{\Theta}$ followed by a subsequent non-orthogonal rotation. In this general case the PCA interpretation would therefore remain but with the original cloud of points projected onto a non-orthogonal axis given by $\boldsymbol{M}^*$. The columns of $\boldsymbol{\Phi}$ could be understood as the individual scores corresponding to each spatial unit when projected onto those non-orthogonal components. The RVA model assumes that each of those columns corresponding to a specific linear combination of diseases (hopefully with a particular sense) follows a spatially structured distribution so, in some sense, RVA performs a spatial PCA of the matrix of log-risks $\boldsymbol{\Theta}$ with non-orthogonal axes.

The non-orthogonal (spatially-correlated) PCA analysis performed in RVA could make more sense than it might seem at a first glance, since geographical patterns of risk factors would be rarely uncorrelated. Think, for example, of the spatial pattern of alcohol and tobacco consumption throughout a region of study. It would be hard to assume that both factors are independent. In that case, if these two risk factors were the two main determinants of the diseases in our study, a simple orthogonal PCA would induce spatially correlated distributions for both a linear combination of these factors (a weighted mean) and the corresponding orthogonal combination for these two variables. Assuming spatial distributions for these two components could not be justified since the second of them is mainly a residual shape component of the PCA, possibly showing weak spatial dependence. In contrast, a non-orthogonal PCA analysis, such as the one performed in the RVA model, would determine the same linear subspace for fitting those (correlated) effects, but without assuming an orthogonal performance between alcohol- and tobacco-related spatial distributions. Thus, these two axes could focus on the separate geographical description of alcohol and tobacco consumption, when the assumption of spatial dependence

for these two patterns is sure to be far more sensible than for the components of the regular orthogonal PCA.

Regarding CVA, Expression (5.4) could be alternatively stated as $\boldsymbol{\Theta}\mathbf{D}^{-1} = \boldsymbol{\Phi}\mathbf{M}^{*} = \boldsymbol{\Phi}\mathbf{Q}\mathbf{R}$ and, thus, CVA performs a matrix decomposition of the scale-standardized matrix $\boldsymbol{\Theta}\mathbf{D}^{-1}$. Since $\mathbf{D}^{-1}\boldsymbol{\Sigma}_b\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{D}(\boldsymbol{M}^{*})'\boldsymbol{M}^{*}\mathbf{D}\mathbf{D}^{-1} = (\boldsymbol{M}^{*})'\boldsymbol{M}^{*} = \mathbf{R}'\mathbf{Q}'\mathbf{Q}\mathbf{R} = \mathbf{R}\mathbf{R}'$, then $\mathbf{R}$ will correspond to the Cholesky upper triangle of the correlation matrix between diseases. In consequence, the columns of $\boldsymbol{\Phi}\mathbf{Q}$ will correspond, respectively, to the individual scores explaining the (standardized) first disease, the individual scores explaining the (standardized) second disease given the first, and so forth. CVA assumes those vectors of scores ($\boldsymbol{\Phi}\mathbf{Q}$) to be orthogonal combinations of common underlying spatial patterns. Those orthogonal combinations mean that all the columns of $\boldsymbol{\Phi}\mathbf{Q}$ share a common distribution (all of them are linear combinations of the same spatial patterns) and this therefore makes the modeling of $\boldsymbol{\Theta}$ order-free with regard to diseases, i.e., invariant to their ordering (Martinez-Beneito, 2013).

A second interesting interpretation of the CVA model also comes from the decomposition $\boldsymbol{\Sigma}_b = \mathbf{D}(\boldsymbol{M}^{*})'\boldsymbol{M}^{*}\mathbf{D} = \mathbf{D}\mathbf{W}\mathbf{D}$, where $\mathbf{W}$ follows a standard Wishart distribution $Wishart(K, \mathbf{I}_J)$. This is a scaled Wishart distribution as defined in Gelman et al. (2014). The scaled Wishart distribution has a clear advantage over the regular Wishart distribution as it separates the modeling of the variance parameters from that of the unscaled covariance structure. This allows it, for example, to be weakly informative on the scale parameters but more informative on the correlation structure of $\boldsymbol{\Sigma}_b$, since being too uninformative on that structure makes the marginal priors of their correlation parameters accumulate most of its mass at their extremes. In contrast, assuming an inverse $Wishart(J+1, \mathbf{I}_J)$ distribution on $\boldsymbol{\Sigma}_b$, which would mean putting flat prior distributions on its correlation parameters, assumes informative priors on its standard deviations (see page 286 in Gelman and Hill (2007)). Thus, the common degrees-of-freedom parameter of

the Wishart distribution seems to introduce modeling conflicts between the correlation and standard deviation parameters of $\boldsymbol{\Sigma}_b$. Mainly for these reasons, some authors advise the use of scaled Wishart priors instead of regular Wishart priors for modeling covariance/precision matrices (Barnard et al., 2000; Gelman and Hill, 2007; Gelman et al., 2014).

From a more practical point of view, the scaled Wishart distribution also allows specific inference to be performed on the different standard deviations $\sigma_1, ..., \sigma_J$ in a direct way. For a Wishart distribution, making inference on different standard deviations for each disease would require to increase the hierarchy of the model by setting $Wishart(K, \mathbf{D})$ and putting an additional layer in the model for $\mathbf{D}$, if the software available allows us to do so. The most popular inference tools for spatial modeling nowadays (`WinBUGS` and `INLA`) have only implemented the Wishart distribution to model precision matrices in multivariate settings. Therefore, the proposed modeling overrides this limitation by building the scaled Wishart distribution by itself.

The Wishart and scaled Wishart distributions are frequently used as priors for precision matrices, instead of for covariance matrices, as we have implicitly assumed in our proposal. In our opinion the main reason for this consensus in the literature could be that Wishart is the conjugate distribution for precision matrices of multivariate Normal variables, what yields substantial benefits in computational and analytical terms. The use of the $Wishart(J + 1, \mathbf{I}_J)$ as a prior distribution for precision matrices is particularly popular as it yields uniform marginal prior distributions on the correlation parameters between diseases (Barnard et al., 2000). Similarly, a scaled $Wishart(J + 1, \mathbf{I}_J)$ prior distribution on the covariance matrix would mean a uniform prior distribution on the partial correlation parameters. Instead, our proposal puts a scaled $Wishart(K, \mathbf{I}_J)$ distribution on the covariance matrix, which means that, for the common case $K = J$, this proposal should not be far from a uniform marginal prior distribution on the partial correlations between

diseases. Nevertheless, Figure 5.2 illustrates the performance of our scaled $Wishart(K, \mathbf{I}_J)$ prior for $\boldsymbol{\Sigma}_b$, for $K = J + 1$, in terms of the correlations between diseases. Each graph in that figure corresponds to the marginal distribution (histogram for 50,000 draws) of the correlation parameter between the first two diseases for scaled $Wishart(J + 1\mathbf{I}_J)$ prior distributions on $\boldsymbol{\Sigma}_b$, for $J = 3, 6, 12$ respectively. Figure 5.2 shows how the prior distributions for these settings concentrates on 0 as we increase the number of diseases. This is in contrast to the scaled $Wishart(J + 1, \mathbf{I}_J)$ prior distribution on the precision matrix, which yielded uniform prior distributions on the marginal correlations independently of J, that is, the number of diseases considered.

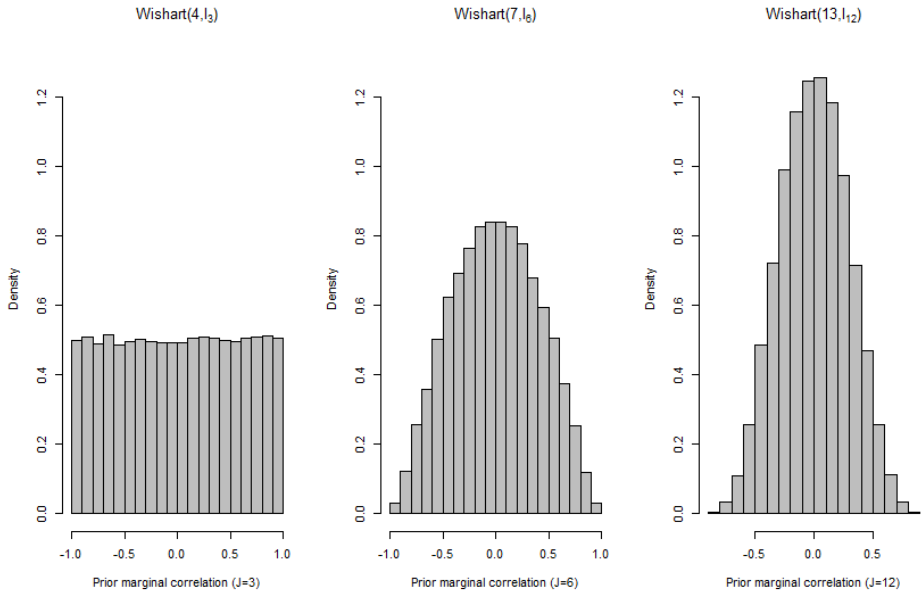


Figure 5.2.: Prior marginal distributions for the correlation for the first two diseases, out of a set of $J = 3, 6, 12$, assuming a $Wishart(J + 1, \mathbf{I}_J)$ distribution for $\boldsymbol{\Sigma}_b$. Histograms correspond to samples of 50,000 draws from the corresponding distribution of that marginal correlation.

Although the preference for small correlations might seem an undesirable effect for the scaled $Wishart(J + 1, \mathbf{I}_J)$ prior distribution on $\boldsymbol{\Sigma}_b$, it could be more desirable than expected. As we increase

the number of diseases, the number of marginal correlations between diseases in a model increases at a quadratic rate. Thus, assuming a uniform prior distribution for the marginal correlations between diseases would also mean a quadratic increase in false 'significant' correlations as the number of diseases increases. In contrast, a prior concentrating its mass on 0 when J grows would avoid this effect. This seems an interesting feature of our proposal as, when we increase the number of diseases in a multivariate study, we would expect the proportion of closely related diseases to go down instead of increasing at a quadratic rate. Hence, the prior structure proposed would perform a kind of multiplicity control on the number of related diseases (Scott and Berger, 2010), thereby inducing a parsimonious fit of the multivariate structure between them.

5.5. Some results of the CVA and RVA M-models

In this section we are going to show several empirical comparisons of the CVA and RVA $\boldsymbol{M}$-models with some other alternatives. First, we are going to explore the performance of these two models in a simulated data set, illustrating the enhanced handling of heteroscedasticity between diseases for these two models. Later, we will revisit the Castellón study mentioned above and we will illustrate how the CVA and RVA $\boldsymbol{M}$-models solve the issues shown in Section 5.3.

5.5.1. An analysis of some simulated data sets

We have performed a simulation to assess the heteroscedastic effect of the new RVA and CVA $\boldsymbol{M}$-models, as compared to the older NVA alternative. Specifically, we have considered the following settings for Castellón, Alicante and Valencia. We have generated an underlying pattern $\boldsymbol{x}$ following a proper CAR distribution of (conditional) standard

deviation σ_x and a spatial correlation parameter of 0.9. We have avoided generating our data from BYM models since, in real settings, the mechanism generating the observed data will never be the same as that used to analyze the data. Moreover, we have generated J additional spatial patterns $\boldsymbol{y}_1, ..., \boldsymbol{y}_J$ also following proper CAR distributions of spatial parameters equal to 0.9. Nevertheless, the (conditional) standard deviation for $\boldsymbol{y}_1$ was set to 1, while for $\boldsymbol{y}_2, ..., \boldsymbol{y}_J$ it was set to 0.2 in order to reproduce a heteroscedastic setting. Thus, J sets of observed counts were generated, supposedly representing different diseases, in the following manner:

$$O_{ij} \sim Poisson(E_{ij}RR_{ij}),\ i = 1, ..., I,\ j = 1, ..., J$$
$$log(RR_{ij}) = x_i + (\boldsymbol{y}_j)_i$$

Therefore, $\boldsymbol{x}$ induces correlation between diseases and the first of the diseases generated will show higher variability than the rest, due to the higher variability of $\boldsymbol{y}_1$. We have considered three different settings for each of our three cities: $\sigma_x = 0$, $\sigma_x = 0.5$ and $\sigma_x = 1$, which supposedly reproduce a gradient of increasing correlations between diseases. Specifically, for $\sigma_x = 0$ the correlation between diseases should be obviously equal to 0 for any two vectors of log-relative risks. For $\sigma_x = 0.5$ the first vector of log-relative risks should show a correlation of 0.42 with the rest of vectors, which should show correlations around 0.86. Finally, for $\sigma_x = 1$ the first vector of log-relative risks should show a correlation of 0.69 with the rest of vectors, which should show correlations around 0.96. According to the results in Botella-Rocamora et al. (2015), where all the correlations between diseases ranged from 0.06 to 0.76, $\sigma_x = 0.5$ would be the most realistic setting in practical terms. Anyway, all 3 settings considered will allow us to explore the effect of having different correlations between diseases in our data sets. We will refer to these three settings as Settings 1, 2 and 3, respectively. Note that the choice of just one common underlying pattern per setting is not motivated by any particular assumption of $\boldsymbol{M}$-models. As a

consequence, we do not find any reason, in principle, why the following results could not be generalized to settings with a higher number of underlying patterns.

Moreover, for each setting and city we have considered two different values of J, 5 and 10, in order to assess the effect that the number of diseases could have on the heteroscedasticity of data fitting. Therefore, we have a total of 18 (=3 settings * 3 cities * 2 values of J) different scenarios. Additionally, the seeds used for each setting, for $J = 5$ and 10, were exactly the same, and so $\boldsymbol{O}_{\cdot 1}, ..., \boldsymbol{O}_{\cdot 5}$ are equal for each city and setting, regardless of J. In this way we want to specifically assess the effect on heteroscedasticity of considering five additional diseases in a multivariate study. Additionally, we have generated five different data sets (replicates) for each scenario. For each of these replicates a different set of relative risks and corresponding observed cases have been generated. For all these replicates of these scenarios we have run: (i) a model with independent BYM models for each disease; (ii) the NVA homoscedastic $\boldsymbol{M}$-model; (iii) the CVA $\boldsymbol{M}$-model; and (iv) the RVA $\boldsymbol{M}$-model. All the $\boldsymbol{M}$-models run have underlying BYM models, as introduced in this paper. In sum, we have run a total of 18(scenarios)*5(replicates)*4(models)=360 models for this simulation study. The full `R` code and material needed to reproduce this simulation study can be found as accompanying material in the supplementary material (Annex B, Section B.2).

Table 5.1 summarizes the analysis that was performed. Each row corresponds to each of the 18 scenarios considered. The first three columns of that table set the scenario corresponding to each row, which is defined by the setting (none/ medium/ high dependence between diseases), city (Castellón, Alicante or Valencia) and number of diseases considered (5/10). Each of the 10 final columns summarize the results obtained for all five replicates of each scenario for the models run. The columns headed with 'Orig.' show the corresponding summary statistic for all five replicates in the original simulated data sets. The first

block of results in Table 5.1 summarizes the standard deviation (for all five replicates) corresponding to the first of the diseases in each scenario. These standard deviations correspond to the set of log-relative risks (their posterior means) for the first simulated disease.The second block of results also summarizes the standard deviations (mean of the standard deviations) but for the rest of the diseases as a whole in the simulation study. These two blocks are intended to illustrate the handling of heteroscedasticity for each of the models considered.

Table 5.1.: Standard deviations for the log-relative risks (their posterior means) for the first and subsequent diseases in the simulation study.

			First disease					Rest of diseases				
Setting	**City**	**Diseases**	**Orig.**	**BYM**	**NVA**	**CVA**	**RVA**	**Orig.**	**BYM**	**NVA**	**CVA**	**RVA**
1	Castellón	5	0.55	0.27	0.13	0.26	0.23	0.10	0.12	0.07	0.12	0.14
		10	0.55	0.27	0.06	0.27	0.18	0.10	0.12	0.04	0.12	0.12
	Alicante	5	0.66	0.22	0.07	0.22	0.17	0.10	0.07	0.03	0.07	0.08
		10	0.66	0.22	0.05	0.22	0.13	0.10	0.08	0.03	0.08	0.08
	Valencia	5	0.61	0.31	0.23	0.31	0.28	0.10	0.06	0.06	0.06	0.07
		10	0.61	0.31	0.09	0.31	0.23	0.10	0.05	0.03	0.05	0.06
2	Castellón	5	0.78	0.41	0.29	0.41	0.37	0.30	0.15	0.14	0.17	0.20
		10	0.78	0.41	0.27	0.43	0.36	0.31	0.17	0.18	0.20	0.24
	Alicante	5	0.77	0.27	0.15	0.27	0.23	0.27	0.11	0.09	0.12	0.14
		10	0.77	0.27	0.16	0.29	0.23	0.27	0.12	0.14	0.16	0.18
	Valencia	5	0.74	0.36	0.33	0.37	0.35	0.28	0.13	0.16	0.15	0.17
		10	0.74	0.36	0.32	0.38	0.34	0.28	0.12	0.19	0.19	0.21
3	Castellón	5	1.32	0.68	0.68	0.72	0.70	0.73	0.33	0.43	0.44	0.46
		10	1.32	0.68	0.67	0.75	0.71	0.74	0.38	0.49	0.50	0.52
	Alicante	5	1.07	0.55	0.53	0.58	0.55	0.60	0.27	0.39	0.40	0.42
		10	1.07	0.55	0.55	0.62	0.57	0.60	0.30	0.44	0.45	0.47
	Valencia	5	1.20	0.70	0.72	0.73	0.72	0.63	0.37	0.48	0.48	0.49
		10	1.20	0.70	0.73	0.75	0.73	0.63	0.38	0.52	0.52	0.53

Table 5.1 illustrates several interesting results. First, we can see how all four models, in general, oversmooth the original data for all the diseases. Nevertheless, focusing on the first disease, the oversmoothing is particularly evident for the NVA $\boldsymbol{M}$-model. The oversmoothing of NVA, in comparison to the rest of the models, diminishes when the correlations between diseases increase, being quite low for Setting 3, although for Setting 1 its effect is quite important. Note that the oversmoothing of the NVA $\boldsymbol{M}$-model for the first disease is slightly alleviated for the RVA $\boldsymbol{M}$-model and even more so for CVA. Note also that the oversmoothing of NVA for $J = 10$ is higher than for 5 for Setting 1 (and maybe slightly so for RVA), although this effect is not so apparent for the rest of the settings. This could explain the great problems found in Castellón for our case study, where 20 diseases were analyzed jointly, thereby making these problems even more worrisome. For BYM and CVA, the differences between $J = 10$ and 5 were irrelevant. This is not surprising for the BYM model, as it treats diseases as being completely independent, but Table 5.1 shows how that same independence effect, in terms of variance, is also achieved for CVA.

Although the non-superiority of the RVA/CVA models for Setting 3 could seem discouraging, in our opinion it is not so worrisome. First, as mentioned previously, Setting 3 reproduces quite high correlations that could be rare to find in practice. Setting 3 has been considered in the study for illustrating the RVA/CVA models performance gradient when correlations between diseases grow. Additionally, we find at least two factors that could explain that apparent loss of the RVA/CVA benefit for Setting 3. First, as designed this simulation study, heteroscedasticity decreases as a function of the settings. Thus, it is straightforward to check that for Setting 1 the standard deviation (between diseases) for the first disease is equal to 1 for the first disease and 0.2 for the rest, while those quantities are equal to 1.41 and 1.02 for Setting 3. Thus, the hypothetical advantage of RVA/CVA for Setting 3 would be lower, as evidenced in Table 5.1. On the other hand, the

variability for the underlying patterns in this study is higher for the correlated settings, being Setting 3 the setting with highest underlying variability. Therefore, heteroscedasticity would be easier to get captured by NVA (and also for the rest of models) in that setting, even though NVA is not particularly devised for taking that feature into account. As a consequence, NVA could reproduce heteroscedasticity better for this setting than for the other two alternatives. Therefore, the non-superiority of RVA and CVA for Setting 3 could be simply a consequence of the particular design assumed for this simulation study.

The columns for the (mean) standard deviations of the rest of the diseases also show some interesting results. Thus, we can see how, for Setting 1, BYM, CVA and RVA show similar variabilities and NVA might show a slight additional oversmoothing. Nevertheless, interestingly, for Setting 2 and more obviously for Setting 3, all three $\boldsymbol{M}$-models show more variability than the BYM model. For Settings 2 and 3, the oversmoothing is reduced as J increases, since more information is shared for a higher number of diseases. This shows the superiority of $\boldsymbol{M}$-models in general when correlated diseases are studied. These models take that correlation into account and are therefore able to alleviate the original oversmoothing of independent BYM models. Additionally, the paper also mentions: 'For these two cities (Alicante and Valencia), both multivariate models take advantage of the additional information provided by the set of diseases considered, depicting more detailed spatial patterns in general than their univariate alternatives'. Thus, those maps (included as supplementary material) show clearer spatial patterns than those drawn from independent BYM models. This result also supports the alleviation of the oversmoothing effect achieved by $\boldsymbol{M}$-models suggested above.

5.5.2. A new analysis of the Castellón mortality data

In this section we return to the geographical analysis of mortality in the city of Castellón and implement the new RVA and CVA

variance-adaptive proposals described in the previous section. In order to evaluate those proposals, we compare the new estimated risks with those obtained with the NVA $\boldsymbol{M}$-model and the univariate BYM models. The models have been executed in `WinBUGS` following the specifications introduced in Subsection 5.3.1. The `R` code for this analysis can also be found in the annex material (Annex B, Section B.2). The MCMC specifications for the models run in this section were basically the same as those used in Subsection 5.3.1.

Figure 5.3 shows the estimated risk maps with the new modeling proposals for AIDS, Cerebrovascular disease and Suicides in men in Castellón. As can be seen in the case of AIDS, the new modeling proposals provide risk maps with greater variability than that obtained with the NVA model and closely similar to those estimated with the univariate BYM model (Figure 5.1). In the case of Cerebrovascular disease and Suicide, the risk maps estimated with the new modeling proposals present a considerable lower variability than the risk maps for AIDS. This shows that both RVA and CVA have solved the problem presented by the original multivariate NVA model, which provided risk maps with a similar variability for all the diseases in the study. Nevertheless, the original patterns estimated by the univariate BYM models seem to be reinforced for both the RVA and the CVA models (mainly for the RVA model), almost certainly as a consequence of sharing information between diseases. RVA and CVA estimates for Valencia and Alicante can also be found as annex material in Annex B, Section B.1. Results for these cities confirm the visual conclusions also drawn for Castellón, although maybe to a lesser extent, since data for these cities are stronger than for Castellón. Thus, the results for these larger cities are far more robust to the multivariate model used to smooth the risks.

Besides the visual comparison of the estimated risk maps with the different modeling proposals, we have also compared the fit of these models in general terms by using the Deviance Information Criterion

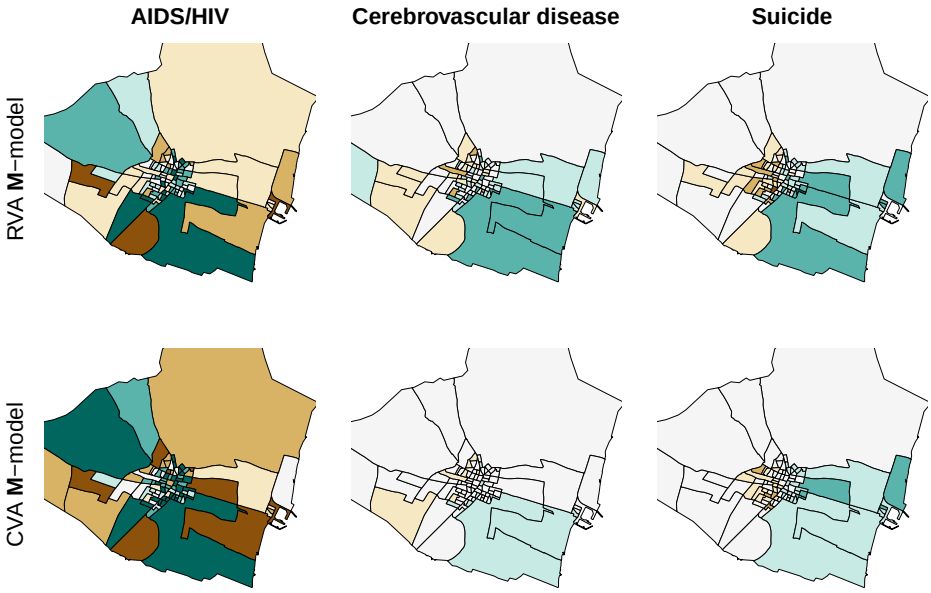


Figure 5.3.: Choropleth maps for the estimated risks using the new heteroscedastic RVA and CVA $\boldsymbol{M}$-models.

(DIC) proposed by Spiegelhalter et al. (2002). The DICs for all models and cities in the study can be found in Table 5.2. As can be observed, the model that provides a better fit in terms of the DIC in all three cities studied is the RVA $\boldsymbol{M}$-model, followed by the CVA $\boldsymbol{M}$-model in two out of three cities in the study. This seems to confirm that, besides the evident visual differences found, the heteroscedatic nature of the RVA and CVA models yields an important enhancement of the fit of the underlying geographical risk patterns.

5.6. Discussion

As described in this paper, the multivariate modeling proposal in Botella-Rocamora et al. (2015) for multivariate spatial studies of diseases presents some limitations when data are weaker. Specifically, in such

Table 5.2.: DICs for the adjusted models in all three cities in the study.

Model	Alicante	Castellón	Valencia
BYM with independent diseases	12964	6173	34270
Fixed effects $\boldsymbol{M}$-model	13212	6675	34416
Random effects $\boldsymbol{M}$-model	12865	6178	34029
RVA $\boldsymbol{M}$-model	**12798**	**6148**	**33918**
CVA $\boldsymbol{M}$-model	12870	6159	34009

situations, the prior structure of the $\boldsymbol{M}$-model can significantly influence the estimated risk patterns for all the diseases considered. As shown, this fact is caused by the single common variance parameter in the $\mathbf{M}$ matrix of this model, which controls the overall variability of all risk patterns fitted. As illustrated, the fixed effects $\boldsymbol{M}$-model has a tendency to yield barely smoothed risk estimates as a consequence of assuming a high prior variance for the log-risks for all diseases. As suggested by a reviewer, a deeper study of this model could perhaps conclude its subsequent impropriety coming from the improper prior distribution of the cells of $\mathbf{M}$. On the other hand, the random effects $\boldsymbol{M}$-model is prone to take all diseases, in terms of variability, to a common point that will be estimated by the model. If the variability of the risk patterns considered was different, these prior assumptions may produce evident misfits in the risk patterns that are estimated. One of the main contributions of this work has been to highlight these limitations, which are particularly worrisome when the original NVA proposal is applied to small regions of study.

In this work we have proposed two modifications of the previous multivariate model that incorporate several different parameters to model the variability of the risks for each disease and which allow us to solve the problems evidenced in the study. These new heteroscedastic proposals allow the spatial patterns for each disease to have greater or

lesser variability when necessary. This has made it possible to obtain more flexible and accurate risk estimates. Additionally, we have also introduced and discussed the formulation of $\boldsymbol{M}$-models based on the popular spatial dependence structure proposed by Besag et al. It is also worth mentioning that, in our opinion, the heteroscedastic proposals in this paper are only advisable for highly multivariate problems, that is, for at least a moderate number of diseases. For multivariate studies of just 2 or 3 diseases, the different variances of the $\mathbf{M}$ matrix in multivariate models would be estimated with just 2 or 3 observations, respectively, making unreliable those estimates. In that case we would rather to use the traditional NVA alternative since then all the cells of the $\mathbf{M}$ matrix would contribute to estimate the common assumed variance.

Regarding the two modeling proposals introduced in this paper, RVA has shown a better performance in empirical terms according to DIC. Thus, for our specific data sets, RVA seems to be more advisable despite the appealing interpretation of the CVA model as a scaled Wishart prior on $\boldsymbol{\Sigma}_b$. In any case, it would be worthwhile performing a more thorough and general comparison between these two models. Beyond these empirical results we also find the RVA model proposed interesting for several reasons. First, the PCA interpretation of the RVA approach seems quite interesting. Further work should be carried out under this approach in order to extract the 'principal maps' underlying this model since, as currently implemented, these factors cannot be identified by the model (some order restriction should be imposed, for example, in the vector of standard deviations $\boldsymbol{\Sigma}$ in order to identify those 'principal maps'). Nevertheless, those 'principal maps' with the municipal scores corresponding to the different principal axes is an interesting idea that is certainly worth exploring. Moreover, the spatial modeling of the 'principal maps' that the RVA model makes is also appealing. Assuming a spatial distribution for these components could be a sensible assumption since these common underlying components could perfectly reflect the spatial distribution of risk factors throughout

the region of study. In contrast, CVA assumes spatial distributions for the residual variability of each disease conditioned to the previous diseases in the study. We find this assumption much less realistic in practical terms. Nevertheless, the CVA proposal is interesting by itself because of its interpretation as a scaled Wishart prior for $\mathbf{\Sigma}_b$. As shown, the CVA model makes it possible to implement the scaled Wishart within some regular Bayesian packages, such as `WinBUGS`, and this could be of interest even beyond the disease mapping literature.

6. On the use of adaptive spatial weight matrices from disease mapping multivariate analyses

In this chapter, we present our paper "On the use of adaptive spatial weight matrices from disease mapping multivariate analyses" by Francisca Corpas-Burgos (Foundation for the Promotion of Health and Biomedical Research of Valencia Region) and Miguel A. Martinez-Beneito (University of Valencia) published in *Stochastic Environmental Research and Risk Assessment* (2020), 34:531–544.

Abstract

Conditional autoregressive distributions are commonly used to model spatial dependence between nearby geographic units in disease mapping studies. These distributions induce spatial dependence by means of a spatial weights matrix that quantifies the strength of dependence between any two neighboring spatial units. The most common procedure for defining that spatial weights matrix is using an adjacency criterion. In that case, all pairs of spatial units with adjacent borders are given the same weight (typically 1) and the remaining non-adjacent units are assigned a weight of 0. However, assuming all spatial neighbors

in a model to be equally influential could be possibly a too rigid or inappropriate assumption. In this paper, we propose several adaptive conditional autoregressive distributions in which the spatial weights for adjacent areas are random variables, and we discuss their use in spatial disease mapping models. We will introduce our proposal in a multivariate context so that the spatial dependence structure between spatial units is shared and estimated from a sufficiently large set of mortality causes. As we will see, this is a key aspect for making inference on the spatial dependence structure. We show that our adaptive modeling proposal provides more flexible and accurate mortality risk estimates than traditional proposals in which spatial weights for neighboring areas are fixed to a common value.

Keywords

Adaptive conditional autoregressive distributions, Gaussian Markov random fields, Multivariate disease mapping, Spatial weights matrix

6.1. Introduction

Disease mapping has attracted considerable attention over the last three decades (Lawson, 2018; Martinez-Beneito and Botella Rocamora, 2019). This area of research pursues the study of the geographical distribution of health-related events, such as mortality from, or incidence of diseases, aggregated over areal units, in order to identify mainly those locations which show higher risks. In disease mapping problems, the units of study usually considered are as small as possible, which can lead to what are known as small areas estimation problems. As a consequence, many modeling proposals have been formulated in order to deal with this problem and thereby derive reliable risks estimates. Most of these models consider dependence among nearby spatial units, assuming them

to show similar risks. Therefore the spatial dependence hypothesis is the main key to improving risks estimates in disease mapping studies.

A large number of disease mapping models have been proposed, most of them following a Bayesian approach; see Besag et al. (1991) or Leroux et al. (1999) for two of the most frequently used models in applied studies. These proposals are frequently specified as generalized linear models that incorporate spatial dependence between nearby geographical units through random effects following some spatial prior distribution. Although some other spatial modeling tools have been also used (Adin et al., 2017), the most popular spatial prior distributions in disease mapping models belong to the family of Conditional Autoregressive (CAR) distributions (Besag, 1974; Besag et al., 1991), also known as Gaussian Markov Random Fields (GMRF) (Rue and Held, 2005). CAR distributions induce spatial dependence by considering a schematic neighborhood structure which accounts for the geographical arrangement of the spatial units. That neighborhood structure is summarized by means of a spatial weights matrix quantifying the relative influence that the random effects of the geographical units have on each other, so those weights should reflect the strength of the dependence between any pair of spatial units. Moreover, that weights matrix is usually sparse, reflecting an implicit Markovian assumption which considers the conditional distribution of any random effect, given its neighbors, independent of the random effects in any other spatial location.

Different proposals of spatial weights matrices for CAR distributions have been made in the literature. By far, the most common procedure is using an adjacency criterion for defining that matrix. In that case all pairs of spatial units with adjacent geographical borders are given the same weight, typically 1, and the remaining non-adjacent units are assigned a weight of 0, reflecting independence given the remaining spatial units (Besag et al., 1991). As pointed out by Raftery and Banfield (1991), this choice could be sensible for regular lattices but less

so for irregular lattices such as those typically used in disease mapping problems. Weighted versions of this common choice also exist which are available in some common Bayesian inference packages, such as `WinBUGS` or `OpenBUGS`; while some others, such as `INLA`, do not allow for this option. This allows the strength of the dependence of nearby pairs of units to be modulated, thereby allowing each neighboring pair of units to show a different strength. However, the CAR distributions in `WinBUGS` or `OpenBUGS` do not allow those weights to be estimated as variables within a model; on the contrary, they have to be supplied as constants to the corresponding model. Best et al. (1999) and Earnest et al. (2007), for example, propose the use of weights matrices with different weights which are a function of the geographic distance between spatial units (usually, the Euclidean distance between their centroids); in this manner random effects of closer geographic units will show stronger dependence. However, it could happen that geography is not necessarily the main determinant of dependence between units; thus areas with similar values of certain covariates, for example, would take similar risks estimates in general even though they are distant. In this regard, Kuhnert (2003) defines the weights matrix of the random effects as a function of the absolute difference between the values of some covariate for the spatial units. Likewise, Earnest et al. (2007) define the weights as a function of both the geographical distance and their similitude in terms of some covariate. A comparison of models of this kind is undertaken in Duncan et al. (2017).

Despite their interest to researchers, the use of the weight matrices above shows some limitations. Firstly, unweighted adjacency-based matrices do not have clear support beyond their simplicity and convenience. In the end, assuming all spatial neighbors in a model to be equally influential is an arbitrary assumption that should be checked in some way. Nevertheless, the mentioned convenience of that choice has led most disease mapping practitioners to accept and use that matrix, without further justification, and to avoid questioning that assumption. On the other hand, the use of functions of geographic

distance for setting weights matrices assumes equal weights for all locations which are equally distant, which could be somewhat simplistic for some settings (consider regions with mountains, rivers or other barriers). Additionally, those distance weighted proposals assume parametric relationships between distances and weights, which could also be rigid or sometimes inappropriate. Finally, the definition of weights as a function of some covariate poses an additional problem since the corresponding covariate may not always be available for all locations. Therefore, the requirements for this option are higher than for pure geometric criteria.

The objective of this work is to propose a procedure for estimating the spatial weights matrix in disease mapping studies solving the issues above. Specifically, we focus on the barely explored adaptive CAR distributions which consider the weights of the spatial weights matrix as additional random variables in the model. Some works can be found in the literature, such as MacNab et al. (2006b); Brezger et al. (2007); Lu et al. (2007); Congdon (2008); Ma et al. (2010), that follow this approach. Our proposal, in contrast to the previous works, estimates a common weights matrix from the joint study of several diseases, which would, presumably, capture the different dependence strengths shown by the neighboring spatial units in the region of study. As we will see, that multivariate feature of our proposal will be a key aspect for its success. The multivariate estimated weights matrix could be subsequently used in future studies on that same region of study. In principle the enhanced weights structure estimated for that region would allow improved risk estimates to be derived incorporating the dependence structure shown by some set of diseases in that region. That spatial structure should reflect physical/social barriers, data artifacts, geographical/geometrical/social features etc., which would be recommendable to consider in subsequent spatial analyses on that same region.

This paper is structured as follows. Section 6.2 introduces some traditional spatial modeling proposals widely used in disease mapping

studies and makes a brief review of the main adaptive CAR models already proposed in the literature. Section 6.3 describes our multivariate modeling proposal for the weights matrix of CAR distributions. Section 6.4 illustrates how the developments proposed at Section 6.3 can be used for estimating the spatial dependence structure in a real setting and how that estimation improves subsequent analyses in comparison to studies with unweighted dependence matrices. Finally, Section 6.5 discusses some results and conclusions drawn from this study.

6.2. Some modeling proposals in disease mapping

6.2.1. Some popular disease mapping models

Disease mapping studies consider regions of interest discretely divided into I spatial units, generally of small size, such as census tracts or municipalities. The main aim of these studies is to determine the geographical distribution of the risks for some disease for these spatial units. The collection of observed cases per spatial unit are jointly denoted by $\boldsymbol{O} = (O_1, \ldots, O_I)'$, where O_i denotes the number of observed cases in the i-th unit. Typically, disease mapping models assume:

$$O_i \sim Poisson(E_i R_i), \quad i = 1, \ldots, I,$$

where $\boldsymbol{E} = (E_1, \ldots, E_I)'$ contains the number of expected cases per spatial unit for the corresponding disease and $\boldsymbol{R} = (R_1, \ldots, R_I)'$ is the collection of location specific risks that we would want to estimate. Typically, the log-risks are modeled as:

$$log(R_i) = \mu + \boldsymbol{z}_i^{'}\boldsymbol{\beta} + \eta_i, \tag{6.1}$$

where μ is an intercept, $\boldsymbol{z}_i$ is a vector of covariates, with $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)'$ being its vector of associated parameters, and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_I)'$ a vector

of random effects. The random effects $\boldsymbol{\eta}$ are introduced to explain the variability that cannot be explained by the covariates and $\boldsymbol{\eta}$ is typically assumed to show spatial dependence since that residual variability could easily exhibit that feature. From now on, for simplicity, we will assume that no covariates are available and the log-risks are simply modeled as the sum of the intercept and the set of random effects.

The random effects vector $\boldsymbol{\eta}$ is habitually modeled by using spatially correlated CAR prior distributions. A particularly popular case of CAR prior distribution is the Intrinsic CAR (Besag et al., 1991) distribution (ICAR from now on), which for a vector $\boldsymbol{\phi}$ may be defined by the following set of I univariate conditional distributions:

$$\phi_i|\boldsymbol{\phi}_{-i}, \sigma_\phi^2 \sim N\left(\frac{1}{w_{i+}}\sum_{k=1}^{I} w_{ik}\phi_k, \frac{\sigma_\phi^2}{w_{i+}}\right),\ i = 1, ..., I. \tag{6.2}$$

In this expression, the subindex in $\boldsymbol{\phi}_{-i}$ denotes all the terms in $\boldsymbol{\phi}$ excepting its i-th component, w_{ik} weighs the contribution of the k-th random effect to the mean of ϕ_i, $w_{i+} = \sum_{k=1}^{I} w_{ik}$ and σ_ϕ^2 is a variance parameter. These conditional distributions can be shown (Besag, 1974) to yield the following joint distribution for $\boldsymbol{\phi}$:

$$\boldsymbol{\phi}|\sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2(\boldsymbol{D} - \boldsymbol{W})^-),$$

where $\boldsymbol{D} = diag(w_{1+}, ..., w_{I+})$, $\boldsymbol{W} = (w_{ik})_{i,k=1}^{I}$ and the superindex in $(\boldsymbol{D} - \boldsymbol{W})^-$ denotes the Moore-Penrose inverse of $\boldsymbol{D} - \boldsymbol{W}$. Covariance between elements of $\boldsymbol{\phi}$ is determined by the spatial weights matrix $\boldsymbol{W}$, whose elements w_{ik} are typically non-zero if areas (i, k) are considered neighbors and zero otherwise. Therefore, if two areas are considered neighbors, their random effects are conditionally dependent, while random effects of non-neighboring areas are conditionally independent. As mentioned previously, the most common approach is to assume that areas (i, k) are neighbors if they share a common border (adjacency) and in that case set $w_{ik} = 1$ for all neighboring pairs of units (i, k). In

that case, the conditional distributions above reduce to simply:

$$\phi_i|\boldsymbol{\phi}_{-i}, \sigma^2_\phi \sim N\left(\frac{1}{n_i}\sum_{k\sim i}\phi_k, \frac{\sigma^2_\phi}{n_i}\right),\ i = 1, ..., I, \tag{6.3}$$

where n_i stands for the number of neighboring areas of unit i and the subindex $k \sim i$ denotes all those units k which are neighbors of i. Now, the conditional mean of ϕ_i is equal to the raw (unweighted) mean of the random effects in its neighboring areas and its conditional variance is inversely proportional to the number of neighbors n_i.

One of the most popular specifications for $\boldsymbol{\eta}$ in disease mapping studies is that introduced in Besag et al. (1991) (BYM from now on). In this proposal, the random effects vector $\boldsymbol{\eta}$ is considered to be the sum of two vectors of random effects $\boldsymbol{\eta} = \boldsymbol{\phi} + \boldsymbol{\theta}$. The term $\boldsymbol{\phi}$, which follows an ICAR distribution as just introduced, will be responsible for inducing spatial dependence on $\boldsymbol{R}$ and accounts for those risk factors of regional scope which take effect on several contiguous spatial units, making them in principle similar. The second term, $\boldsymbol{\theta}$, whose components follow independent Normal distributions of mean zero and common variance σ^2_θ, accounts for risk factors of very limited geographical scope that take an effect just on isolated areal units, making their risks different to those of their surrounding units. Thus, this second term induces additional unstructured variability in $\boldsymbol{\eta}$. The amount of spatial/unstructured variability in $\boldsymbol{R}$ depends on the balance between σ_ϕ and σ_θ, which is determined by the model/data itself. If the former has higher (respectively lower) values, in comparison to the latter, the final pattern will show substantial spatial dependence (respectively independence).

A second popular CAR prior distribution for inducing spatial correlation on the random effects vector $\boldsymbol{\eta}$ in Expression (6.1) is that introduced in Leroux et al. (1999). In contrast to the BYM model, $\boldsymbol{\eta}$ in this alternative proposal is not the sum of two additional components. In this case, the determination of the amount of spatial/unstructured

variability is controlled by a spatial correlation parameter $\rho \in [0,1]$ so that the special case of $\rho = 0$ simplifies to a model with independent random effects and $\rho = 1$ corresponds to the ICAR distribution above. All intermediate values of $\rho \in (0,1)$ induce patterns mixing both sources of dependence. Specifically, for the Leroux et al. proposal, the prior conditional distributions corresponding to η_i are given by:

$$\eta_i|\boldsymbol{\eta}_{-i}, \rho, \sigma_\eta^2 \sim N\left(\frac{\rho}{\rho w_{i+} + 1 - \rho}\sum_{k=1}^{I} w_{ik}\eta_k, \frac{\sigma_\eta^2}{\rho w_{i+} + 1 - \rho}\right),\ i = 1, ..., I.$$

Note the obvious coincidence of this proposal with a weighted CAR distribution for $\rho = 1$ and with a heterogeneous Normal distribution for $\rho = 0$. For the usual assumption of $w_{ik} = 1$ for adjacent spatial units, and 0 otherwise, the Leroux et al. proposal reduces to:

$$\eta_i|\boldsymbol{\eta}_{-i}, \rho, \sigma_\eta^2 \sim N\left(\frac{\rho}{\rho n_i + 1 - \rho}\sum_{k\sim i} \eta_k, \frac{\sigma_\eta^2}{\rho n_i + 1 - \rho}\right),\ i = 1, ..., I.$$

In the same manner as for the ICAR distribution, these conditional distributions yield a joint Normal distribution, specifically:

$$\boldsymbol{\eta}|\rho, \sigma_\eta^2 \sim N_I(\mathbf{0}, \sigma_\eta^2((1-\rho)\boldsymbol{I}_I + \rho(\boldsymbol{D} - \boldsymbol{W}))^-),\ i = 1, ..., I,$$

where, as for BYM, $\boldsymbol{W} = (w_{ik})_{i,k=1}^I$ denotes the spatial weights matrix considered.

As described in the introduction, setting the same weights $w_{ik} = 1$ to all the random effects of adjacent locations in CAR distributions could be an inappropriate or rigid assumption. This makes all neighboring regions equally influential on any particular risk, which may not correspond to reality. In order to solve this limitation, models with alternative stochastic weight matrices have been proposed and are reviewed in the next subsection.

6.2.2. Adaptive CAR distributions

A few CAR models with adaptive weights matrices can be found in the Bayesian disease mapping literature. The goal of these proposals is to model spatial correlation through the fitting of an stochastic spatial weights matrix $\boldsymbol{W}$. This approach is undertaken within Bayesian hierarchical models where the corresponding CAR distributed random effects are defined as $\boldsymbol{\eta}|\boldsymbol{W}, \sigma_\eta^2, ...$ and an additional layer is considered in the model for estimating the elements in $\boldsymbol{W}$. Next, we briefly summarize several of the contributions in this area. There is obviously a huge body of literature proposing the stochastic modeling of variances or covariances between variables in many different contexts: state space models (Carter and Kohn, 1996), function estimation (Lang et al., 2002), etc. Nevertheless, we will restrict the review below just to CAR spatial models in order to focus on the particular topic that we are discussing where the variance matrix is defined as a function of a particular weights matrix $\boldsymbol{W}$.

Wombling (Lu and Carlin, 2005; Lu et al., 2007; Ma and Carlin, 2007; Ma et al., 2010) would be a first attempt of stochastic modeling of the spatial weights matrix $\boldsymbol{W}$ in CAR distributions. Specifically, Wombling assumes the cells of $\boldsymbol{W}$ as binary stochastic values, which are modeled as Bernoulli distributions. The probability of $w_{ij} = 1$ for any pair of spatial units (i, j) could be modeled by means of a logistic regression as a function of some covariates (Lu et al., 2007), such as the adjacency matrix of the area of study or some other related quantity. Obviously, the number of elements in that logistic regression will increase quadratically as a function of the number of spatial units in the study, which could be a problem for large lattices. In addition, a large collection of sensible covariates would be required under this approach in order to define a rich enough spatial weights matrix. More flexible alternatives are also considered for estimating $\boldsymbol{W}$ within the Wombling approach, although in this case only the weights of the cells corresponding to adjacent elements in the lattice are estimated. In this case, since w_{ij} are modeled as Bernoulli variables for adjacent units, this procedure

will prune the adjacency graph originally considered. Ma et al. (2010) proposes a spatial Ising model (see for example Geman and Geman (1984) or Green and Richardson (2002)), which favors contiguous edges in the graph (those sharing one of their nodes) to take the same values in the weights matrix. This proposal has also been applied to multivariate data sets, for the joint study of three diseases, as in Ma and Carlin (2007). Although this proposal seems much more flexible than the naive use of covariates for modeling $P(w_{ij} = 1)$, the binary treatment of the elements of the weights matrix in Wombling studies seems somewhat restrictive. Moreover, the use of Ising models for estimating the non-zero cells of $\boldsymbol{W}$ induces important computational problems (Ma and Carlin, 2007) for estimating the penalizing parameter of that model, at least in the multivariate case. This forces this parameter to be fixed/tuned according to previous runs of the models. However, as reported by the authors, the fit of this model even for a fixed penalizing parameter becomes challenging for medium/large lattices.

On the other hand, MacNab et al. (2006b) and Congdon (2008) consider adaptive versions of the Leroux et al. CAR prior distribution. This approach could be also used for estimating weights matrices in CAR distributions as will become evident in the next section. These proposals allow the spatial correlation parameter ρ of Leroux et al. to vary for each geographical unit. Specifically, MacNab et al. (2006b) propose defining an unweighted (adjacency-based) spatial process as the following set of conditional distributions:

$$\eta_i|\boldsymbol{\eta}_{-i}, \boldsymbol{\rho}, \sigma_\eta^2 \sim N\left(\frac{\rho_i}{\rho_i n_i + 1 - \rho_i}\sum_{k\sim i}\eta_k, \frac{\sigma_\eta^2}{\rho_i n_i + 1 - \rho_i}\right),\ i = 1, ..., I.$$

The problem with this proposal is that this set of conditional distributions does not yield a valid CAR prior distribution since the symmetry condition (Besag and Kooperberg, 1995), necessary for $\boldsymbol{\eta}$ to have a symmetric covariance matrix, does not hold in this case. Interestingly, regarding the spatial correlation parameters $\boldsymbol{\rho} = (\rho_1, ..., \rho_I)$, MacNab et al. (2006b) mention that 'the analysis

showed very little prior-to-posterior updating for the ρ_js, an indication that the data did not provide enough information for useful posterior inference'. Congdon (2008), also in the unweighted case for simplicity, proposes the following set of conditional distributions:

$$\eta_i|\boldsymbol{\eta}_{-i}, \boldsymbol{\rho}, \sigma_\eta^2 \sim N\left(\frac{\rho_i}{\rho_i n_i + 1 - \rho_i}\sum_{k\sim i}\rho_k\eta_k, \frac{\sigma_\eta^2}{\rho_i n_i + 1 - \rho_i}\right),\ i = 1, ..., I,$$

which fulfills the mentioned symmetry condition. Although this proposal is supposed to extend the Leroux et al. CAR distribution to having different correlation parameters $\rho_i, i = 1, ..., I$, strikingly it does not coincide with that proposal when all those ρ_i take a single common value ρ. Moreover, this process yields the following joint covariance matrix:

$$\sigma_\eta^2(diag(\mathbf{1}_I - \boldsymbol{\rho})\boldsymbol{I}_I + diag(\boldsymbol{\rho})(\boldsymbol{D} - \boldsymbol{W}diag(\boldsymbol{\rho})))^-.$$

In this case the covariance matrix is not a combination of $\boldsymbol{I}_I$ and $\boldsymbol{D}-\boldsymbol{W}$ since this latest term is replaced by $\boldsymbol{D} - \boldsymbol{W}diag(\boldsymbol{\rho})$. As a consequence, this proposal does not seem such a straightforward generalization of the Leroux covariance matrix. Congdon (2008) suggests several different prior distributions for the components of $\boldsymbol{\rho}$, such as beta, probit-normal or logit-normal distributions which allow further modeling of these variables. This adaptive CAR distribution is proposed and used in the univariate context where a single spatial pattern is studied.

Brezger et al. (2007) propose an alternative adaptive model which makes it possible to make inference on the spatial weights matrix of ICAR prior distributions. The context of this paper is spatio-temporal modeling in human brain mapping, but their ideas could be also useful for regular disease mapping studies. Brezger et al. uses an ICAR as prior distribution for the coefficients of some basis of functions modeling the time trend for a set of brain voxels; a different ICAR distribution is used for each element in that basis. Under this approach, the cells of the weights matrices for those ICAR random effects are considered equal to

0 for all non-adjacent pairs of units and are modeled as positive random variables, following an informative Gamma(0.5,0.5) prior distribution for the adjacent areas. Posterior sensitivity to that informative prior distribution is not assessed in the paper. Additionally, the likelihood for the Brezger et al. proposal is Normal so the applicability of their model and results to regular disease mapping studies, usually with Poisson or binomial likelihoods, should be further explored.

6.3. A new adaptive CAR distribution and its use in multivariate models

As an alternative to the commonly used adjacency criterion, which considers all the weights in $\boldsymbol{W}$ as fixed binary quantities, we propose to model the spatial weights as random variables within the model so as to allow variability between them. We begin by describing two proposal in the simplest univariate case for both ICAR and Leroux et al. spatial distributions. Subsequently, we will describe their equivalents in the context of the multivariate study of several diseases since, as we will argue, this is the appropriate context where adaptive proposals should be implemented.

6.3.1. Univariate case

We start first by introducing the estimation of spatial weights matrices for ICAR distributions. Let $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_I)'$ be a random effects vector with ICAR distribution, that is, $\boldsymbol{\phi}|\sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2(\boldsymbol{D} - \boldsymbol{W})^-)$. For this expression we will assume that $\boldsymbol{D}$ and $\boldsymbol{W}$ are defined according to adjacency between spatial units, i. e. $\boldsymbol{D} = diag(n_1, \cdots, n_I)$ for n_i the number of neighbors of unit i and $\boldsymbol{W} = (w_{ik})$ where $w_{ik} = 1$ if (i, k) are adjacent units and 0 otherwise.

Let us now consider a random vector $\boldsymbol{c} = (c_1, \cdots, c_I)'$ of positive values, a new spatial weights matrix $\boldsymbol{W}^*(\boldsymbol{c}) = diag(\boldsymbol{c})^{1/2}\boldsymbol{W}diag(\boldsymbol{c})^{1/2}$ and $\boldsymbol{D}^* = diag(w^*_{1+}, ..., w^*_{I+})$. With this, we propose the following adaptive CAR prior distribution:

$$\boldsymbol{\phi}|\boldsymbol{c}, \sigma^2_\phi \sim N_I(\boldsymbol{0}, \sigma^2_\phi(\boldsymbol{D}^* - \boldsymbol{W}^*(\boldsymbol{c}))^-)$$

$$c_i \sim Gamma(\alpha, \alpha).$$

The elements of the vector $\boldsymbol{c}$ are assumed to be positive since the non-zero weights of the new spatial weights matrix $\boldsymbol{W}^*$ are $w^*_{ij} = (c_i c_j)^{1/2}$ so, in this manner they will all be well defined and positive. Accordingly, we have used a Gamma prior distribution for its elements, which seems a natural choice. The Gamma distribution considered has mean 1, in accordance with the value of the non-zero cells of $\boldsymbol{W}$ when an adjacency criterion is considered. Thus, $\boldsymbol{W}^*(\boldsymbol{c})$ will be on average equal to $\boldsymbol{W}$, although its non-zero weights will not necessarily have to be equal to 1. Hence the new adaptive distribution will be more flexible than the regular ICAR distribution. Note that, as defined, the (a priori) standard deviation of any element of $\boldsymbol{c}$ is equal to $\alpha^{-0.5}$, which could guide us to set a prior distribution for this parameter. In fact, we have considered a uniform prior distribution on $\alpha^{-0.5}$, with lower and upper limits intended to make it vague, in order to complete the hierarchical structure above.

Alternatively, the definition of the adaptive ICAR distribution above could be restated as a set of conditional distributions $\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{c}, \sigma^2_\phi,\ i = 1, ..., I$, of mean

$$E[\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{c}, \sigma^2_\phi] = \frac{1}{w^*_{i+}} \sum_{k=1}^{I} w^*_{ik}\phi_k = \frac{c_i^{1/2} \sum_{k\sim i} c_k^{1/2}\phi_k}{c_i^{1/2} \sum_{k\sim i} c_k^{1/2}} = \frac{\sum_{k\sim i} c_k^{1/2}\phi_k}{\sum_{k\sim i} c_k^{1/2}} \qquad (6.4)$$

and variance

$$Var[\phi_i|\boldsymbol{\phi}_{-i}, \boldsymbol{c}, \sigma_\phi^2] = \frac{\sigma_\phi^2}{w_{i+}^*} = \frac{\sigma_\phi^2}{c_i^{1/2} \sum_{k\sim i} c_k^{1/2}}. \tag{6.5}$$

These two expressions provide some quite valuable insights about on the model proposed. The expected value in Expression (6.4) is just a weighted mean of the random effects for the corresponding neighbors. The weights in that expression are given by the vector $\boldsymbol{c}$, thus if c_i had a low value for some i, that region will have a low contribution to the means of its surroundings units. Additionally, Expression (6.5) suggests that if c_i is low, then the conditional variance of ϕ_i will be in contrast high. Thus, if c_i was low, these two expressions suggest that it is as if spatial unit i was 'disconnected' from its spatial neighbors, since ϕ_i will be less influential on them and will have higher variance, allowing it to move independently from the rest of the units. Conversely, if c_i was high, unit i will become more influential on its neighbors and will take a value in close agreement with them. Therefore, in some manner, the adaptive ICAR distribution would impose a tighter dependence between this unit and its neighbors.

Besides the enhanced interpretation just made, the conditional statement of the adaptive ICAR distribution above allows its implementation in conventional Bayesian software packages such as `WinBUGS`, `JAGS` ... Additional care should be taken for the adaptive ICAR distribution since sum-to-zero restrictions are, in general, imposed on ICAR distributions in order to solve the rank-deficiency of its precision matrix (Besag and Kooperberg, 1995). This can be done in practice, in a computationally convenient manner, by imposing $\sum_i \phi_i \sim N(0, \epsilon)$ for some small value ϵ. Details of the coding of this constraint for the adaptive ICAR distribution can be found in the supplementary material of the paper (Annex C, Section C.2).

In the case of the Leroux et al. model $\boldsymbol{\phi}$ is distributed as $\boldsymbol{\phi}|\rho, \sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2((1-\rho)\boldsymbol{I}_I + \rho(\boldsymbol{D}-\boldsymbol{W}))^-)$. Following the development

above, several adaptive versions of the Leroux et al. distribution could be made. For example, let us assume $\boldsymbol{\phi}|\rho, \boldsymbol{c}, \sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2((1-\rho)diag(\boldsymbol{c}^{1/2}) + \rho(\boldsymbol{D}^* - \boldsymbol{W}^*(\boldsymbol{c})))^-)$, where $\boldsymbol{D}^*$ and $\boldsymbol{W}^*(\boldsymbol{c})$ are defined as for the adaptive ICAR distribution. In this manner, for $\rho = 1$ this distribution would be equivalent to an adaptive ICAR distribution, while for $\rho = 0$ it would yield a collection of independent Normal random effects with adaptive (heteroscedastic) variance. If preferred, an alternative formulation of adaptive Leroux distribution could be derived as $\boldsymbol{\phi}|\rho, \boldsymbol{c}, \sigma_\phi^2 \sim N_I(\mathbf{0}, \sigma_\phi^2((1-\rho)\boldsymbol{I}_I + \rho(\boldsymbol{D}^* - \boldsymbol{W}^*(\boldsymbol{c})))^-)$ which for $\rho = 0$ yields non-adaptive (homoscedastic) independent random effects. Nevertheless, we will focus in the first of these options as it seems more flexible and appealing, from our perspective. For that proposal, the conditional mean and variance of the random effect ϕ_i can be expressed as:

$$\begin{aligned} E[\phi_i|\boldsymbol{\phi}_{-i}, \rho, \boldsymbol{c}, \sigma_\phi^2] &= \frac{\rho}{(1-\rho)c_i^{1/2} + \rho w_{i+}^*} \sum_{k=1}^{I} w_{ik}^* \phi_k \\ &= \frac{\rho c_i^{1/2}}{(1-\rho)c_i^{1/2} + \rho c_i^{1/2} \sum_{k\sim i} c_k^{1/2}} \sum_{k\sim i} c_k^{1/2} \phi_k \\ &= \frac{\rho}{1-\rho+\rho \sum_{k\sim i} c_k^{1/2}} \sum_{k\sim i} c_k^{1/2} \phi_k \end{aligned}$$

and

$$Var[\phi_i|\boldsymbol{\phi}_{-i}, \rho, \boldsymbol{c}, \sigma_\phi^2] = \frac{\sigma_\phi^2}{(1-\rho)c_i^{1/2} + \rho w_{i+}^*} = \frac{\sigma_\phi^2}{c_i^{1/2}(1-\rho+\rho \sum_{k\sim i} c_k^{1/2})}.$$

Once again, these conditional expressions allow the adaptive Leroux CAR distribution to be implemented in conventional Bayesian software, such as `WinBUGS`. Note that similar, although somewhat different, adaptive CAR distributions have been already proposed in MacNab (2018). Nevertheless, those proposals had as a goal the formulation of more general (adaptive) CAR distributions. Our goal will now be to estimate spatial weight matrices $\boldsymbol{W}^*(\boldsymbol{c})$ suitable to be used in

subsequent spatial analyses in that same area of study. It is hoped that $\boldsymbol{W}^*(\boldsymbol{c})$ will capture the geometric/demographic/geographic features of the region of study which make some neighboring units more similar to their neighbors than others.

6.3.2. Multivariate case

Although the univariate models above seem quite appealing, one could be concerned about including another level in the hierarchy of the model containing that adaptive CAR distribution. Moreover, that additional layer would contain as many variables as observations in the univariate model, so this modification increases the number of variables in the model in a quite significant manner. As a consequence, data in univariate disease mapping models may be not strong enough as to make inference on vector $\boldsymbol{c}$ possible. As mentioned earlier, this was already suggested by MacNab et al. (2006b) and we agree with that point of view. For example, let us assume that we performed an univariate disease mapping study with an adaptive CAR distribution where the weights vector $\boldsymbol{c}$ should be estimated. Let us also assume that the number of observed events for spatial unit i was abnormally higher than the corresponding number of expected cases. This would make the corresponding log-risk ϕ_i take a large positive value. It seems clear that, in this case, if c_i was low, this would help ϕ_i to reach that goal by allowing it to have a more independent performance, in comparison to its neighbors, and a higher variance. But, what makes the risk of this disease so strange for this spatial unit? Is it the specific particularities of the disease under study in that precise unit (ϕ_i) or the spatial unit itself that, for some structural (geographical, social, environmental, etc.) reason, is far different in general to its surrounding spatial units (c_i)? With a single observation per spatial unit the model does not have enough information to distinguish these two settings and therefore to estimate $\boldsymbol{c}$ properly. In contrast, if we had several risk estimates for several diseases ϕ_{ij} depending on a common weights matrix $\boldsymbol{W}^*(\boldsymbol{c})$,

we would be able to know if the risk of the original disease in the i-th spatial unit was really different, or the differential factor was the spatial unit. In the first case, among all the log-risks of the i-th spatial unit, only that corresponding to the original disease should take a high value and c_i should not therefore take a low value since that unit does not have a different performance in general than its neighbors. In the second, all (or most of) the log-risks for the i-th spatial unit would take extreme values and c_i should take a low value in order to accommodate that behavior. As a consequence, the use of adaptive CAR distributions in the context of multivariate studies could improve the fit of the spatial weights vector $\boldsymbol{c}$ to a considerable extent. This issue will become clearer in the real case study in the next section.

In accordance to the previous paragraph, we introduce now a multivariate model integrating adaptive CAR distributions, one per disease, with a common spatial weights matrix. This formulation allows an appropriate estimation of the vector $\boldsymbol{c}$ and therefore an appropriate estimation of the weights matrix $\boldsymbol{W}^*(\boldsymbol{c})$ corresponding to the set of diseases and region of study considered. The following formulation implements an adaptive BYM model per disease, although a similar formulation could be analogously proposed for the case of the Leroux et al. CAR distribution.

Let O_{ij} represent the observed number of cases for the i-th spatial unit and j-th disease, $i = 1, ..., I$, $j = 1, ..., J$. We will assume:

$$O_{ij} \sim Poisson(E_{ij}R_{ij}),$$

where E_{ij} is the number of expected cases, and R_{ij} the relative risk for the corresponding spatial unit and disease. In accordance with the univariate model, the log-risks can be expressed as:

$$log(R_{ij}) = \mu_j + \eta_{ij}, \tag{6.6}$$

where μ_j stands for an intercept modeling the mean of the log-risks

for each disease and η_{ij} are random effects accounting for spatial or unstructured variability for those risks. We will model the columns of $\boldsymbol{\eta} = (\eta_{ij})$ by means of a BYM model, i.e. $\eta_{ij} = \phi_{ij} + \theta_{ij}$, where θ_{ij} are independent Normal random effects and the columns of $\boldsymbol{\phi} = (\phi_{ij})$ follow adaptive ICAR distributions, all of them depending on a common weights matrix $\boldsymbol{W}^*(\boldsymbol{c})$ with a common weights vector $\boldsymbol{c}$, as described in the univariate case. Also, in parallel to the univariate case, the components of $\boldsymbol{c}$ will be assumed to follow a $Gamma(\alpha, \alpha)$ distribution. Different standard deviations would be considered for the columns of the $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ matrices in order to allow the relative risk geographical patterns to show more or less spatial dependence. The prior distribution for α will be chosen similarly to the univariate case above.

Note that the model just introduced, although posed in a multivariate setting, does not take into account the hypothetical dependence between diseases that these could show. In that case, that proposal could incorporate and take advantage of that dependence. In fact, we have explored that possibility by proposing a multivariate $\boldsymbol{M}$-model (Botella-Rocamora et al., 2015) with a common adaptive spatial weights matrix and therefore a common spatial dependence structure. We have not noticed any evident benefit of that proposal in terms of the estimation of the spatial weights matrix in comparison to the proposal above. Since our goal in this paper is focused on that estimation, we have preferred to pose the 'independent' multivariate version in order to keep its content simpler. In any event, if the main goal was to estimate the risks $\boldsymbol{R}$ for several diseases with an adaptive spatial dependence structure, the use of pure multivariate models, as the $\boldsymbol{M}$-model mentioned, would obviously yield a significant improvement.

6.4. Application

6.4.1. Spatial weights matrices estimation from multivariate data sets

In this section, we evaluate the performance of the multivariate adaptive spatial model described in the previous section in some real scenarios. The main data set for this analysis corresponds to the observed and expected deaths in the city of Valencia (Spain), for a total of 15 different mortality causes in men for the period from 1996 to 2015. Mortality data are available for each of the 531 census tracts of Valencia, the geographical unit for this analysis. The main goal of this analysis is to estimate a suitable weights matrix for the Valencian census tracts that reflect the dependence structure of mortality causes in general over the whole city. We will use the multivariate adaptive extension developed in Subsection 6.3.2 for both the BYM and Leroux models, and the mortality data set described to estimate that matrix. Subsequently, in the next subsection, the estimated spatial structure matrix will be used in posterior univariate analyses in order to assess the improvement that its use could bring, in comparison to the traditional adjacency criterion that assumes fixed weights, equal to 1, for each adjacent pair of units.

Both BYM and Leroux et al. multivariate models were run in `WinBUGS` and the corresponding `R` code for all the analyses in this section can be found as annex material to this paper in the Annex C, Section C.2. For each model, three chains were run with 200,000 iterations, whose first 50,000 iterations were used as a burn-in period. Of these, one of every 150 iterations was saved yielding a final sample size of 3,000 iterations. Convergence was assessed by means of the Brooks-Gelman-Rubin statistic (we required this to be lower than 1.1 for each variable in the model) and the effective sample size (required to be at least 100 for each variable in the model).

We start by taking a look to the estimated weights vector $\boldsymbol{c}$ for our analyses. These weights should reflect the strength of spatial dependence

between each neighboring pair of spatial units. For the BYM model, the values of the spatial weights $\boldsymbol{c}$ (their posterior means) range from 0.098 to 2.042, with a mean value of 1.240, while for the Leroux et al. model these values range from 0.027 to 1.886, with a mean value of 1.264. Figure 6.1 shows the corresponding c_i for each geographic unit for both models run. The census tracts with dark red color represent those with a lower estimated spatial weight c_i. According to the comments of Section 6.3, these census tracts should show a different behavior in comparison to their surrounding census tracts and, as a consequence, the model tries to separate/isolate those units. In contrast, the census tracts in yellow are those that have been found to have a greater influence on the risk of the surrounding areas or, in other words, those most dependent on their neighbors. As shown in this figure, both adaptive proposals of the BYM and Leroux models estimate a closely similar spatial dependence structure for the region of study. The correlation between the estimated spatial weights vector $\boldsymbol{c}$ for the adaptive BYM and Leroux models is 0.956, which shows the agreement of the spatial dependence structure estimated for both models.

We have explored the results in Figure 6.1 in order to interpret the spatial dependence structure estimated by the models. In particular, we have observed that the census tracts with lowest c_i values have certain peculiarities that make them special with respect to their adjacent units. On the one hand, residential homes for elderly or socially excluded people are frequently located in some of those "special" census tracts. As a consequence, these units often show higher observed deaths than expected for most of the mortality causes considered, which makes them exhibit a different behavior from those of their neighbors. Anyway, if our main goal was just to detect and model spatial units of this kind, with a different behaviour than their neighbors, the use of models accounting for discontinuities (Knorr-Held and Raßer, 2000; Denison and Holmes, 2001; Adin et al., 2019b) would be possibly a more suitable modeling choice. On the other hand, new building areas and socially marginal regions of the city also frequently show c_is in the darkest red zones. The

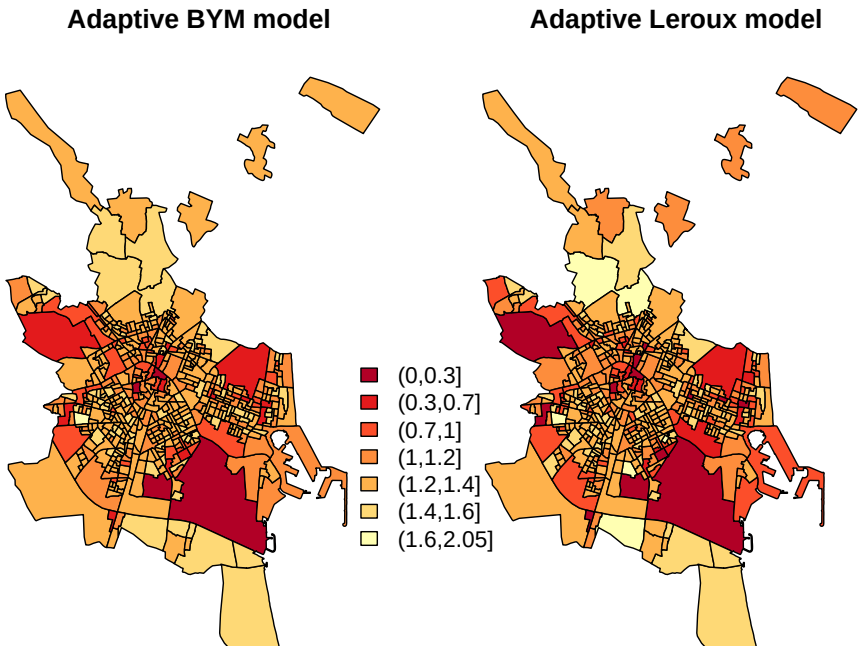


Figure 6.1.: Estimated spatial weight c_i for each census tract of Valencia according to all 15 diseases in the data set. Each choropleth map corresponds to either BYM (left) or Leroux et al. (right) models for the log-relative risks.

use of a broad set of mortality data, with 15 diseases, has allowed the models to identify those census tracts with these particularities that lead them to exhibit a very particular behavior in terms of mortality. That behavior requires an adaptation of the spatial weights matrix, otherwise their risks would be oversmoothed and estimated more similarly to their neighbors than they should. The low estimated value of their spatial weights will allow them to show the separate behavior that they require. As will be shown in the next subsection, a more flexible modeling of the risks is obtained in this manner, avoiding the excessive smoothing problems that many of the most frequent disease mapping models entail (Richardson et al., 2004; Botella-Rocamora et al., 2012). In contrast, Figure 6.1 shows some other regions where high spatial weights have

been fitted. Note that several of these units are located in spatial units at the borders of the graph where the geometry of the graph would impose lesser spatial dependence. See for example the yellow unit in the north of the city or those in the southeast of the city, which connect some other spatial units in its south which are not completely shown at the choropleth maps. The high values of the spatial weights vector seem to be used to connect more tightly those regions of the map in the outskirts that would otherwise have an excessively independent behavior, preventing them from being isolated. Thus, the adaptive proposal run seems to change some geometric properties of the graph that could make some census tracts less connected to the rest of the graph than would be desirable.

The two tables shown in the Annex C reinforce also these results. The first of these tables show, for each disease, the mean absolute difference between the risks of the adjacency and adaptive models. These results are separately presented for the regions with low, medium and high weights. These results point out that the main risk differences for both models occur there where the spatial weights vector takes more extreme values, more different to 1. Thus, this is where these two models particularly differ. On the other hand, the second of these tables show the risk differences between each area and its neighbours for the regions taking low and high weights, respectively. In average, the spatial units taking low weights show higher differences in comparison to their neighbours than the adjacency based model. Thus, these low weights allow these units to have a more independent performance. In contrast, the spatial units with high weights have risk estimates more similar to their neighbours than the corresponding adjacency based estimates. Therefore, the performance of these regions is just the opposite of that of the regions taking low spatial weights.

Similar conclusions are drawn from a parallel analysis that we have made for the whole of Spain at the municipal level (see more detailed results for this analysis in the supplementary material). In this case,

a multivariate analysis of 18 mortality causes has been carried out and a common spatial structure is estimated from this analysis. As a summary, 7 of the 10 municipalities with the lowest spatial weights (those which are disconnected) are municipalities in the Costa Blanca, a Spanish region with a large foreigner community of elderly retired people from northern Europe who have moved to this Spanish region. The presence of this community has been shown to have a clear impact on the mortality of this region; see Zurriaga et al. (2008). The three remaining municipalities with low spatial weights are all provincial capitals, which correspond to municipalities with substantially higher population than their neighbors. The spatial adaptive model used seems to have been sensitive to both data artifacts, making these municipalities different to their neighbors. On the contrary, we have found that 8 out of the 10 municipalities with the highest spatial weights belong to coastal municipalities, that is, they are placed at the borders of the graph of the region of study, which seems to confirm the impression that we have drawn from the Valencia city data set.

Before concluding the study of the estimated spatial weights vector, we want to illustrate the importance of the multivariate feature of the models implemented for that estimation. Figure 6.2 shows, once again for the BYM model in the Valencia city data set, the variability (standard deviation) of the estimated vector $\boldsymbol{c}$ (its posterior mean) as a function of the number of diseases considered. Thus, for one disease we have run 15 different models, one per disease, and we have repeated this for another 15 (randomly chosen) pairs of diseases, 15 trios and so forth until reaching groups of 14 diseases. The black line in Figure 6.2 connects the mean of the observed standard deviations of $\boldsymbol{c}$ for each number of diseases considered. The gray band delimits the minimum and maximum standard deviations observed for $\boldsymbol{c}$ for each number of diseases. Figure 6.2 shows how the multivariate model proposed describes an increasing trend for the variability of $\boldsymbol{c}$ as a function of the number of diseases considered. Thus, for the univariate studies, $\boldsymbol{c}$ hardly learns from the data, which means that the resulting spatial weights

matrix closely resembles the adjacency based weights matrix. In other words, the adaptive feature of the model has hardly any effect when a low number of diseases is considered. By way of contrast, Figure 6.2 shows substantial variability in $\boldsymbol{c}$ when the number of diseases is higher. Thus, in summary, Figure 6.2 clearly points out the need to perform multivariate studies if inference is pursued for the spatial dependence structure of a region of study.

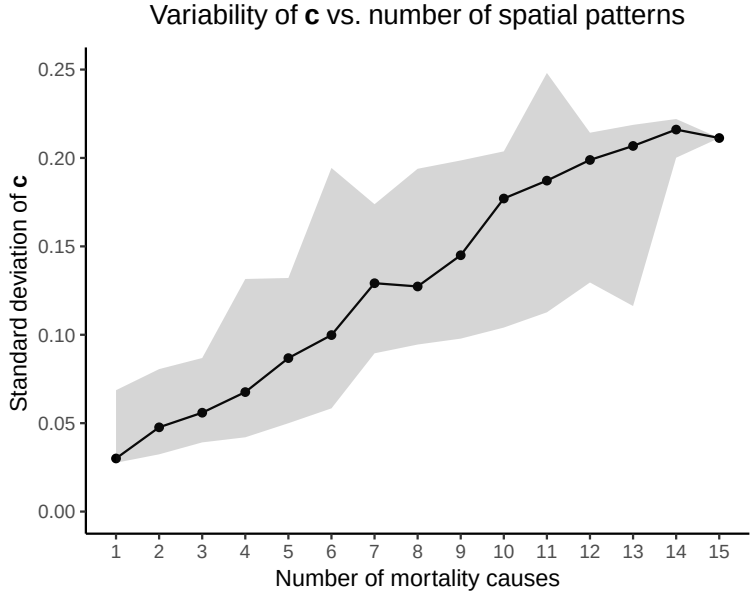


Figure 6.2.: Variability of $\boldsymbol{c}$ (standard deviation) in the Valencia city data set when estimated with the adaptive multivariate BYM model as a function of the number of diseases considered in the analysis. The black line connects the mean of the observed standard deviations of $\boldsymbol{c}$ and the gray band delimits the minimum and maximum observed standard deviations of $\boldsymbol{c}$ for each number of diseases.

6.4.2. Use of the estimated spatial weights matrix in subsequent univariate studies

Once the spatial weights matrix of the spatial random effects has been estimated for a region, it could be used for subsequent univariate disease mapping analyses on that same region. That is, the multivariate estimates above of $\boldsymbol{W}^*(\boldsymbol{c})$ could be used as estimates of the spatial dependence structure for later uses, instead of the traditional (although arbitrary) adjacency matrix $\boldsymbol{W}$. It is hoped that $\boldsymbol{W}^*(\boldsymbol{c})$ would have captured the geographical structure and particularities of mortality in that region of study. In this section, we are going to assess that procedure on our data set comparing the use of the $\boldsymbol{W}^*(\boldsymbol{c})$ matrices estimated in the previous subsection with the most traditional procedure which uses the adjacency criterion. Specifically, for all 15 diseases in our data set we have fitted univariate BYM and Leroux et al. models assuming either the spatial dependence structure estimated from the multivariate analysis above or the traditional adjacency-based weights matrix. Next, we compare the results of both analyses for each mortality cause according to the Standardized Mortality Ratios (SMR) of both alternatives and also according to the Conditional Predictive Ordinate (CPO) and the Deviance Information Criterion (DIC) proposed by Spiegelhalter et al. (2002).

In order to make a fair comparison, avoiding the use of the data twice (once for estimating $\boldsymbol{c}$ and once for estimating the SMRs with the corresponding univariate models), we have used different $\boldsymbol{W}^*(\boldsymbol{c})$ in our comparisons. Thus, for the case of AIDS mortality, for example, we have estimated $\boldsymbol{c}$ with a multivariate study of 14 diseases, all excepting AIDS, which avoids using the data twice for the univariate (preestimated) adaptive analyses. We have repeated this procedure for all 15 causes of mortality considered. Interestingly, the correlations between the spatial weights vectors $\boldsymbol{c}$ for the analyses with 14 mortality causes and that with 15 mortality causes are rather high, ranging from 0.94 to 0.97 depending on the cause removed. As a consequence, we would not

expect important differences if $\boldsymbol{W}^*(\boldsymbol{c})$ had been estimated just once with all 15 diseases.

Figure 6.3 shows the estimated Standardized Mortality Ratios (SMR) with the BYM model for the Valencian census tracts for cirrhosis mortality (similar choropleth maps for the remaining of diseases are shown in the supplementary material of this paper). The map on the left side corresponds to the model assuming an adjacency-based relationship between spatial units, while that on the right side uses the spatial weights matrix previously estimated. The Leroux et al. model provides similar results and these are also shown in the supplementary material of this paper.

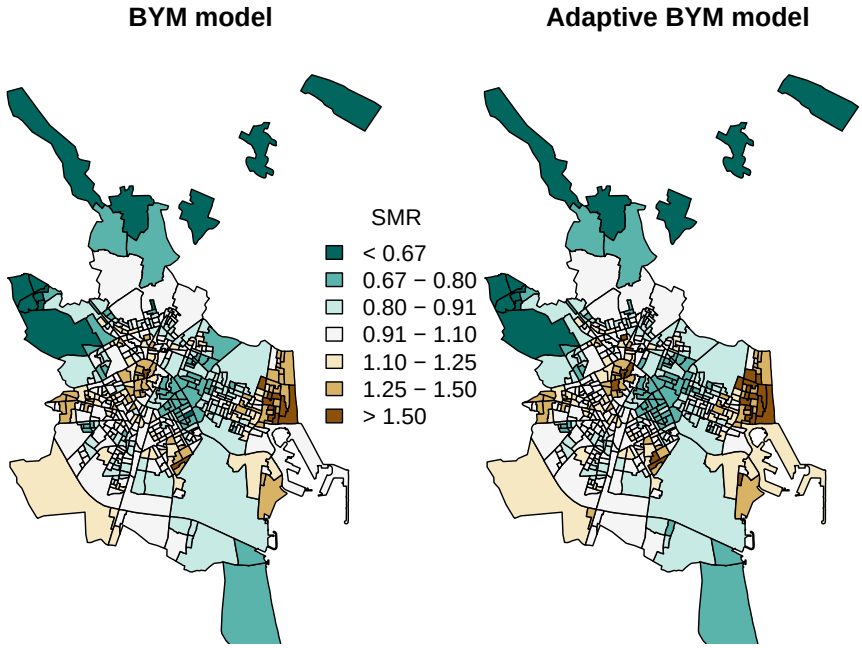


Figure 6.3.: Standardized Mortality Ratios for Cirrhosis in Valencia estimated with the BYM model and with spatial weights matrices of either unitary weights (left) or using the values obtained from the multivariate analysis of 14 diseases (all mortality causes of study except Cirrhosis) (right).

As can be seen, both models provide risk maps with similar spatial patterns. However, the model using the adaptive spatial weights reproduces higher variability than its adjacency-based alternative allowing some of Valencia's neighborhoods to be reproduced more clearly and making it possible for some census tracts to reproduce more extreme risks. In particular, the adaptive analysis depicts more clearly some particular high risk zones scattered throughout the city which usually correspond to high deprivation areas. Thus, as previously pointed out, the adaptive weights avoid the excessive smoothing of the SMRs previously described in the literature by allowing additional flexibility wherever it is required according to previous information synthesized in the previous multivariate adaptive analysis.

Afterwards, we have compared the fit of the adaptive vs. the non-adaptive weights models according to the CPO and DIC criterion. Table 6.1 shows the DIC and CPO of the BYM and Leroux models for both spatial weights matrices considered. For each model and mortality cause we have marked the proposal providing a better fit, according to DIC and CPO, in bold. As can be observed, according to DIC (CPO), BYM and Leroux et al. models with adaptive weight matrices provide a better fit than the corresponding adjacency based model in 14 (13) and 13 (9), respectively, out of the 15 mortality causes considered. This confirms that the greater flexibility of the adaptive models really improves the SMRs estimates in comparison to the traditional adjacency-based analyses. In addition, as it can be confirmed at the supplementary material, those mortality causes with a more substantial improvement in terms of DIC or CPO in general coincide with those showing a spatial pattern of higher variability (AIDS or COPD for example). Thus, an improvement is achieved mainly when there is spatial variability to be explained, otherwise the gain achieved is very modest as might seem logical.

Table 6.1.: DIC and CPO for the BYM and Leroux at al. models with adaptive and unweighted spatial weights matrices.

Causes		Adjacency BYM model	Adaptive BYM model	Adjacency Leroux model	Adaptive Leroux model
AIDS	DIC	1648.11	**1631.87**	1653.57	**1647.39**
	CPO	-857.98	**-842.16**	-865.67	**-853.49**
Stomach cancer	DIC	1771.57	**1771.43**	1770.46	**1770.30**
	CPO	-884.84	**-884.77**	-884.56	**-884.49**
Colorectal cancer	DIC	2354.53	**2354.42**	2354.07	**2353.52**
	CPO	**-1176.82**	-1176.95	-1176.69	**-1176.65**
Lung cancer	DIC	2861.52	**2857.85**	2872.61	**2872.34**
	CPO	-1430.85	**-1429.09**	**-1436.90**	-1438.01
Prostate cancer	DIC	2126.24	**2126.08**	2124.63	**2124.51**
	CPO	-1062.46	**-1062.34**	-1061.91	**-1061.83**
Bladder cancer	DIC	**2051.48**	2052.19	2056.07	**2053.80**
	CPO	**-1027.02**	-1027.55	-1029.16	**-1027.80**
Hematological cancer	DIC	1955.37	**1955.17**	1953.59	**1953.26**
	CPO	-977.43	**-977.34**	-976.51	**-976.35**
Mellitus diabetes	DIC	1976.97	**1975.01**	1978.35	**1976.38**
	CPO	-987.946	**-987.03**	-988.85	**-987.54**
Dementia	DIC	2335.17	**2333.21**	**2341.74**	2342.03
	CPO	-1168.47	**-1168.05**	**-1172.21**	-1172.41
Ischemic heart disease	DIC	3055.33	**3048.46**	3061.86	**3061.53**
	CPO	-1537.51	**-1535.37**	**-1541.93**	-1542.05
Ictus	DIC	2662.95	**2659.86**	2665.66	**2664.25**
	CPO	-1331.24	**-1329.64**	-1332.84	**-1332.18**
COPD	DIC	2681.81	**2668.87**	2698.39	**2679.29**
	CPO	-1348.43	**-1342.04**	-1359.55	**-1347.98**
Liver cirrhosis	DIC	2130.23	**2128.75**	**2139.20**	2141.70
	CPO	-1071.21	**-1070.22**	**-1075.66**	-1077.16
Suicides	DIC	1488.94	**1488.86**	1488.43	**1488.36**
	CPO	-744.36	**-744.33**	**-744.09**	-744.25
Traffic accidents	DIC	1506.68	**1506.17**	1506.76	**1506.75**
	CPO	-752.97	**-752.80**	**-753.13**	-753.22

6.5. Discussion

As described in this paper, CAR distributions are usually considered to model the spatial dependence between geographic units in disease mapping studies. Although CAR distributions are undoubtedly useful and powerful tools, the parameterization used to induce dependence by means of its structure matrix is usually arbitrary. We have introduced a procedure to estimate that spatial weights matrix according to retrospective multivariate data. As shown, our adaptive procedure makes CAR models more flexible and improves the fit of subsequent analysis adopting the estimated weights matrix, which in principle should have captured the particularities that mortality data could show in that region. Additionally, the multivariate character of our proposal has shown itself to be an indispensable tool for appropriately estimating the spatial structure of the data.

The methodology introduced could have several different uses. First, the multivariate adaptive model introduced could be used in multivariate studies with adaptive spatial structures. These models should provide more accurate risk estimates that could take advantage of the adaptive character of the spatial dependence considered. In any event, if that was the main goal of our analysis, a multivariate model, considering the dependence between mortality causes, would be much more advisable. A second use of adaptive CAR models would be the one emphasized in this paper, that is, making inference on the spatial weights matrix of a region of study. In this case, we would be more interested in the values of the weights $\boldsymbol{c}$ than the own risks. As a consequence, that vector $\boldsymbol{c}$, and thus the adaptive weights matrix could be later used in subsequent enhanced spatial disease mapping studies with a non-arbitrary spatial structure based on previous data and knowledge.

In this sense, we find it convenient to mention a couple of limitations of the proposed methodology. Our adaptive model proposes a kind of meta-analysis of the spatial structure of several causes of death. It would be obviously convenient that these causes of death were as

homogenous as possible. In an ideal situation, all of them should be cardiovascular diseases, or tumoural causes of death ... since these settings should probably share a common spatial dependence structure, as assumed by our model. Considering congenital deaths or senility, for example, as causes of death with a common spatial structure could be possibly a bit risky since that assumption would be hard to maintain in that case. Although we do not find any reason why adjacency could be a better option in this setting. Anyway, this limitation should be born in mind when using the estimated spatial weights matrix in new studies, since the new causes of death in those cases should be as related as possible to those used for estimating the spatial weights matrix.

In the same manner, as suggested by one reviewer, it would be convenient to bear in mind that adding covariates to disease mapping problems could possibly change the spatial dependence structure of the region of study. For example, if an adaptive spatial dependence structure gives a low weight to a particular spatial unit, separating it from its neighbours, this could be also reproduced by a covariate taking in this spatial unit very different values than in its neighbours. Thus, a weights matrix that could be suitable for disease mapping studies for a region of study could be not so good for ecological regression studies on that same region.

Although in principle the main uses of our model would be those mentioned in the previous paragraphs, we have also found a third practical use of the model that we did not expect. This use would be quality control of systematic problems that could be present in health data sets. Being more precise, the Valencia city mortality data used in Section 6.4 belongs to a large Spanish project studying mortality in large cities, the MEDEA project. All the deaths in that data set have been geocodified by using several geocoding tools, in particular the Google geocoding API and a second geocoding tool (Cartociudad) of the Spanish Geographic National Institute. These tools, as with any other geocoding tool, are not perfect and they have errors for some

particular streets, groups of cases that are geocodified in the city center etc. that could distort the spatial analyses of that data base. We have found that the multivariate adaptive model on those data bases give low spatial weights to those census tracts with systematic geocoding errors since their mortality data are somewhat different from their surrounding areas. This has allowed us to distill those errors by focusing on those census tracts with low spatial weights and no potential alternative explanation (no residential homes, no socially marginal areas, no new building areas, etc.) for them. In most cases we have found that those regions contained some geocoding error. Note that the results shown in Section 6.4 correspond to the distilled database without geocoding errors, which have been already fixed otherwise Figure 6.1 would have also pointed out the census tracts with geocoding errors. This is just a single example of the many uses that adaptive CAR models could have in practice. This work illustrates just some potential uses of adaptive CAR models, although we suspect there are many more than those that we have found. We encourage readers to keep exploring new potential uses of this approach.

Acknowledgements

The autors acknowledge the support of the research Grant PI16/01004 (co-funded with FEDER grants) of Instituto de Salud Carlos III and the predoctoral contract UGP-15-156 of FISABIO.

Conflict of interest

The authors declare that they have no conflict of interest.

7. The Spanish National Atlas of Mortality (ANDEES)

In this chapter, we introduce the Spanish National Atlas of Mortality (ANDEES). This chapter is divided into 4 sections. Section 7.1 introduces and motivates the work that we have done in this atlas. Section 7.2 describes the data, methodology followed and sketches some of its main features. Section 7.3 shows a sample of the more interesting results, in our opinion, that can be consulted in ANDEES. Finally, Section 7.4 discusses some conclusions from our experience developing ANDEES.

7.1. Introduction

The geographical distribution of mortality has been the object of interest in many epidemiological studies. Just in Spain, small areas mortality studies have been carried out for many of the regions into which Spain is divided (Benach et al., 2004; Martínez-Beneito et al., 2005; Esnaola et al., 2010; Ocaña et al., 2010) or they have paid particular attention to the distribution of mortality within large cities (Borrell et al., 2009, 2010; Puigpinos-Riera et al., 2011; Aguilar-Palacio et al., 2017). While these studies do have their own interest, they show just a part of the picture for the whole of Spain when, evidently, the risk in any of these regions is not independent of the risks in the surrounding regions. Moreover, if the region of study in a spatial analysis is too small to capture the

geographical variability of the disease, the geographical distribution of risk could seem flat, when it would emerge if the region of analysis was larger. Accordingly, we could be missing an opportunity to capture that variability and thereby take advantage of its knowledge it contains, since the underlying risk factors that could be causing that variability could remain unknown. As a consequence, a small areas mortality study at the national level could be quite interesting to undertake, mainly if it is comprehensive enough to provide a general view of mortality for a variety of causes throughout the whole country.

Some initiatives have already been undertaken in order to explore the geographical distribution of mortality for the whole of Spain (Benach et al., 2001; López-Abente et al., 2006) at a small area (municipal or small municipality aggregates) level. Despite having some interesting aspects, these works are already a bit outdated, in terms of the period of study (1987-1995 in Benach et al. (2001) and 1989-1998 in López-Abente et al. (2006)). Moreover, these two works were originally published as traditional printed books. Obviously, the richness of municipal mortality maps gets a bit constrained by this publishing format which does not allow such a large amount of results to be explored in close detail. Fortunately, new statistical dissemination tools such as `Shiny` (Chang et al., 2020) or `Tableau` have emerged in recent years, making it possible to consider new publishing formats for studies of this kind. These new formats permit the use of more modern tools, such as interactive maps, charts and tables facilitating a more effective dissemination of the results of these studies. Additionally, `Shiny` (or some other similar tools) enable the web publishing of the mortality projects that could be developed, making this kind of analysis accessible even to the general public, in contrast to the hardcopy print format of the prior mortality atlases already published. Finally, the web publishing of the broad mortality studies that we are talking about, would make it possible to include a far higher number of causes of death. Mortality studies with hundreds, or even thousands, of deaths are not feasible in paper format, but web applications are able to handle such an amount of

information easily and in a user-friendly way. All of this supports the idea of using updated mortality data and these modern statistical tools to develop comprehensive small areas mortality studies at the national level. This study could be of great interest for both epidemiologists (or health professionals in general) and the general public.

ANDEES is an interactive web application that allows for the visualization of the spatial and spatio-temporal distribution of mortality throughout the whole of Spain during the period 1989-2014. The spatial unit of analysis in ANDEES is the municipality, specifically a total of 8,063 municipalities throughout Spain over the period of analysis. The annual average population of these municipalities ranges from 5 to 3,061,610 inhabitants during the period of study, thus risk-smoothing statistical models are absolutely required in order to draw minimally reliable results. Independent geographical patterns are estimated in ANDEES for men, women and for both sexes together; accordingly, three spatial patterns are estimated for each cause of death, obviously whenever this makes sense, since prostate cancer, for example, is not considered in women or both sexes for evident reasons. ANDEES considers a total of 102 causes of death, which correspond to all the mortality groups considered by the Spanish National Statistical Institute (INE).

Given the large volume of information generated in ANDEES, the final design of this mortality atlas has been developed as a web application. In this manner, we avoid the mentioned drawbacks of traditional publishing formats. The final version of ANDEES can be consulted at the following URL:

http://atlasnacional.fisabio.es.

ANDEES has been developed by the *Bayensians* research group of the FISABIO Foundation and the Valencian General Directorate of Public Health. The authors of ANDEES are, in this order: Francisca Corpas Burgos, Carlos Vergara Hernández, Paloma Botella Rocamora,

Jordi Pérez Panadés, Hèctor Perpiñán Fabuel and Miguel Ángel Martínez Beneito. ANDEES has been partially funded by: UGP-15-156 grant from FISABIO, PI16/01004 grant from Instituto de Salud Carlos III and MTM2013-42323-P grant from the Ministerio de Economía y Competitividad.

7.2. Methodology

7.2.1. Data

Two data sets have been used as the main sources for the development of this atlas. The first of them contains all the deaths that occurred in Spain during the period 1989-2014. For each of these deaths, we have the following individual information: sex, age, INE code of the municipality of residence, year and cause of death. These data have been provided by the INE and tabulated according to sex, five-year age group (considering a final age group of 85 years or more), municipality and cause of death, for the total period 1989-2014 and considering eight triennial periods, from 1991-1993 to 2012-2014. The causes of death studied correspond to each of the 102 causes of death set by the INE in its *Abbreviated List*. This full list, with the CIE9 and CIE10 codifications corresponding to each of these causes of death, can be found at: https://www.ine.es/daco/daco42/sanitarias/lista_reducida_CIE10.pdf.

Of the previous groups of death, only those sexes or the combination of both with at least 10,000 deaths during the entire study period have been finally studied, since the causes with a lower number of deaths could yield unreliable results. In this manner, we guarantee at least one observed death, on average, per municipality during the period of study, which seems a reasonably safe limit over which the smoothing methods used should yield reliable results.

The second main data set for ANDEES contains information on the population (number of people at risk) for each region and year of

study. This second database is tabulated according to the following variables: sex, age (five-year groups), INE code for each municipality and year. This information has been obtained from the municipal population register published annually by the INE. These data are available annually at the INE website starting in 1998 (https://www.ine.es/dynt3/inebase/es/index.htm?padre=517&capsel=525). Additionally, the INE website also has this information available for 1996. The population for 1997 has been estimated by geometric interpolation of the populations published for 1996 and 1998. For the years between 1989 and 1995, the municipal populations have also been estimated by geometric interpolation of the population data from the 1991 and 2001 censuses, since population register data were not available before 1996. Both censuses were previously calibrated to the population registries data sources comparing the 2001 population gap for both sources, thereby correcting the census data in order to adapt it to the other source. This should eliminate the bias that could exist between both sources for Spain as a whole; that same correction would be also applied to municipal data. The same correction was also applied to the 1991 census data and therefore it would be taken into account for the interpolated interim data. Population data for each age group, sex, year and municipality have been used to calculate the number of expected deaths per municipality and period, which will be used later for disease mapping models.

Spain was divided into 8,119 municipalities in 2014; however, not all of them have been considered as units of analysis in ANDEES. This is because some of these municipalities were created after 1989, so they have not existed throughout the whole period of study. As a consequence, no observed and expected cases would be available for all the years of the period of study for these municipalities, which could lead to analysis problems mainly for the spatio-temporal study. As a solution, we have decided to merge these municipalities into the municipality that they come from, and consider them as a single unit of study, summing their observed and expected cases for all the years

of analysis. In the end, a total of 8,063 municipalities (or aggregates in some cases) have been considered for the analyses of ANDEES. The cartographic information used, also downloaded from the INE website (https://www.ine.es/censos2011_datos/cen11_datos_resultados_seccen.htm), corresponds to the Spanish municipalities for 2011 (latest update available) and has been suitably modified taking into account the aggregations of municipalities that we have just commented.

7.2.2. Spatial and spatio-temporal modeling of mortality risks

The mortality risk indicators studied in ANDEES have been estimated from the observed and expected deaths in each of the municipalities of Spain, for the whole period of analysis 1989-2014 and for the eight triennial periods that make up the whole study period. As for the rest of the works carried out in this thesis, smoothed SMR have been the main mortality indicator used to study the mortality risks for each municipality. Smoothed SMRs have been estimated by means of spatial and spatio-temporal smoothing models. On the one hand, the spatial smoothing model used to smooth the SMRs for the whole period of study has been the one proposed by Besag et al. (1991) (BYM), already described in detail in the previous chapters of this thesis. Thus, we will not introduce this model once again in this chapter. On the other hand, due to the great length of the study period, it is also interesting to study mortality over shorter periods of time and analyze its evolution over time. To do so, the observed and expected deaths for each of the municipalities of Spain, disaggregated in the eight triennial periods already mentioned, have been considered. In this case, mortality throughout Spain during the whole period of study is used as the reference population for deriving the expected deaths for all the subperiods and municipalities. The use of spatio-temporal smoothing models allows the study to be disaggregated into several study subperiods, providing updated estimates of the risks, instead of

global estimates summarizing the entire period. In this way, we avoid the bias that occurs when considering risks as static amounts over time, which show temporal variations (Ocaña Riola, 2007).

Within the spatio-temporal modeling literature, the Martinez-Beneito et al. (2008) proposal has been particularly well received and has been repeatedly used to explore the evolution of mortality in different studies (Zurriaga et al., 2008, 2010; Gracia et al., 2017; Marco et al., 2017; Morris et al., 2019). The spatio-temporal smoothing of the SMRs in ANDEES has been carried out according to this proposal. The spatio-temporal model proposed by Martinez-Beneito et al. (2008) considers the joint use of spatial smoothing tools and time series models, specifically auto-regressive processes. That is, in this *auto-regressive model*, a spatio-temporal structure is defined in which the risks are spatially and temporally dependent at the same time, favouring SMRs that correspond to either nearby locations or consecutive time periods to take similar values.

Specifically, the auto-regressive spatio-temporal model assumes that the observed deaths in each municipality and time period O_{it} follow a Poisson distribution:

$$O_{it} \sim Poisson(E_{it}R_{it}), \quad i = 1, \ldots, I, \quad t = 1, \ldots, T,$$

where E_{it} and R_{it} are the expected number of deaths and the relative risk in each geographic unit and time period under study, respectively. For the first time period, the logarithm of the relative risks is defined as the sum of an intercept, a spatial random effect, and a heterogeneous random effect as follows:

$$\begin{aligned} &log(R_{i1}) = (\mu + \alpha_1) + \left(1 - \rho^2\right)^{-1/2} \cdot (\phi_{i1} + \theta_{i1}), \\ &\boldsymbol{\phi}_1 = (\phi_{11}, \ldots, \phi_{I1}) \sim ICAR(\sigma_\phi^2), \\ &\theta_{i1} \sim \mathcal{N}(0, \sigma_\theta^2), \quad i = 1, \ldots, I. \end{aligned} \tag{7.1}$$

As in the BYM model, the spatial random effect is considered to follow

an ICAR distribution and an additional heterogeneous random effect of constant variance is also considered. The parameter ρ corresponds to the temporal correlation, which controls the strength of the smoothed SMRs temporal dependence for each municipality. The intercept for these log-risks is decomposed into two terms: μ, which is the average value of the log-risks for all the geographic units and periods, and α_1, which is the mean deviation from the average level that occurs for the smoothed SMRs for the first subperiod of study.

For the following subperiods, the logarithm of the relative risks are defined as follows:

$$\begin{aligned} &log(R_{it}) = (\mu + \alpha_t) + \rho \cdot (log(R_{i(t-1)}) - \mu - \alpha_{t-1}) + \phi_{it} + \theta_{it}, \\ &\boldsymbol{\phi}_t = (\phi_{1t}, \ldots, \phi_{It}) \sim ICAR(\sigma^2_\phi), \\ &\theta_{it} \sim \mathcal{N}(0, \sigma^2_\theta), \quad i = 1, \ldots, I, \quad t = 2 \ldots, T. \end{aligned} \tag{7.2}$$

In this way, the risk in each geographic unit and period not only depends on the risk in its neighboring units, but also depends on its own risk in previous periods. That temporal dependence is defined by means of a first order autoregressive time series, while geographical dependence is induced by including spatial random effects to model the temporal evolution between consecutive periods. Thus, neighboring areas have similar risk evolutions in the same manner that they have similar geographical risk estimates.

Additionally, the vector of period-specific intercepts

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_T) \sim ICAR(\sigma^2_\alpha),$$

is assumed to also follow an ICAR distribution, considering the consecutive time periods as neighbors. This induces a smooth evolution in the average smoothed SMRs over all the time periods considered, which seems a reasonable assumption from an epidemiological point of view.

Finally, we complete our model with the following prior distributions:

$$\mu \sim \mathcal{U}(-\infty, \infty)$$

$$\rho \sim \mathcal{U}(-1, 1)$$

$$\sigma_\phi, \sigma_\theta, \sigma_\alpha \sim U(0, 100).$$

These prior distributions for the overall intercept (μ) and, for the standard deviation of the random effects (σ_ϕ, σ_θ and σ_α) are intended to be vague. The prior distribution of the time correlation parameter (ρ) has been chosen in order to reproduce a stationary time series. In equation (7.1), the term $(1 - \rho^2)^{-1/2}$ is introduced so that the variance-covariance matrix of $O_{\cdot 1}$ coincides with that same matrix for the rest of the periods, which is the stationary covariance matrix of the multivariate series $(O_{\cdot t})_{t=2}^{\infty}$. The expression of the overall variance-covariance matrix for all the smoothed SMRs can be found in Martinez-Beneito et al. (2008), which is just a Kronecker product of the spatial and temporal dependence structures.

The models described above have been used in the studies of mortality data for continental Spain. For the rest of Spain (islands and autonomous cities), alternative models have been run in which there is spatial dependence only for the municipalities on the island of Mallorca. For the rest of the islands and autonomous cities, only heterogeneous random effects have been considered, since the geographical extent of these regions means the spatial dependence hypothesis makes little sense. Moreover, avoiding spatial random effects in those regions also avoids the particular problems that ICAR distributions exhibit when used on disjointed sets of regions (Hodges et al., 2003). In any event, for each sex and cause of death, three different intercepts were used for the Canary Islands, Balearic Islands and Ceuta-Melilla, respectively, since we assumed these regions to be distant and different enough as to need these different parameters.

The mentioned spatial and spatio-temporal models have been run in

WinBUGS (Lunn et al., 2000), using the statistical package R with some particular libraries. The enormous volume of data to be analyzed as well as the number of estimates to be obtained for each analysis have been a clear computational challenge. Thus, some tools have been necessary to speed up the computing. Specifically, we have made use of the Pbugs package of R (https://github.com/fisabio/pbugs), which has allowed us to automate the calls to WinBUGS from R, running in parallel each of the different chains involved in the MCMC process, reducing computation times by a factor of the number of chains run. In the case of the spatial analysis, three chains were run for each model with 11,000 iterations, where the first 10% of these were discarded as a burn-in period. Of these, one out of every 29 iterations was saved, thereby yielding a final sample size of 1,026 iterations. In the case of the spatio-temporal analysis, five chains were run for each model with 15,000 iterations. The first 5,000 iterations of each chain were discarded in this case as a burn-in period. Of these, one out of every 50 iterations was saved, thereby yielding a final sample size of 1,000 iterations. Convergence was assessed by means of the Brooks-Gelman-Rubin statistics (we required this to be lower than 1.1 for each variable) and the effective sample size (required to be at least 100 for each retrieved variable) (Brooks and Gelman, 1998; Gelman et al., 2014). These statistics are implemented in the R2WinBUGS package (Sturtz et al., 2005) of R, that the Pbugs library makes use of.

Once the posterior distributions of the smoothed SMRs were estimated from the BYM model and the spatio-temporal auto-regressive proposal, we calculated their posterior means and their mass above 1 (risk excess probability) for each municipality and period (in the case of spatio-temporal modeling). These values are the two main statistics finally shown in the maps drawn in ANDEES.

In addition to these statistics, and to facilitate the visualization and understanding of the results of the spatio-temporal analysis, ANDEES allows the estimated smoothed SMRs from that analysis

to be broken down into separate components: spatial, temporal and spatio-temporal. ANDEES also permits each of these components to be reproduced by separating or removing its effect over the spatio-temporal smoothed SMRs for each municipality and subperiod. The mentioned spatial component of the spatio-temporal SMRs would be the average geographic pattern of those SMRs over all the subperiods that make up the global period of study. This component would be quite similar to the SMR maps represented in the spatial analysis. The temporal component of the SMRs would represent the mean temporal evolution of the SMRs throughout the study period, for the whole period of study. Finally, the mentioned spatio-temporal component would represent the changes produced in the smoothed SMRs for each subperiod and municipality, beyond the sum of the spatial and temporal components already described. Specifically, as described in Adin et al. (2017) and Martinez-Beneito and Botella Rocamora (2019), if $\mu^* = \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=1}^{T}\log(R_{it})$ is the average log-risk for all the spatial units and periods, the estimated log-risks for each municipality i and subperiod t can be decomposed, respectively, into the following spatial, temporal and spatio-temporal components:

$$\begin{aligned}
\xi_i^* &= \frac{1}{T}\sum_{t=1}^{T}\log(R_{it}) - \mu^*, \\
\gamma_t^* &= \frac{1}{I}\sum_{i=1}^{I}\log(R_{it}) - \mu^*, \\
\delta_{it}^* &= \log(R_{it}) - \xi_i^* - \gamma_t^* - \mu^*.
\end{aligned}$$

It can be easily checked (see for example the references above) that the estimated log-risks can be decomposed as the sum of these patterns, i.e.:

$$\log(R_{it}) = \mu^* + \xi_i^* + \gamma_t^* + \delta_{it}^*.$$

As mentioned, ANDEES allows the representation of different maps showing the SMRs and the probabilities of excess risk, including/excluding the different components that make up the log-risks

decomposition above. Thus, for example, if the SMR maps in a spatio-temporal model showed a strong temporal component, making the risks high in the first period, in general, and low in the last period, it might be wise to remove the temporal effect in both representations. In this manner, it would be possible to observe the spatial pattern for each period beyond the strong temporal trend mentioned. Once the temporal component was removed, we could find a strong spatial component underlying all the resulting maps. In that case, it could perhaps be useful to remove the spatial component common to all these maps in order to see more specific risk variations of more specific interest for some particular locations and periods. This illustrates how playing with these components may make it possible to visualize some particular features that would otherwise remain unnoticed. The risk excess probabilities maps for this representation are sensitive to the components of the SMRs considered, thus if the temporal component is removed, the risk excess probability of any observation would really measure $P(\xi_i^* + \delta_{it}^* > 0)$ for the corresponding municipality.

7.2.3. Some further non-statistical details of ANDEES

Beyond the data modelling component of ANDEES, this project also has an important data visualization component that we are going to describe now. The web application that hosts and allows the results of the above models to be visualized has been developed using the `Shiny` package of `R`. Nowadays, this package is becoming nowadays very popular and several applications for spatial and spatio-temporal data analysis and visualization have already used it (Moraga, 2017; Adin et al., 2019a). ANDEES enables user interaction through several control widgets and creates interactive visualizations of the data and results according to those controls. Specifically, ANDEES allows different selection and visualization criteria for the maps to be set, showing the spatial and spatio-temporal distribution of risks. These selection criteria permit the specification of the sex, cause of death and study

period (in the case of spatio-temporal analyses) of the map to be represented. The visualization criteria enable the graphical output of the results to be configured, for example, by specifying the indicator that will be represented on the maps (smoothed SMRs or risk excess probabilities), establishing the cut-off points or color palettes for the choropleth maps, etc. Additionally, as already mentioned, in order to facilitate the interpretation of the spatio-temporal analyses, it is also possible to select which components of the SMRs (spatial, temporal or spatio-temporal) will be included in the corresponding map.

The main exploration tools implemented in ANDEES are:

- Choropleth maps with the smoothed SMRs and risk excess probabilities for the different causes of death, sexes and periods of study (in the spatio-temporal case). These maps support interactive panning and zooming, which is very convenient for exploring particular regions in detail. In addition, when clicking on a municipality, specific information shows up with its name, province and the corresponding smoothed SMR and risk excess probability estimates.

- Line plots showing the temporal evolution by subperiods of the average risk for each province and the whole country, for the selected sex and cause of death. These plots also include support for interactive features such as panning, zooming and series highlighting.

- Data tables containing numeric information on the estimates of interest. These tables support filtering, pagination and sorting which would be helpful for locating the information from one particular municipality or identifying those municipalities with highest or lowest risk estimates.

The `Leaflet` package (Cheng et al., 2019) has been used for building the choropleth maps of ANDEES. This package allows interactivity to be added to these maps, as well as to represent them over underlying

cartography layers, which makes it possible to place each municipality in its geographical context. Some of the `Leaflet` functions used have had to be optimized and recoded in order to speed up the rendering of the maps, as it seems that these functions were not designed to handle such an amount of data. These improvements have reduced the rendering time of each map in the atlas from about 8 seconds to just 2 seconds, approximately. Line plots with the risks evolution for the different subperiods of the study have been drawn by using the `Plotly` package (Sievert et al., 2020). Finally, data tables for displaying the estimates of interest are shown with the help of the `DT` package (Xie et al., 2020), which allows interactive features such as ordering by the different columns of the table.

In addition to these results, ANDEES also provides further information that may be of interest for exploring the results in greater detail. These additional tools contain, among others:

- Dispersion plots for the smoothed municipal SMRs against some municipal features, such as the number of inhabitants, municipal average income, longitude or latitude. These plots make it possible to address the potential relationship between these factors and the risks of death.

- Density plots for the smoothed SMRs. This yields deeper insights into the shape of that distribution, which choropleth maps hardly allow knowing about. Moreover, these plots allow the length of the tails of that distribution to be known, which is obviously of interest from an epidemiologic point of view.

- Age-specific mortality rates per 100,000 inhabitants, for each sex separately and for both together. This information makes it possible to know the age distribution of each mortality cause and compare it between sexes.

All the results shown in ANDEES can be downloaded by users as image files, in the case of maps and figures, and csv/Excel files in the case of data tables.

7.3. Results: Some interesting mortality geographic patterns

In this section we provide a brief description of some of the results shown in ANDEES. Specifically, we show the estimated geographic mortality patterns for all causes of death jointly and for some specific causes, separately for men and women. Likewise, we also present the changes in the mortality risks at the provincial and national levels during the period of the study, which allows us to identify areas with a different behavior in comparison to the general trend. Our goal is not to make an exhaustive interpretation of the maps and figures shown, but to present an overview with some of the results obtained. In this way, we attempt to illustrate the potential use of ANDEES for other health professionals with a particular interest in some specific causes of death, and the ways in which we would like people to make use of this tool.

7.3.1. All-causes mortality

Figure 7.1 shows the all-causes mortality risk maps for the whole period 1989-2014 and the evolution of such risks at the provincial and national levels in the different subperiods of study for both men and women, respectively. As can be observed (Figures 7.1a and 7.1b), all-causes mortality maps show clear inequalities, mainly marked by a north-south pattern. In the case of men (Figure 7.1a), the highest mortality occurs especially in the southwestern half of the peninsula (Extremadura and western Andalucía). In contrast, the areas with the lowest mortality are found in the north and north-east side of Spain (northern Meseta and

below the Pyrenees). In the case of women (Figure 7.1b), the highest mortality occurs in the southern half of the peninsula, especially in the southern and western part of Andalucía (Huelva, Sevilla, Cádiz and Málaga). On the other hand, the areas with the lowest mortality are once again located in the north and north-east of Spain. Overall, both two maps are quite similar in terms of evidencing similar geographical patterns for both sexes. This higher mortality for all causes in men and women in the southern areas of the country compared to mortality in the northern zone could be reflecting, in part, the existing socioeconomic inequalities between these regions (Benach and Yasui, 1999).

The general evolution of all-causes mortality at the provincial and national levels, in men (Figure 7.1c) and women (Figure 7.1d), shows a clear downward trend throughout the period of study. This overall downtrend in general mortality throughout the whole country could be explained by the public health and sanitary improvements that have occurred over the years of this long period of study. The spatio-temporal results show how, in general, the north-south geographic pattern found is maintained in all the subperiods of the study, and areas with an evolution different from the general one are not found. In other words, the spatio-temporal interaction is quite mild for all-causes mortality. Therefore, we do not include the maps for each one of the subperiods of analysis, since they are very similar (leaving apart the effect of the overall trend for the whole of Spain).

7.3.2. Malignant tumor of the trachea, bronchi and lung mortality

Figure 7.2 shows the mortality risk maps for malignant tumors of the trachea, bronchi and lung (simply lung cancer from now on) for the period 1989-2014 and its evolution at the provincial and national levels for both men and women. Of the 102 causes of death studied, lung cancer represents the first cause of death in men, accounting for 8.3% of all deaths. In contrast, lung cancer in women is the twenty-first

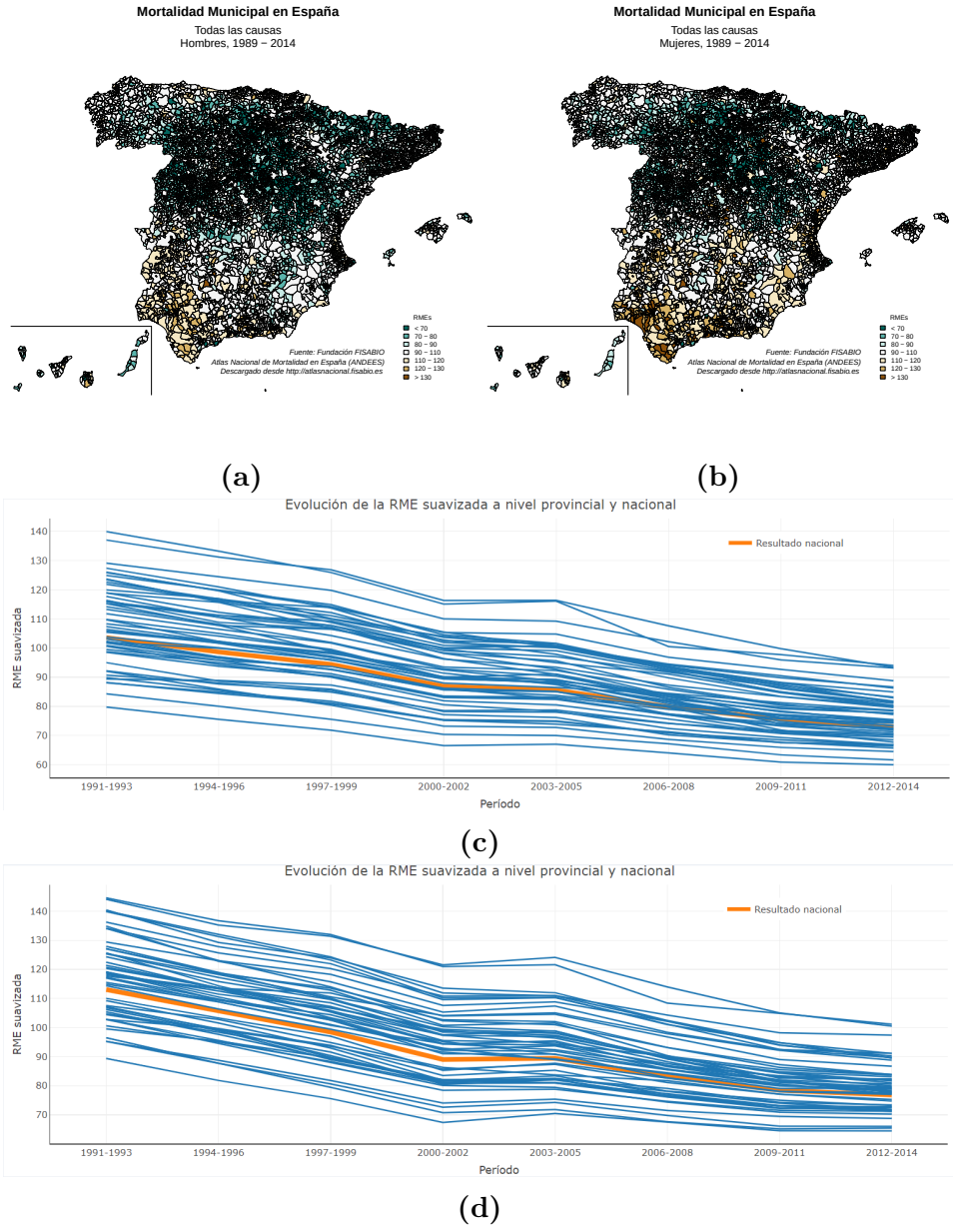


Figure 7.1.: All-causes mortality risk maps in the period 1989-2014 and the evolution of risks in men and women.

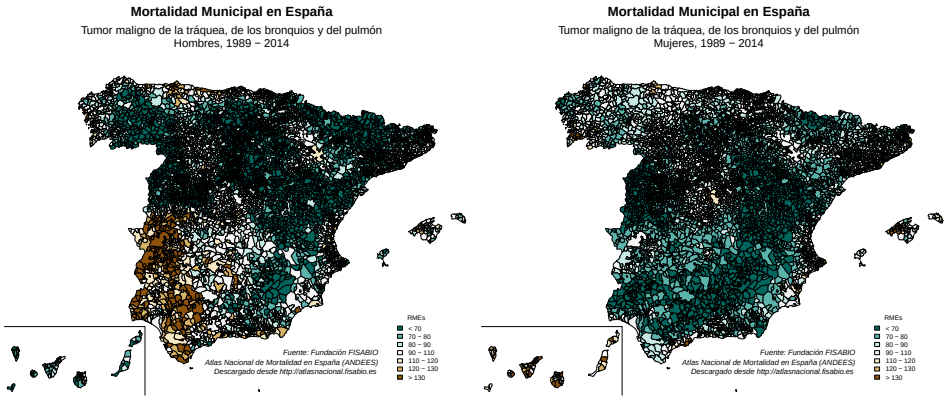
cause of death, accounting for 1.3% of all deaths, and the fifth tumoural mortality cause. As can be seen, the areas with the highest mortality risks for this cause in men (Figure 7.2a) are mainly concentrated in the southwestern part of the peninsula (Extremadura, Huelva, Sevilla and Cádiz), also in some areas of the Mediterranean coast and in some areas of Asturias. We found that nine out of the ten municipalities with the highest lung cancer mortality risks are in Extremadura, the community with the highest percentage of smokers according to the report on 'Smoking and Cancer in Spain' published in 2018 by the Spanish Association Against Cancer (AECC) (https://www.aecc.es/sites/default/files/content-file/Informe-tabaquisimo-cancer-20182.pdf) and historically a region hosting tobacco farming lands. In contrast, the areas with the lowest mortality risks are located in the northern half of the peninsula, as well as the northeast of Andalucía, east of Castilla-La Mancha and the Canary Islands. In the case of women (Figure 7.2b), the areas showing the highest mortality rates are scattered around many parts of the country, with the highlighted areas being the Canary Islands, some areas of Madrid, Bizkaia and Pontevedra, and coastal zones of Málaga, Alicante and western Mallorca. Some of the municipalities with the highest mortality risks correspond to tourist areas located on the coast of Malaga, Alicante and the Canary and Balearic Islands. During the period of study, large communities of elderly foreigners from northern Europe, whose women historically smoked in larger numbers than in Spain (Graham, 1996), have settled in these areas. It has been previously suggested that the presence of these communities could be having a clear impact on the mortality of these regions, it least in the case of Alicante (Zurriaga et al., 2008), due to the different historical smoking habits of these groups of women.

The evolution of lung cancer mortality risks at the provincial and national levels in men (Figure 7.2c) remains stable in most provinces in the periods between 1991 and 2008, exhibiting a slight downward trend from that point on. The provinces with the highest mortality (Huelva, Sevilla, Cádiz, Cáceres and Badajoz) show the sharpest decrease, which

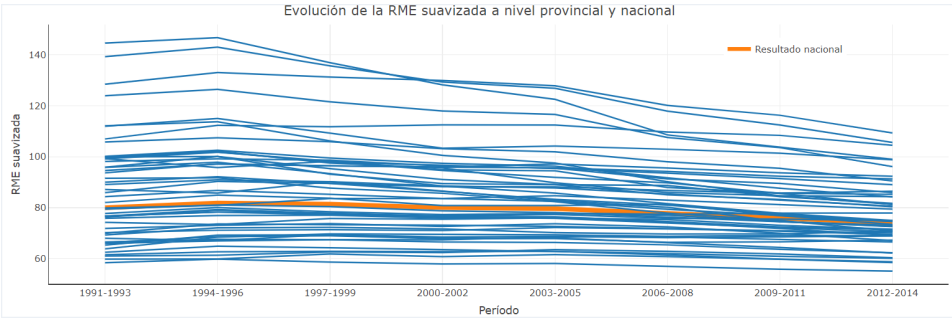
starts at the beginning of the period of study. Thus the geographical variability of risks in men decreases with the evolution of the period of study. For women (Figure 7.2d), lung cancer mortality shows a clear upeard trend from the year 2000. The spatio-temporal results show how, in general, the geographical mortality pattern found for each sex for the entire study period is maintained (except for the main time trend in women) with slight variations in the different subperiods for some particular locations (maps not shown).

7.3.3. Malignant tumor of the stomach mortality

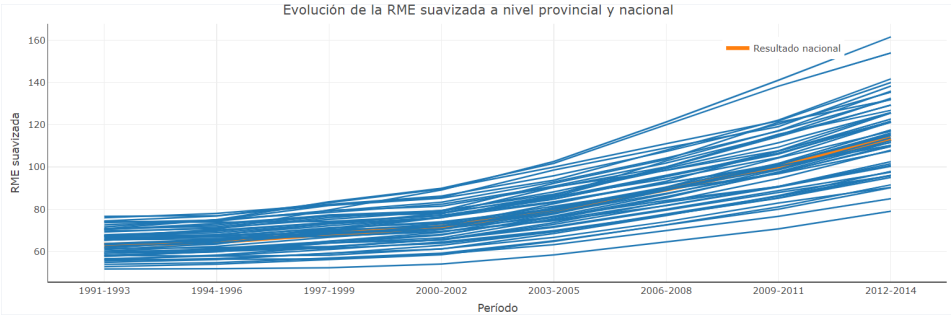
Figure 7.3 shows the malignant tumor of the stomach (just stomach cancer from now on) mortality risk maps for the period 1989-2014 and the evolution of those risks at the provincial and national levels in men and women. Of the 102 causes of death evaluated, stomach cancer represents the fourteenth cause of death in men, accounting for 1.9% of all deaths, and the fourth tumoural cause. For women, stomach cancer is the twenty-first cause of death, accounting for 1.4% of all deaths, and also the fourth leading tumoural cause. As can be seen, the areas with the highest mortality risks for this cause in men (Figure 7.3a) are mainly concentrated in Castilla y León and some neighboring areas. The Galician Atlantic coast, Cáceres, Ciudad Real and areas of the northern interior of Cataluña also show high mortality. The lowest relative risks are concentrated on the Canary and Balearic Islands, and more dispersed on the Mediterranean coast. In the case of women (Figure 7.3b), stomach cancer mortality shows a similar pattern to that of men, although it extends to a greater extent in areas of southern Galicia, northern Cataluña and Ciudad Real. This marked mortality pattern in the interior of Spain for both sexes could be associated with food consumption or food production tasks in these rural areas, where more cured meat and less fruits and vegetables are consumed than in the coastal zones (López-Abente et al., 2014). Overall, the geographical



(a) **(b)**



(c)



(d)

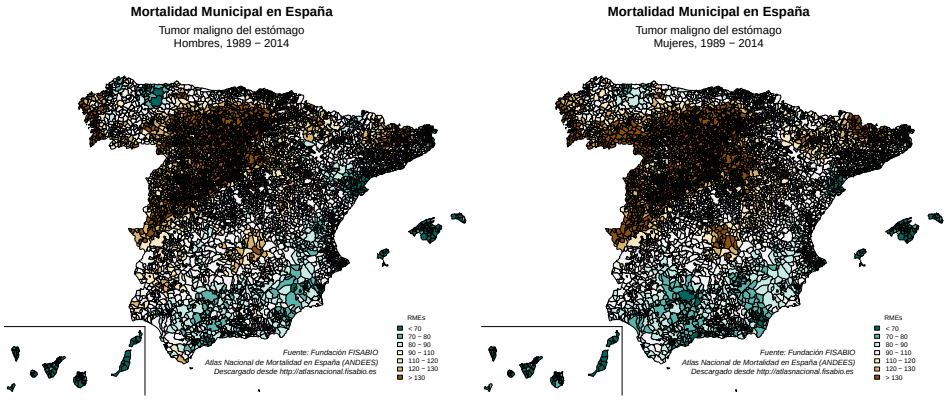
Figure 7.2.: Lung cancer mortality risk maps in the period 1989-2014 and the evolution of risks in men and women.

pattern found for stomach cancer is one of the more striking patterns in ANDEES because of the high geographical inequalities evidenced.

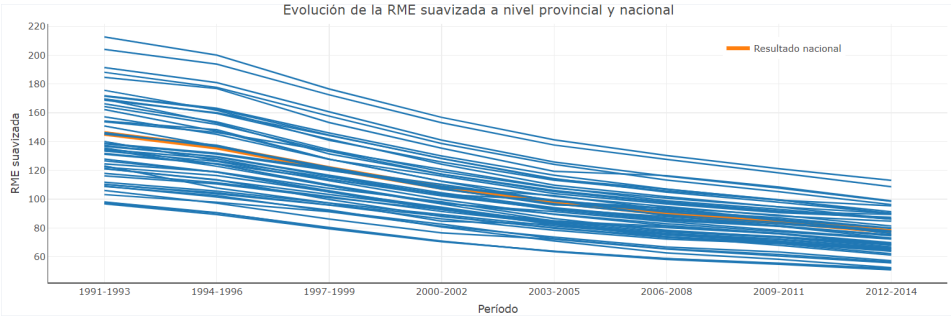
The evolution of stomach cancer mortality risks at the provincial and national levels for both sexes (Figures 7.3c and 7.3d) shows a downward trend. The spatio-temporal analyses show how the geographic pattern found in both sexes is maintained for the different subperiods of the study, and areas with a significant departure from the overall mortality trend for this disease are not found (maps not shown).

7.3.4. Diabetes mellitus mortality

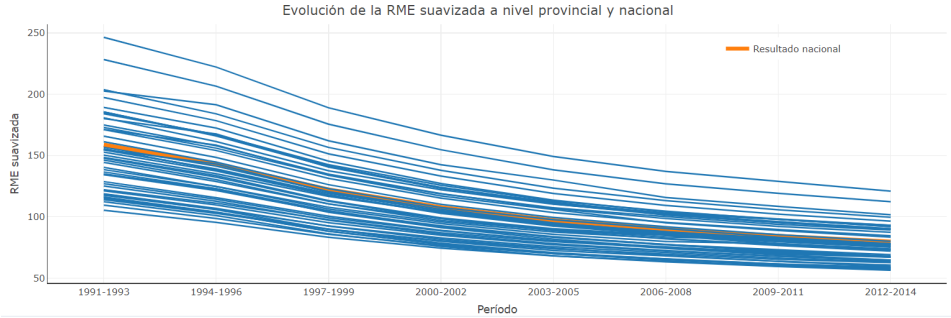
Figure 7.4 shows the diabetes mellitus (simply diabetes from now on) mortality risk maps in the period 1989-2014 and the temporal evolution of these risks at the provincial and national levels for both sexes. Of the 102 causes of death studied, diabetes represents the fifteenth cause of death in men, accounting for 1.9% of all deaths. For women, diabetes is the ninth cause of death, accounting for 3.4% of all deaths. As observed, the geographical distribution of the mortality relative risks for this cause in men (Figure 7.4a) shows lower mortality in the northern half of the peninsula and higher mortality in the southern half. The areas with the highest mortality are especially concentrated in the Canary Islands, Sevilla, Cádiz and zones of Jaén, Ciudad Real and Valencia, while those with the lowest mortality are located in the eastern provinces of Castilla y León, Galicia and Madrid. For women (Figure 7.4b), the territorial distribution also follows a north-south pattern, where the areas with the highest mortality are in the south: Extremadura, Andalucía, southern Castilla-La Mancha, Murcia, Valencia and the Canary Islands. Low risk areas are observed in the northeast of the Meseta (Soria, Segovia, Burgos) and some areas of Teruel, León and Galicia. In short, we find that the diabetes mortality risks for both sexes show a clear north-south gradient. The Canary Islands stands out as the region with the highest rates in Spain. Diabetes is a disease associated with junk food and the



(a) **(b)**



(c)



(d)

Figure 7.3.: Stomach cancer mortality risk maps in the period 1989-2014 and the temporal evolution of risks in men and women.

Canary Islands is precisely the region with the most obese population in Spain (Marcelino-Rodríguez et al., 2016).

The evolution in the diabetes mortality risk at the provincial and national levels in men (Figure 7.4c) has remained practically stable throughout the period between 1991 and 2005. Starting in 2005, we observe a slight downtrend, except in Las Palmas and Santa Cruz de Tenerife where the trend is upwards. For women (Figure 7.4d), the evolution of the mortality risk is similar to that of men. These figures show different spatio-temporal evolutions for the different regions of Spain. Figures 7.5 and 7.6 shows the mortality risk maps (only spatio-temporal interaction) for some subperiods in men and women, respectively. The rest of the subperiods not shown (for a question of space) do not add anything different to the maps shown in those figures. These figures allow us to identify those areas that have followed a different evolution to the country as a whole. For both sexes, we find that the municipalities with the starkest mortality risks decrease are located in areas of Extremadura, southern and western Andalucía, Madrid and coastal zones of Alicante, Valencia, Galicia and Cantabria. In constrast, we find that the municipalities with the steepest increases in risk are located in areas of the Canary Islands, Cataluña and Castilla y León. In the rest of the municipalities the risks remain stable during the different subperiods considered. In any case, diabetes is a clear example of spatio-temporal interaction in the evolution of mortality risks.

7.3.5. Leukemia mortality

Figure 7.7 shows the leukemia mortality risk maps for the period 1989-2014 and the evolution of such risks at the provincial and national levels for both sexes. Of the 102 causes of death studied, leukemia represents the thirty-fourth cause of death in men, accounting for 0.9% of all deaths. For women, leukemia is the thirty-fifth cause of death, accounting for 0.7% of all deaths. As can be observed, the mortality risk

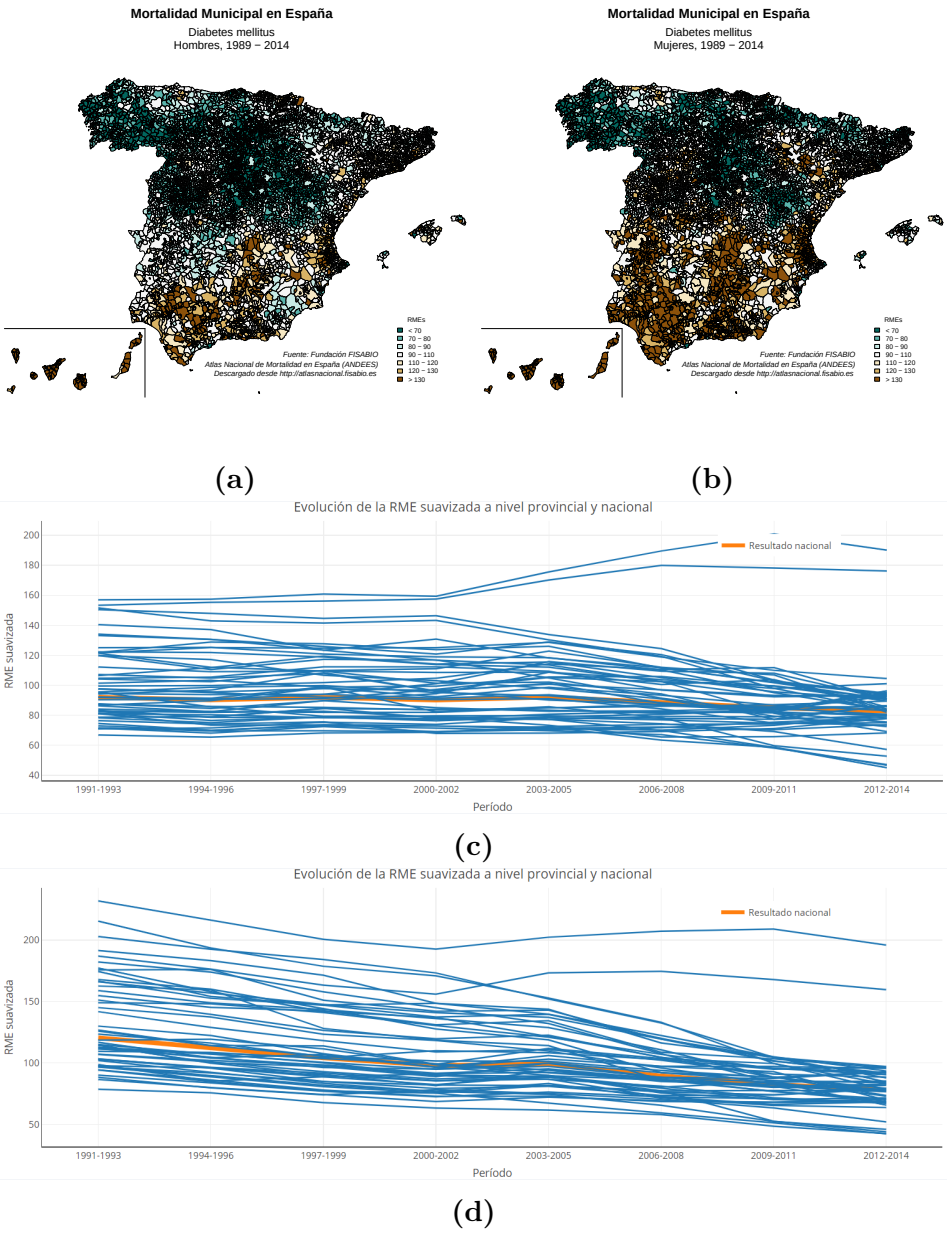


Figure 7.4.: Diabetes mortality risk maps in the period 1989-2014 and the evolution of risks in men and women.

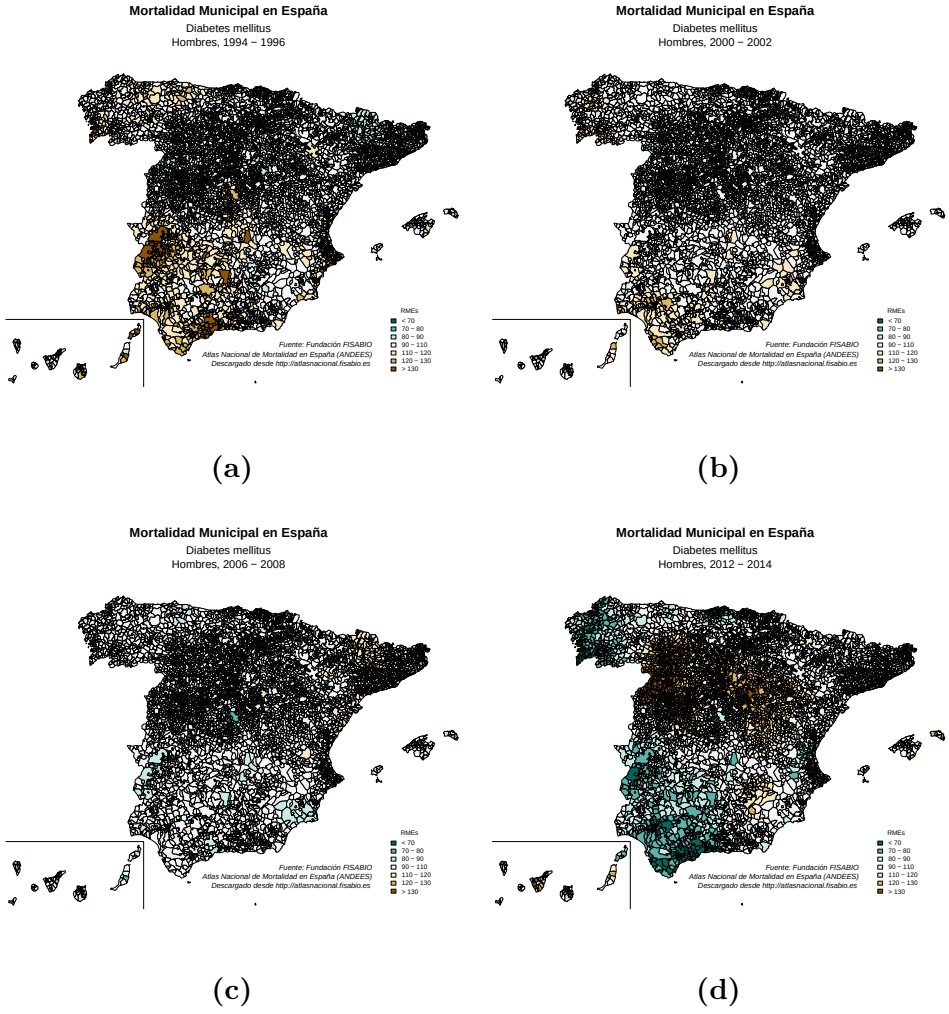
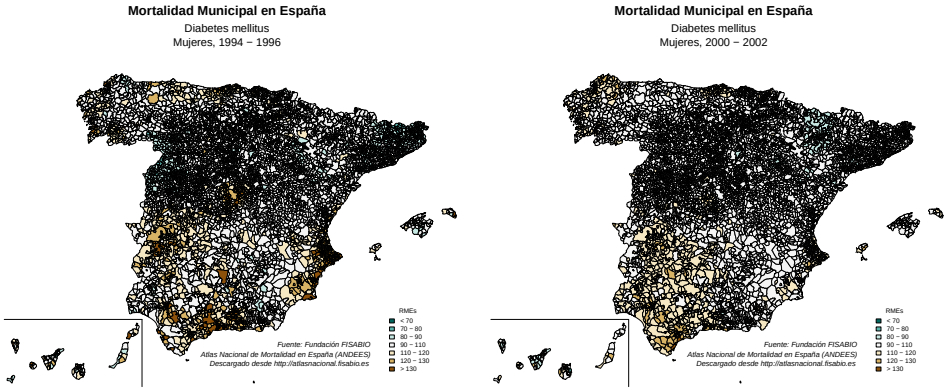
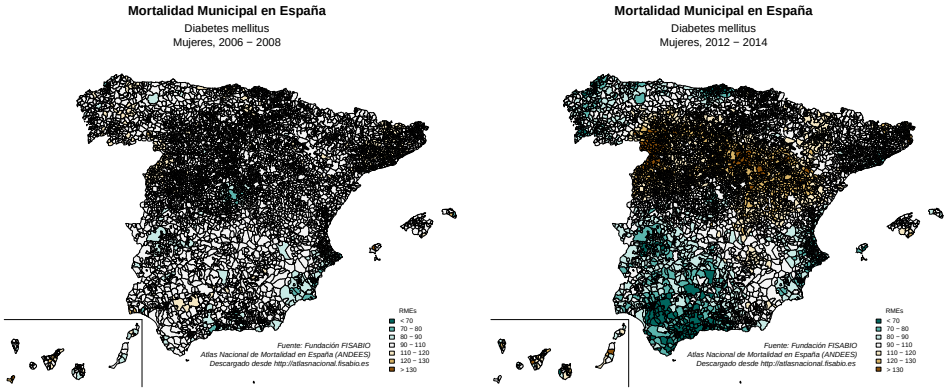


Figure 7.5.: Diabetes mortality risk maps (only spatio-temporal interaction) in some study subperiods in men.

distribution for this cause of death in men (Figure 7.7a) shows hardly any variability. We find that the areas with the lowest relative risks are found in areas of the Canary and Balearic Islands, Galicia and eastern Castilla y León. On the other hand, the regions with higher risks are found in areas of Cáceres, Barcelona and Córdoba. For women (7.7b),



(a) **(b)**



(c) **(d)**

Figure 7.6.: Diabetes mortality risk maps (only spatio-temporal interaction) in some study subperiods in women.

the geographic pattern is completely flat, with no areas with particularly higher/lower risks standing out from the whole country. Hardly any municipalities show significant risks ($P(SMRs > 100)$ above 0.95 or below 0.05) for the whole of Spain for either sex. In view of these results, we conclude that mortality from leukemia is distributed homogeneously

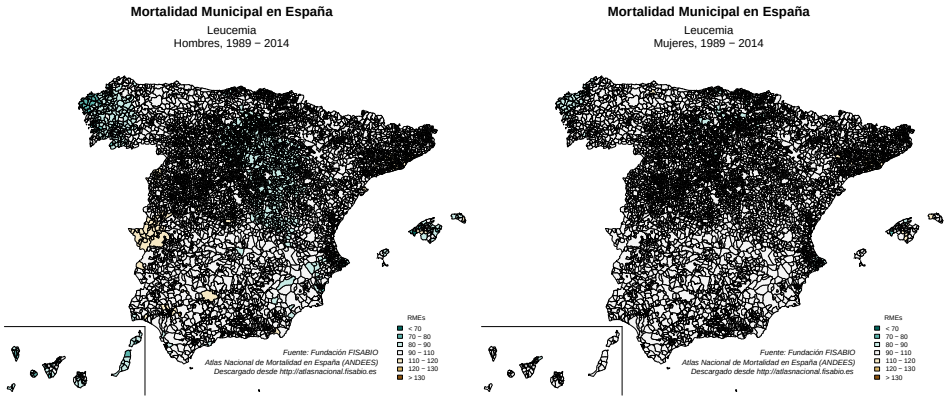
throughout the national territory, unlike mortality from the rest of the causes presented, which exhibits a marked spatial pattern. In other words, leukemia is a good example of an evenly distributed cause of mortality.

The spatio-temporal mortality evolution of leukemia in men (Figure 7.7c) and in women (Figure 7.7d) shows a downward trend throughout the study period. The spatio-temporal analyses show how the flat geographical distribution of risks for both sexes is maintained throughout the whole period, without any significant departure from this trend (maps not shown).

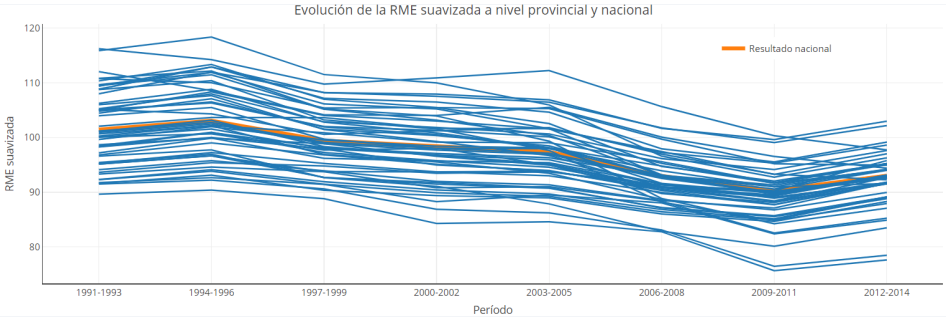
7.3.6. AIDS mortality

Figure 7.8 shows the AIDS mortality risk map for the period 1989-2014 and the evolution of those risks at the provincial and national levels in men. The risk map of mortality from AIDS in women is not available due to the small number of deaths (lower than 10,000) for that sex. Of the 102 causes of death studied, AIDS represents the thirty-second cause of death in men, accounting for 0.9% of all deaths. As can be seen, the areas with the highest mortality risks for this cause in men (Figure 7.8a) are mainly concentrated in zones of Sevilla and the Andalusian coast, Valencia and the Levantine coast, Asturias, Madrid and Barcelona.

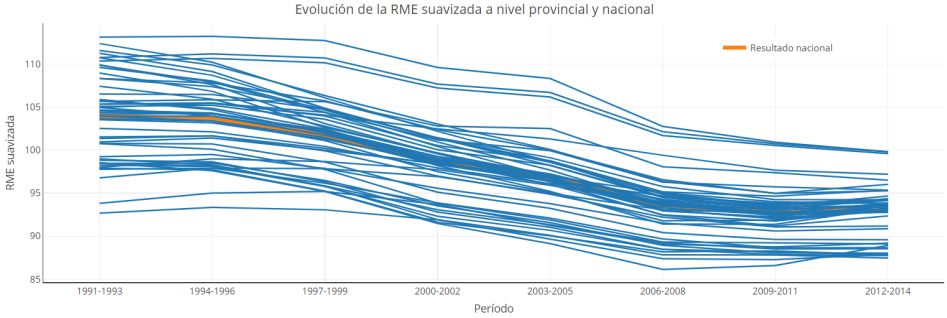
The temporal evolution of AIDS mortality in men (Figure 7.8b) shows a significant increase from 1991 to 1996. From then on, mortality exhibits a clear downward trend. The spatio-temporal results show the existence of an important spatio-temporal interaction. Figure 7.9 represents AIDS mortality risk maps (only spatio-temporal interaction) in men for four out of the eight subperiods. The rest of the subperiods not shown (for a question of space) do not add anything different to the maps shown in those figures. As previously mentioned, these figures allow for the identification of those areas with a more unique behavior in each subperiod. We found that the municipalities with the sharpest risk



(a) **(b)**



(c)



(d)

Figure 7.7.: Leukemia mortality risk maps in the period 1989-2014 and the evolution of risks in men and women.

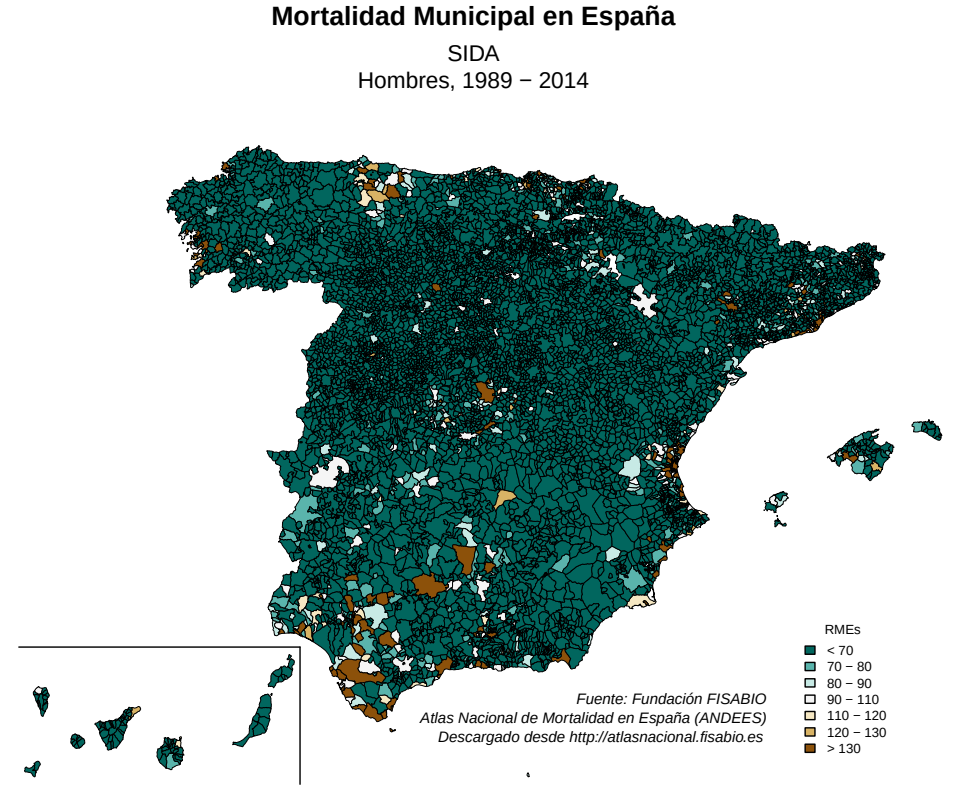
decrease are located in Cataluña, Valencia, Madrid, Vizkaia, Guipuzkoa and Navarra. These correspond in several cases to the more urban zones of Spain. In contrast, we find that the municipalities with the steepest increase are located in Andalucía, Extremadura, western Castilla y León and Galicia. In view of these results, it seems as if AIDS mortality shifts from urban areas to more socio-economically disadvantaged areas over the period of study.

7.4. Conclusions

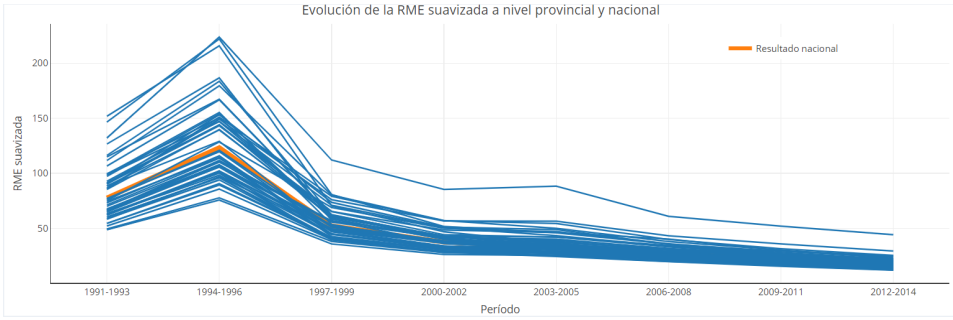
Mortality atlases on small geographic areas have proven to be a very useful tool for understanding the geographical distribution of diseases, identifying high-risk areas and implementing social and public health programs and interventions (Benach de Rovira and Martínez Martínez, 2013). In Spain, geographic mortality analyses have been very popular and widespread. Thus, if we just limit ourselves to the mortality atlases published in Spain in recent years, we could find at least nine relevant works: Benach et al. (2001); López-Abente et al. (2001); Martínez-Beneito et al. (2005); López-Abente et al. (2006); Ocaña et al. (2007); Borrell et al. (2009); Esnaola et al. (2010); Ocaña et al. (2010); Benach de Rovira and Martínez Martínez (2013). In these atlases, different levels of disaggregation (census tract, municipal, provincial), different regions (large cities, autonomous communities, national) and different methodologies (Bayesian spatial modeling, empirical-Bayesian random effects models, frequentist estimation) are used to study mortality. All these works have contributed to a better understanding of the geographical distribution of mortality in Spain, with important repercussions from the social, public health, clinical or healthcare points of view. Furthermore, these works have also had a notable scientific impact, as evidenced by the number of international publications that have used specific results from these works.

Despite the importance of those works, ANDEES is currently the largest and most updated mortality atlas developed in Spain, with data on almost 9.5 million deaths between 1989 and 2014. This tool permits the interactive visualization of the geographical distribution and temporal evolution of mortality risks over a long time period, enabling a detailed exploration of these results. ANDEES results show the existence of very different mortality geographic patterns depending on the cause, sex and period of study analyzed. We hope that the periodic updating of the information contained in ANDEES will make it possible to identify future geographic inequalities in the health of the population and to evaluate public health interventions. In summary, the use of modern statistical dissemination tools has made it possible to bring a new modern concept of mortality atlas to this study. In our opinion, this is the greatest value of this work.

ANDEES smoothing methodologies have made it possible to estimate the geographical distribution of mortality risks and their temporal evolution with an adequate degree of detail and reliability, providing an updated view of mortality distribution when one focuses on the latest subperiods. However, as we have shown in this thesis, the spatial and spatio-temporal models used to estimate risks could be extended by implementing more complex and flexible models that enable a deeper understanding of the distribution of diseases. As future work, we would be interested in implementing the previous models developed in this thesis and some combinations of them, such as multivariate adaptive hurdle models, to the geographic analysis of comprehensive Spanish mortality data.



(a)



(b)

Figure 7.8.: AIDS mortality risk maps in the period 1989-2014 and the evolution of risks in men.

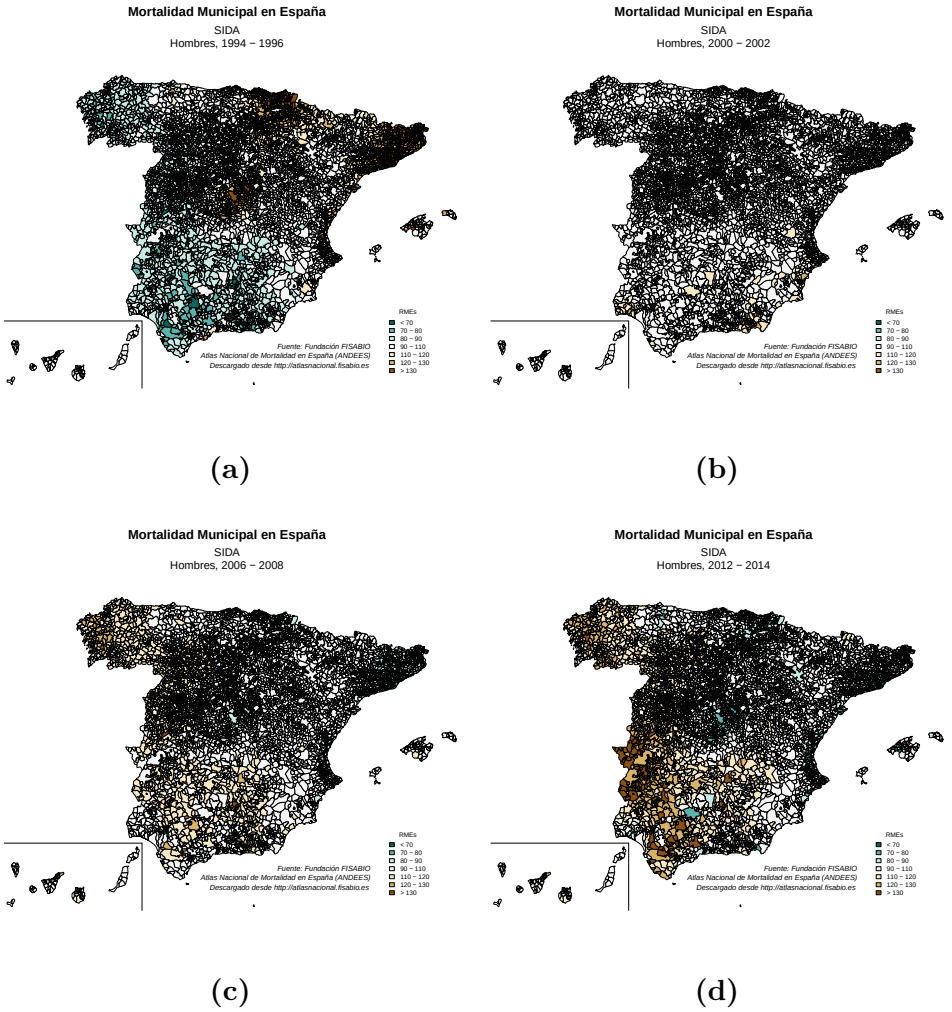


Figure 7.9.: AIDS mortality risks maps (only spatio-temporal interaction) in some study subperiods in men.

8. Conclusions and future work

In this thesis we implement, apply and evaluate some relevant models previously proposed in the disease mapping literature for estimating the geographical distribution of risks in different scenarios. After evaluating the behavior of these modeling proposals for mortality studies in different contexts, we found that they present important limitations. Therefore, our main goal in this thesis, at least in Chapters 4 to 6, has been to develop new proposals or extend those evaluated in order to resolve the limitations found.

First, regarding zero excess modeling, we show how zero excesses problems can be often found in practice when data are modeled using standard disease mapping models. In the Valencian Mortality Dataset used in Chapter 4, a relevant proportion of the diseases studied have been found to show zero excesses. Thus, as evidenced, zero excesses require attention for mortality geographic studies in general and specific models are needed to deal with this problem since, otherwise, maps with oversmoothed risks in the geographical units with low expected counts will be obtained. In this regard, we find that naive zero-inflation and hurdle models that propose to handle this lack of fit without an explicit modeling of the probabilities of zeroes, do not fix zero excesses problems well enough and are clearly unsatisfactory. Results sharply suggest the need for an explicit modeling of the probabilities that should vary across areal units. Unfortunately, as we prove in several theoretical results, these more flexible modeling strategies can easily lead to improper posterior distributions or arbitrary posterior distributions. This makes that modeling quite tricky, and caution has to be taken in order to

avoid flawed modeling proposals. Our results determine some ZIP and hurdle specific proposals, frequently proposed in the literature, that should be avoided in general. We finally propose several valid modeling alternatives that do not present the above problems and that are suitable for fitting zero excesses. We show that those proposals fix zero excesses problems and correct the mentioned oversmoothing of risks in low populated units, depicting geographic patterns more suited to the data.

In our work about multivariate disease mapping, we find that the Botella-Rocamora et al. (2015) proposal for the joint spatial modeling of several diseases shows some limitations when data are weaker. Specifically, in such situations, the prior structure of the $\boldsymbol{M}$-model can significantly influence the estimated risk patterns for all the diseases considered. As we show, this fact is caused by the single common variance parameter in the $\mathbf{M}$ matrix of this model, which controls the overall variability of all risk patterns fitted. If the variability of the risk patterns considered was different, this prior assumptions may produce evident misfits in the risk patterns that are estimated. One of the main contributions of this work has been to highlight these limitations, which are particularly worrisome when the original NVA proposal is applied to small regions of study. In this thesis, we propose two modifications of the previous multivariate model that incorporate several different parameters to model the variability of the risks for each disease and which allow us to solve the problems evidenced in the multivariate mortality study in the city of Castellón. These new heteroscedastic proposals allow the spatial patterns for each disease to have greater or lesser variability when necessary. This made it possible to obtain more flexible and accurate risk estimates.

In our work about adaptive spatial dependence, we propose a procedure to estimate the spatial weights matrix in CAR distributions according to retrospective multivariate data. As we show, our adaptive procedure makes CAR models more flexible and improves the fit of

subsequent analysis adopting the estimated weights matrix, which in principle should have captured the particularities that mortality data could show in that region. Additionally, the multivariate character of our proposal has shown itself to be an indispensable tool for appropriately estimating the spatial structure of the data.

The methodology introduced could have several different uses. First, the multivariate adaptive model introduced could be used in multivariate studies, considering dependence also between mortality causes, with adaptive spatial structures. These models should provide more accurate risk estimates that could take advantage of the adaptive character of the spatial dependence considered. A second use of adaptive CAR models would be the one emphasized in our work, that is, making inference on the spatial weights matrix of a region of study. As a consequence, that adaptive weights matrix could be later used in subsequent enhanced spatial disease mapping studies with a non-arbitrary spatial structure based on previous data and knowledge. We have also found a third practical use of our adaptive model. This use would be quality control of systematic problems that could be present in health data sets. Specifically, the mortality data of Valencia city used in the analysis in Chapter 6 belongs to a large Spanish project studying mortality in large cities, the MEDEA project. All the deaths in that data set have been geocodified by using several geocoding tools, in particular the Google geocoding API and a second geocoding tool (Cartociudad) of the Spanish Geographic National Institute. These tools, as with any other geocoding tool, are not perfect and they have errors for some particular streets, groups of cases that are geocodified in the city center, etc. that could distort the spatial analyses of that data base. We have found that the multivariate adaptive model on those data bases gives low spatial weights to those census tracts with systematic geocoding errors since their mortality data are somewhat different from their surrounding areas. This has allowed us to distill those errors (and correct them) by focusing on those census tracts with low spatial weights and no potential alternative explanation (no residential homes,

no socially marginal areas, no new building areas, etc.) for them.

Finally, the Spanish National Atlas of Mortality (ANDEES) developed allows us to know at the municipal level the geographical distribution and the temporal evolution of mortality due to a large set of causes of death throughout Spain. The results shown in ANDEES show the existence of very different geographic patterns of mortality depending on the cause, sex and period of study analyzed. This tool will allow researchers and public health experts to examine geographic patterns of diseases and detecting high-risk areas that are not evident through other types of analysis. The results presented can play a crucial role in the search for risk factors, as well as in the establishment of priorities and guide social and health policies.

ANDEES leaves open many possible lines of future work. On the one hand, we would like to update periodically the results of the atlas by incorporating the subsequent mortality data after 2014 as they are published. On the other hand, we would also like to implement other more complex and flexible models in order to deepen the understanding of the geographical distribution of diseases. Specifically, we would be interested in implementing each of the models developed in this thesis at a national massive level. Thus, we would evaluate (and fix) the possible existence of zeroes excess problems in each of the analyzed data sets. Furthermore, multivariate modeling considering groups of diseases that could have common risk factors would greatly improve the geographic estimation of risks by making use of alternative sources of information. Similarly, adaptive spatial modeling would also allow obtaining risk maps with greater variability, allowing municipalities with special characteristics to show the separate behavior that they require. Finally, the combination of spatio-temporal modeling also with these proposals, those that show a more evident improvement on the spatial analysis, would allow obtaining an updated and more precise view of the risks. The implementation of some of these models for the analysis of mortality in the whole of Spain could give rise to

challenging computational problems, given the large size of the study region considered and the large number of geographical patterns to be estimated in a single model. As a consequence, another future line of work would be solving such computational problems by exploring different computing tools and optimizing the implementation of each of the proposed models.

A. Supplementary material to the paper: *"Some findings on zero-inflated and hurdle Poisson models for disease mapping"*

A.1. Theoretical results

Then, we shows four different general results on the modeling of zero-inflated and hurdle Poisson models with either fixed or random effects. At the end we draw two corollaries with some specific results of particular interest for the paper above.

Result 1. *Let* $\mathbf{O} = \{O_i : i = 1, ..., I\}$ *be independent observations from the hurdle Poisson model*

$$O_i \sim (1 - \pi_i(\boldsymbol{u}, \sigma))^{1_{\{0\}}(O_i)} \left(\pi_i(\boldsymbol{u}, \sigma) \frac{Poi(O_i \mid E_i R_i)}{1 - Poi(0 \mid E_i R_i)} \right)^{1_{(0,\infty)}(O_i)},$$

where

$$\pi_i(\boldsymbol{u}, \sigma) = F(\sigma \boldsymbol{z}_i \boldsymbol{u}), \;\; with \;\; \boldsymbol{u} \sim f(\boldsymbol{u}) = N_I(\mathbf{0}, \boldsymbol{I}),$$

being F *a distribution function with* $F(-x) = 1 - F(x)$ *and* $\{\boldsymbol{z}_i : i = 1, ..., I\}$ *a set of* I*-dimensional vectors. Let also* $\boldsymbol{Z}^\star$ *be the* $I \times I$ *matrix with rows* $\boldsymbol{z}_i^\star$ *defined as* $\boldsymbol{z}_i$ *if* $O_i = 0$ *or* $-\boldsymbol{z}_i$ *if* $O_i > 0$*. Assume that* σ*,* $\boldsymbol{u}$ *and* $\boldsymbol{R}$ *are independent a priori and* σ *follows an improper prior distribution* $f(\sigma)$*. Let*

$$\mathcal{C} = \{\boldsymbol{v} \in \mathcal{R}^I : \;\; \boldsymbol{Z}^\star \boldsymbol{v} \leq 0\},$$

if the following condition is satisfied

$$dimension\,(\mathcal{C}) = I, \tag{A.1.1}$$

then the posterior distribution $f(\boldsymbol{u}, \sigma, \boldsymbol{R} \mid \boldsymbol{O})$ *is improper independently on the prior distribution* $f(\boldsymbol{R})$ *assumed for* $\boldsymbol{R}$*.*

Proof. The proof uses a similar technique as the proof for impropriety of posterior distributions in Bernoulli experiments derived in Natarajan and McCulloch (1995) (Theorem 1.i).

We have to show that the integral

$$\int L(\boldsymbol{u}, \sigma, \boldsymbol{R}; \boldsymbol{O})\, f(\boldsymbol{u}, \sigma, \boldsymbol{R})\, d\boldsymbol{u} d\sigma d\boldsymbol{R}$$

diverges, where L is the likelihood function, that is,

$$L(\boldsymbol{u}, \sigma, \boldsymbol{R}; \boldsymbol{O}) = \prod_{i=1}^{I} (1-\pi_i(\boldsymbol{u}, \sigma))^{1_{\{0\}}(O_i)} \Big(\pi_i(\boldsymbol{u}, \sigma) \frac{Poi(O_i \mid E_i R_i)}{1 - Poi(0 \mid E_i R_i)}\Big)^{1_{(0,\infty)}(O_i)}$$

and f is the prior distribution. It can be easily seen that the integral above is:

$$\int \prod_{\{i:O_i=0\}} \frac{Poi(O_i \mid E_i R_i)}{1 - Poi(0 \mid E_i R_i)} \Big\{ \int \int_{\mathcal{R}^I} \prod_{\{i:O_i=0\}} (1-F(\sigma \boldsymbol{z}_i \boldsymbol{u})) \prod_{\{i:O_i>0\}} F(\sigma \boldsymbol{z}_i \boldsymbol{u}) f(\sigma) f(\boldsymbol{u}) d\boldsymbol{u} d\sigma \Big\} f(\boldsymbol{R})\, d\boldsymbol{R}.$$

Bearing in mind that $F(-x) = 1 - F(x)$ the inner integral above results:

$$\int \int_{\mathcal{R}^I} \prod_{\{i:O_i=0\}} (1 - F(\sigma \boldsymbol{z}_i^* \boldsymbol{u})) \prod_{\{i:O_i>0\}} F(-\sigma \boldsymbol{z}_i^* \boldsymbol{u}) f(\sigma) f(\boldsymbol{u}) d\boldsymbol{u} d\sigma =$$

$$\int \int_{\mathcal{R}^I} \prod_{i=1}^{I} (1 - F(\sigma \boldsymbol{z}_i^* \boldsymbol{u})) f(\sigma) f(\boldsymbol{u}) d\boldsymbol{u} d\sigma \geq \int \int_{\mathcal{C}} \prod_{i=1}^{I} (1 - F(\sigma \boldsymbol{z}_i^* \boldsymbol{u})) f(\sigma) f(\boldsymbol{u}) d\boldsymbol{u} d\sigma \geq$$

$$\geq \int f(\sigma) d\sigma \int_{\mathcal{C}} \frac{1}{2^I} f(\boldsymbol{u}) d\boldsymbol{u} = \frac{1}{2^I} \int f(\sigma) d\sigma.$$

The last integral obviously diverges if $f(\sigma)$ is improper.

□

Result 2. *Let* $\mathbf{O} = \{O_i : i = 1, ..., I\}$ *be independent observations from the hurdle Poisson model*

$$O_i \sim (1 - \pi_i(\boldsymbol{\beta}))^{1_{\{0\}}(O_i)} \left(\pi_i(\boldsymbol{\beta}) \frac{Poi(O_i \mid E_i R_i)}{1 - Poi(0 \mid E_i R_i)}\right)^{1_{(0,\infty)}(O_i)},$$

where

$$\pi_i(\boldsymbol{\beta}) = F(\boldsymbol{x}_i \boldsymbol{\beta}),$$

$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$ are J-dimensional vectors of known covariates and F a distribution function. Suppose that $\boldsymbol{\beta}$ is, a priori, independent of $\boldsymbol{R}$ with prior distribution

$$\boldsymbol{\beta} \sim \prod_{j=1}^{J} f_j(\beta_j), \quad \beta_j \in \mathcal{R}.$$

If for any $1 \leq j^* \leq J$, $x_{ij^*} > 0$ for all i with $O_i > 0$ and negative otherwise (respectively $x_{ij^*} > 0$ for all i with $O_i = 0$ and negative otherwise) and $\int f(\beta_{j^*}) d\beta_{j^*}$ diverges for large positive (respectively negative) values of β_{j^*} then the posterior distribution $f(\boldsymbol{\beta}, \boldsymbol{R} \mid \boldsymbol{O})$ is improper independently on the prior distribution $f(\boldsymbol{R})$ assumed for $\boldsymbol{R}$.

Proof. Let us assume the case $x_{ij^*} > 0$ for all i with $O_i > 0$ and negative otherwise. The likelihood function can be put as

$$L(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) = \prod_{\{i:O_i=0\}} (1 - \pi_i(\boldsymbol{\beta})) \prod_{\{i:O_i>0\}} \pi_i(\boldsymbol{\beta}) \frac{Poi(O_i \mid E_i R_i)}{1 - Poi(0 \mid E_i R_i)} =$$

$$\left(\prod_{\{i:O_i=0\}} (1 - F(\boldsymbol{x}_i \boldsymbol{\beta})) \prod_{\{i:O_i>0\}} F(\boldsymbol{x}_i \boldsymbol{\beta}) \right) \prod_{\{i:O_i>0\}} \frac{Poi(O_i \mid E_i R_i)}{1 - Poi(0 \mid E_i R_i)}.$$

Thus,

$$\int_R L(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) f(\beta_{j^*}) d\beta_{j^*} > \int_0^\infty L(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) f(\beta_{j^*}) d\beta_{j^*} \propto$$

$$\int_0^\infty \prod_{\{i:O_i=0\}} (1 - F(\boldsymbol{x}_i \boldsymbol{\beta})) \prod_{\{i:O_i>0\}} F(\boldsymbol{x}_i \boldsymbol{\beta}) f(\beta_{j^*}) d\beta_{j^*}.$$

Since F is a distribution function is also, in particular, an increasing function. Moreover, as $x_{ij^*} > 0$ for all i with $O_i > 0$ and negative otherwise we have that, if $\boldsymbol{\beta}_0 = (\beta_1, ..., \beta_{j^*-1}, 0, \beta_{j^*+1}, ..., \beta_J)$, this

expression is greater than

$$\prod_{\{i:O_i=0\}} (1 - F(\boldsymbol{x}_i\boldsymbol{\beta}_0')) \prod_{\{i:O_i>0\}} F(\boldsymbol{x}_i\boldsymbol{\beta}_0') \int_0^\infty f(\beta_{j^*}) d\beta_{j^*}$$

which diverges due to the prior impropriety of $f(\beta_{j^*})$ for large positive values.

The proof for the case $x_{ij^*} > 0$ if $O_i = 0$ and negative otherwise is analogous. □

Result 3. *Let* $\mathbf{O} = \{O_i : i = 1, ..., I\}$ *be independent observations from the ZIP model*

$$O_i \sim (1 - \pi_i(\boldsymbol{u}, \sigma))1_{\{0\}}(O_i) + \pi_i(\boldsymbol{u}, \sigma)\, Poi(O_i \mid E_i R_i),$$

where

$$\pi_i(\boldsymbol{u}, \sigma) = F(\sigma \boldsymbol{z}_i^t \boldsymbol{u}), \;\; \boldsymbol{u} \sim f(\boldsymbol{u}) = N_I(\mathbf{0}, \boldsymbol{I}_I),$$

being F *a distribution function with* $F(-x) = 1 - F(x)$ *and* $\{\boldsymbol{z}_i, i = 1, ..., I\}$ *a set of* I*-dimensional vectors. Let also* $\boldsymbol{Z}^\star$ *be the* $I \times I$ *matrix with rows* $\boldsymbol{z}_i^\star$ *defined as* $\boldsymbol{z}_i$ *if* $O_i = 0$ *or* $-\boldsymbol{z}_i$ *if* $O_i > 0$*. Assume that* σ*,* $\boldsymbol{u}$ *and* $\boldsymbol{R}$ *are independent a priori and* σ *follows an improper prior distribution* $f(\sigma)$*. Let*

$$\mathcal{C} = \{\boldsymbol{v} \in \mathcal{R}^I : \;\; \boldsymbol{Z}^\star \boldsymbol{v} \leq 0\},$$

if the following condition is satisfied

$$dimension\,(\mathcal{C}) = I,$$

then the posterior distribution $f(\boldsymbol{u}, \sigma, \boldsymbol{R} \mid \boldsymbol{O})$ *is improper independently on the prior distribution* $f(\boldsymbol{R})$ *assumed for* $\boldsymbol{R}$*.*

Proof. The Likelihood function for this model can be expressed as:

$$L(\boldsymbol{u},\sigma,\boldsymbol{R};\boldsymbol{O}) = \prod_{i=1}^{I}(1-\pi_i(\boldsymbol{u},\sigma)+\exp(-E_iR_i))^{1_{\{0\}}(O_i)}(\pi_i(\boldsymbol{u},\sigma)Poi(O_i|E_iR_i))^{1_{(0,\infty)}(O_i)} \geq$$

$$\prod_{i=1}^{I}(1-\pi_i(\boldsymbol{u},\sigma))^{1_{\{0\}}(O_i)}(\pi_i(\boldsymbol{u},\sigma)Poi(O_i|E_iR_i))^{1_{(0,\infty)}(O_i)},$$

which is proportional, as a function of $\boldsymbol{u}$ and σ to the likelihood function of the hurdle Poisson model. Since the conditions of this results are the same than for Result 1 and there $\int f(\boldsymbol{u},\sigma,\boldsymbol{R}\mid\boldsymbol{O})d\boldsymbol{u}d\sigma$ diverged, it follows that $f(\boldsymbol{u},\sigma,\boldsymbol{R}\mid\boldsymbol{O})$ is now improper as a direct consequence of that Result. □

Result 4. *Let* $\mathbf{O} = \{O_i : i = 1, ..., I\}$ *be independent observations from the ZIP model*

$$O_i \sim (1-\pi_i(\boldsymbol{\alpha}))1_{\{0\}}(O_i) + \pi_i(\boldsymbol{\alpha})\, Poi(O_i \mid E_iR_i),$$

where

$$\pi_i(\boldsymbol{\alpha}) = F(\boldsymbol{x}_i\boldsymbol{\beta}),$$

$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iJ})$ *are* J*-dimensional vectors of known covariates and* F *a distribution function. Suppose that* $\boldsymbol{\beta}$ *is, a priori, independent of* $\boldsymbol{R}$ *with prior distribution*

$$\boldsymbol{\beta} \sim \prod_{j=1}^{J} f_j(\beta_j), \quad \beta_j \in \mathcal{R}.$$

If for any $1 \leq j^* \leq J$, $x_{ij^*} > 0$ *for all* $i = 1, \ldots, I$ *(respectively* $x_{ij^*} < 0$*) and* $\int f(\beta_{j^*})d\beta_{j^*}$ *diverges for large positive (respectively negative) values of* β_{j^*} *then* $f(\boldsymbol{\beta},\boldsymbol{R}\mid\boldsymbol{O})$ *is improper independently on the prior distribution* $f(\boldsymbol{R})$ *assumed for* $\boldsymbol{R}$*.*

Proof. Let us assume the case $x_{ij^*} > 0$ for all i. The likelihood function is

$$L(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) = \prod_{i=1}^{I} \Big((1 - \pi_i(\boldsymbol{\alpha})) 1_{\{0\}}(O_i) + \pi_i(\boldsymbol{\alpha})\, Poi(O_i \mid E_i R_i) \Big).$$

The above expression corresponds to a sum with 2^I positive terms. Obviously if the integral of any one of these terms times the prior is divergent then the posterior distribution would be improper. One of these terms in the likelihood is

$$L_1(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) = \prod_{i=1}^{I} Poi(O_i \mid E_i R_i) \pi_i(\boldsymbol{\alpha}) = \prod_{i=1}^{I} Poi(O_i \mid E_i R_i) F(\boldsymbol{x}_i \boldsymbol{\beta}).$$

The function $F(\cdot)$ is increasing and therefore, as $x_{ij^*} > 0$, is also an increasing function of β_{j^*}. Then, if $\boldsymbol{\beta}_0 = (\beta_1, \ldots, \beta_{j^*-1}, 0, \beta_{j^*+1}, \ldots, \beta_J)$

$$L_1(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) \geq L_1(\boldsymbol{\beta}_0, \boldsymbol{R}; \boldsymbol{O})$$

for any $\boldsymbol{\beta}$ with $\beta_{j^*} > 0$. Then

$$\int L_1(\boldsymbol{\beta}, \boldsymbol{R}; \boldsymbol{O}) f(\boldsymbol{\beta})\, d\beta_{j^*} \geq L_1(\boldsymbol{\beta}_0, \boldsymbol{R}; \boldsymbol{O}) \int_0^{\infty} f_{j^*}(\beta_{j^*})\, d\beta_{j^*},$$

and, since $f_{j^*}(\beta_{j^*})$ diverges for large positive values, $f(\boldsymbol{\beta}, \boldsymbol{R} \mid \boldsymbol{O})$ is improper.

The proof of the Result for $x_{ij} < 0$ is analogous. □

Corollary 1. *Let us consider a hurdle Poisson model with Poisson means modeled as a BYM model and probabilities of zeroes as*

$$logit(\pi_i) = \boldsymbol{x}_i \boldsymbol{\beta} + v_i,$$

for $\boldsymbol{v} \sim N_I(\boldsymbol{0}, \sigma^2 \boldsymbol{C})$ *a vector of random effects with* $\boldsymbol{C}$ *a symmetric, positive-definite structure matrix and* $\boldsymbol{x}_i$ *a vector of covariates of length*

J. Then,

1. *If $f(\sigma)$ is improper then $f(\boldsymbol{\beta}, \sigma, \boldsymbol{u}, \boldsymbol{R}|\boldsymbol{O})$ diverges regardless of $f(\beta, \boldsymbol{u}, \boldsymbol{R})$.*

2. *Let us assume $\boldsymbol{\beta}$ to be a priori independent with $\beta_j \sim f(\beta_j)$ $j = 1, ..., J$, and there is a j^* $(1 \leq j^* \leq J)$ with $x_{ij^*} > 0$ when $O_i > 0$ and negative otherwise (respectively $x_{ij^*} > 0$ when $O_i = 0$ and negative otherwise). If $\int f(\beta_{j^*}) d\beta_{j^*}$ diverges for large positive (respectively negative) values, then $f(\boldsymbol{\beta}, \sigma, \boldsymbol{u}, \boldsymbol{R}|\boldsymbol{O})$ is improper regardless of $f(\sigma, \boldsymbol{u}, \boldsymbol{R})$.*

3. *Both previous results also hold for:*

   - *probit or tobit link functions for modelling $\boldsymbol{\pi}$.*
   - *non-Poisson discrete likelihoods (such as binomial or negative-binomial).*
   - *other spatial structures, beyond BYM, for the Poisson means.*

Proof.

Proof for item 1:

This is just a particular case of Result 1 for $F(x) = antilogit(x)$, which is the distribution function for a logistic density. Moreover, the linear term for $\boldsymbol{\pi}$ is a bit different since we have now random effects v_i instead of $\sigma \boldsymbol{z}_i \boldsymbol{u}$ and fixed effects.

Regarding $\boldsymbol{v}$, if $\boldsymbol{C} = \boldsymbol{\Lambda} \boldsymbol{D}^2 \boldsymbol{\Lambda}'$ is the eigendecomposition of $\boldsymbol{C}$, then $\boldsymbol{v} = \sigma \boldsymbol{Z} \boldsymbol{u}$ for $\boldsymbol{u} \sim N_I(\boldsymbol{0}_I, \boldsymbol{I}_I)$ and $\boldsymbol{Z} = \boldsymbol{\Lambda} \boldsymbol{D}$. In this case the matrix $\boldsymbol{Z}^*$ in Result 1 would be $\boldsymbol{Z}^* = \boldsymbol{L} \boldsymbol{Z}$ for $\boldsymbol{L}$ a diagonal matrix with $L_{ii} \in \{-1, 1\}$ for all i. Since $\boldsymbol{C}$ is positive definite then $\boldsymbol{Z}$ is of full

rank. Besides, since both $\boldsymbol{L}$ and $\boldsymbol{Z}$ are full rank, then $\boldsymbol{Z}^*$ is also full rank and therefore regular so $\boldsymbol{y}\boldsymbol{Z}^* = \boldsymbol{0}$ iif $\boldsymbol{y} = \boldsymbol{0}$. Natarajan and McCulloch (1995) (Section 2.4) state that Condition (A.1.1) holds iif there is no nonnegative vector, $\boldsymbol{y} \neq \boldsymbol{0}$ such that $\boldsymbol{y}\boldsymbol{Z}^* = \boldsymbol{0}$ so, according to this criterion, that condition holds also for this Corollary.

Finally, regarding the fixed effects in the linear term for $\boldsymbol{\pi}$. The last integral in Result 1 would be now of the form

$$\int \int_{\mathcal{R}^I} \prod_{i=1}^{I} (1 - F(\boldsymbol{x}_i\boldsymbol{\beta} + \sigma \boldsymbol{z}_i^* \boldsymbol{u})) f(\sigma) f(\boldsymbol{u}) d\boldsymbol{u} d\sigma$$

which by similar bounding arguments as those used in the last part of the proof of Result 1, is a divergent integral.

Proof for item 2:
This is just a particular case of Result 2 for $F(x) = antilogit(x)$ and with an additional random effects term in the linear predictor. This term would not interfere at all in last integral of the proof of that result which makes the posterior distribution improper. So this result keeps being valid with the additional random effects term.

Proof for item 3:
First note that using probit or tobit link functionswould be equivalent to consider normal or t probability density functions for $F()$. So these would be also particular cases of Result 1. Note also that the Poisson likelihood does not have any effect on the posterior impropriety of the proofs of Results 1 and 2 so this could also be changed to binomial or negative-binomial distributions, for example. Finally, note that the BYM model for the Poisson means is irrelevant for the posterior impropriety in these models since the impropriety comes from their zero-specific terms. □

Corollary 2. *Let us consider a ZIP model with Poisson means modeled as a BYM model and probabilities of extra-Poisson zeroes as*

$$logit(\pi_i) = \boldsymbol{x}_i\boldsymbol{\beta} + v_i,$$

for $\boldsymbol{v} \sim N_I(\boldsymbol{0}, \sigma^2\boldsymbol{C})$ *a vector of random effects with* $\boldsymbol{C}$ *a symmetric, positive-definite full-rank structure matrix and* $\boldsymbol{x}_i$ *a vector of covariates of length* J*. Then,*

1. *If* $f(\sigma)$ *is improper then* $f(\boldsymbol{\beta}, \sigma, \boldsymbol{u}, \boldsymbol{R}|\boldsymbol{O})$ *diverges regardless of* $f(\beta, \boldsymbol{u}, \boldsymbol{R})$*.*

2. *Let us assume* $\boldsymbol{\beta}$ *to be a priori independent with* $\beta_j \sim f(\beta_j)$ $j = 1, ..., J$*, and there is a* j^* $(1 \leq j^* \leq J)$ *with* $x_{ij^*} > 0$ *for* $i = 1, ..., I$ *(respectively* $x_{ij^*} < 0$ *for* $i = 1, ..., I$*). If* $\int f(\beta_{j^*})d\beta_{j^*}$ *diverges for large positive (respectively negative) values then* $f(\boldsymbol{\beta}, \sigma, \boldsymbol{u}, \boldsymbol{R}|O)$ *is improper regardless of* $f(\sigma, \boldsymbol{u}, \boldsymbol{R})$*.*

3. *Both previous results also hold for:*

   - *probit or tobit link functions for modelling* $\boldsymbol{\pi}$*.*
   - *non-Poisson discrete likelihoods (such as binomial or negative-binomial).*
   - *other spatial structures beyond BYM for the Poisson means.*

Proof.
Follow the same argument than for Corollary 1 applied to Results 3 and 4.

□

A.2. Additional results

A.2.1. Observed and predicted zeroes for models in Section 4.2

Table A.1.: Observed zeroes for each data set and posterior predicted zeroes for each model. Values in the Obs. zeroes column correspond to the real observed zeroes for each data set. For the 3 columns on the right, numbers correspond to the posterior predictive median for this same quantity for each model run and the corresponding unilateral 95% posterior predictive interval. Bold fonts denote those combinations of models and data sets evidencing zero excesses according to their predictive intervals.

Sex & Cause	Obs. zeroes	BYM	ZIP	Hurdle
(Men, All tumours)	4	2 [0,5]	3 [0,5]	5 [0,11]
(Women, All tumours)	7	6 [0,10]	6 [0,10]	8 [0,15]
(Men, Mouth)	216	**196 [0,211]**	**199 [0,215]**	216 [0,242]
(Men, Stomach)	105	**91 [0,103]**	**92 [0,104]**	105 [0,127]
(Women, Stomach)	150	137 [0,151]	138 [0,152]	150 [0,173]
(Men, Colorectal)	73	**58 [0,68]**	**59 [0,69]**	74 [0,93]
(Women, Colorectal)	74	72 [0,82]	73 [0,83]	74 [0,93]
(Men, Colon)	96	**79 [0,91]**	84 [0,96]	96 [0,119]
(Women, Colon)	98	91 [0,102]	92 [0,104]	99 [0,119]
(Men, Rectum)	201	**180 [0,196]**	**183 [0,199]**	202 [0,228]
(Women, Rectum)	234	223 [0,239]	225 [0,242]	235 [0,262]
(Men, Liver)	156	**138 [0,153]**	**139 [0,153]**	157 [0,182]
(Women, Liver)	188	176 [0,191]	178 [0,193]	188 [0,214]
(Women, Vesicle)	243	239 [0,255]	241 [0,256]	244 [0,270]
(Men, Pancreas)	179	**163 [0,178]**	165 [0,181]	179 [0,205]
(Women, Pancreas)	194	180 [0,196]	184 [0,201]	194 [0,220]
(Men, Larynx)	214	**186 [0,203]**	**187 [0,203]**	214 [0,240]
(Men, Lung)	34	**25 [0,32]**	**25 [0,33]**	34 [0,47]
(Women, Lung)	199	188 [0,203]	189 [0,205]	199 [0,224]
(Women, Breast)	80	73 [0,85]	74 [0,85]	80 [0,101]
(Women, Uterus)	188	187 [0,201]	188 [0,204]	188 [0,215]
(Women, Ovary)	201	195 [0,210]	195 [0,211]	202 [0,227]
(Men, Prostate)	62	**51 [0,60]**	53 [0,63]	63 [0,81]
(Men, Bladder)	123	**104 [0,117]**	**105 [0,119]**	124 [0,146]

Sex & Cause	Obs. zeroes	BYM	ZIP	Hurdle
(Men, Lymphatic)	176	168 [0,183]	170 [0,184]	176 [0,201]
(Women, Lymphatic)	213	**185 [0,201]**	**191 [0,207]**	213 [0,240]
(Men, Leukemia)	196	**179 [0,193]**	182 [0,197]	196 [0,222]
(Women, Leukemia)	210	223 [0,240]	226 [0,241]	209 [0,236]
(Men, Diabetes)	97	**82 [0,94]**	**83 [0,95]**	97 [0,121]
(Women, Diabetes)	56	46 [0,56]	49 [0,59]	56 [0,74]
(Men, Hypertensive)	171	157 [0,171]	159 [0,175]	172 [0,197]
(Women, Hypertensive)	116	104 [0,117]	107 [0,120]	117 [0,136]
(Men, Ischemic)	8	8 [0,12]	8 [0,12]	8 [0,17]
(Women, Ischemic)	21	16 [0,22]	16 [0,22]	22 [0,34]
(Men, Cerebrovascular)	9	9 [0,13]	9 [0,13]	9 [0,17]
(Women, Cerebrovascular)	7	7 [0,11]	7 [0,10]	8 [0,16]
(Men, Atherosclerosis)	131	128 [0,144]	130 [0,144]	131 [0,155]
(Women, Atherosclerosis)	103	95 [0,109]	99 [0,113]	104 [0,125]
(Men, Other Cardiovascular)	16	12 [0,16]	12 [0,17]	17 [0,28]
(Women, Other Cardiovascular)	7	7 [0,11]	7 [0,11]	7 [0,15]
(Men, Pneumonia)	85	80 [0,93]	81 [0,93]	86 [0,107]
(Women, Pneumonia)	84	86 [0,97]	87 [0,98]	85 [0,105]
(Men, COPD)	27	21 [0,27]	21 [0,28]	27 [0,40]
(Women, COPD)	104	**87 [0,99]**	**90 [0,102]**	105 [0,127]
(Men, Cirrhosis)	104	93 [0,106]	95 [0,106]	104 [0,126]
(Women, Cirrhosis)	184	169 [0,184]	171 [0,186]	185 [0,211]

Mortality causes with bold font stand for those causes with zero excesses according to BYM.

A.2.2. Observed and predicted zeroes for models in Section 4.4

Table A.2.: Observed zeroes for each data set and posterior predicted zeroes for each model. Values in the Obs. zeroes column correspond to the real observed zeroes for each data set. For the 5 columns on the right, numbers correspond to the posterior predictive median for this same quantity for each model run and the corresponding unilateral 95% posterior predictive interval. Bold fonts denote those combinations of models and data sets evidencing zero excesses according to their predictive intervals.

Sex & Cause	Obs. zeroes	BYM	FE	NFE	HGeo	ZGeo
(Men, All tumours)	4	2 [0,5]	4 [0,8]	4 [0,9]	4 [0,9]	3 [0,7]
(Women, All tumours)	7	6 [0,10]	7 [0,12]	7 [0,12]	6 [0,12]	8 [0,12]
(Men, Mouth)	216	**196 [0,211]**	216 [0,235]	216 [0,234]	215 [0,234]	210 [0,228]
(Men, Stomach)	105	**91 [0,103]**	105 [0,121]	105 [0,123]	102 [0,119]	102 [0,117]
(Women, Stomach)	150	137 [0,151]	150 [0,169]	150 [0,169]	149 [0,168]	148 [0,163]
(Men, Colorectal)	73	**58 [0,68]**	73 [0,87]	73 [0,88]	70 [0,85]	71 [0,85]
(Women, Colorectal)	74	72 [0,82]	74 [0,89]	74 [0,89]	74 [0,88]	77 [0,89]
(Men, Colon)	96	**79 [0,91]**	95 [0,113]	96 [0,111]	98 [0,113]	90 [0,106]
(Women, Colon)	98	91 [0,102]	98 [0,114]	98 [0,115]	93 [0,110]	101 [0,114]
(Men, Rectum)	201	**180 [0,196]**	201 [0,220]	201 [0,220]	199 [0,220]	199 [0,215]
(Women, Rectum)	234	223 [0,239]	234 [0,255]	234 [0,255]	235 [0,255]	231 [0,248]
(Men, Liver)	156	**138 [0,153]**	157 [0,173]	157 [0,175]	152 [0,171]	155 [0,171]
(Women, Liver)	188	176 [0,191]	188 [0,208]	188 [0,206]	185 [0,204]	188 [0,205]
(Women, Vesicle)	243	239 [0,255]	243 [0,263]	243 [0,262]	241 [0,261]	248 [0,265]

Sex & Cause	Obs. zeroes	BYM	FE	NFE	HGeo	ZGeo
(Men, Pancreas)	179	**163 [0,178]**	180 [0,197]	179 [0,199]	177 [0,196]	179 [0,195]
(Women, Pancreas)	194	180 [0,196]	194 [0,214]	193 [0,213]	196 [0,215]	191 [0,207]
(Men, Larynx)	214	**186 [0,203]**	214 [0,234]	214 [0,233]	206 [0,226]	214 [0,229]
(Men, Lung)	34	**25 [0,32]**	34 [0,45]	33 [0,45]	33 [0,44]	30 [0,39]
(Women, Lung)	199	188 [0,203]	198 [0,218]	200 [0,218]	198 [0,217]	196 [0,213]
(Women, Breast)	80	73 [0,85]	80 [0,94]	80 [0,95]	78 [0,93]	81 [0,93]
(Women, Uterus)	188	187 [0,201]	187 [0,208]	189 [0,209]	190 [0,210]	193 [0,208]
(Women, Ovary)	201	195 [0,210]	201 [0,220]	200 [0,221]	196 [0,217]	208 [0,224]
(Men, Prostate)	62	**51 [0,60]**	62 [0,76]	62 [0,75]	64 [0,77]	58 [0,71]
(Men, Bladder)	123	**104 [0,117]**	124 [0,141]	123 [0,141]	122 [0,139]	120 [0,135]
(Men, Lymphatic)	176	168 [0,183]	175 [0,195]	176 [0,194]	174 [0,194]	179 [0,195]
(Women, Lymphatic)	213	**185 [0,201]**	213 [0,232]	213 [0,232]	211 [0,232]	208 [0,225]
(Men, Leukemia)	196	**179 [0,193]**	196 [0,217]	196 [0,216]	195 [0,216]	192 [0,210]
(Women, Leukemia)	210	223 [0,240]	210 [0,230]	210 [0,231]	211 [0,230]	231 [0,247]
(Men, Diabetes)	97	**82 [0,94]**	97 [0,114]	97 [0,113]	95 [0,111]	95 [0,109]
(Women, Diabetes)	56	46 [0,56]	56 [0,70]	56 [0,69]	58 [0,72]	52 [0,63]
(Men, Hypertensive)	171	157 [0,171]	171 [0,191]	170 [0,192]	172 [0,190]	167 [0,183]
(Women, Hypertensive)	116	104 [0,117]	116 [0,133]	116 [0,133]	118 [0,136]	113 [0,128]
(Men, Ischemic)	8	8 [0,12]	7 [0,14]	8 [0,14]	7 [0,13]	9 [0,14]
(Women, Ischemic)	21	16 [0,22]	20 [0,29]	20 [0,30]	20 [0,29]	20 [0,27]
(Men, Cerebrovascular)	9	9 [0,13]	9 [0,15]	9 [0,15]	9 [0,15]	10 [0,14]

Sex & Cause	Obs. zeroes	BYM	FE	NFE	HGeo	ZGeo
(Women, Cerebrovascular)	7	7 [0,11]	7 [0,12]	7 [0,12]	6 [0,11]	8 [0,12]
(Men, Atherosclerosis)	131	128 [0,144]	131 [0,151]	130 [0,149]	137 [0,156]	133 [0,148]
(Women, Atherosclerosis)	103	95 [0,109]	103 [0,121]	103 [0,119]	114 [0,132]	100 [0,113]
(Men, Other Cardiovascular)	16	12 [0,16]	16 [0,24]	16 [0,23]	16 [0,24]	14 [0,20]
(Women, Other Cardiovascular)	7	7 [0,11]	7 [0,12]	7 [0,12]	6 [0,11]	8 [0,12]
(Men, Pneumonia)	85	80 [0,93]	84 [0,98]	84 [0,100]	83 [0,98]	88 [0,99]
(Women, Pneumonia)	84	86 [0,97]	84 [0,100]	83 [0,100]	86 [0,103]	89 [0,101]
(Men, COPD)	27	21 [0,27]	27 [0,37]	27 [0,37]	28 [0,37]	25 [0,33]
(Women, COPD)	104	**87 [0,99]**	104 [0,121]	104 [0,120]	105 [0,122]	100 [0,115]
(Men, Cirrhosis)	104	93 [0,106]	104 [0,121]	103 [0,120]	103 [0,121]	101 [0,115]
(Women, Cirrhosis)	184	169 [0,184]	184 [0,205]	184 [0,204]	184 [0,203]	180 [0,199]

Mortality causes with bold font stand for those causes with zero excesses according to BYM.

A.2.3. Model selection criteria (DIC) for models in Section 4.4

Table A.3.: DICs for all models and data sets with their corresponding deviances and number of effective parameters.

	BYM			FE			NFE			HGeo			ZGeo		
Sex & Cause	D	pD	DIC	D	pD	DIC	D	pD	DIC	D	pD	DIC	D	pD	DIC
(Men, All tumours)	3623.3	265.1	3888.4	3629.7	267.1	3896.8	3624.5	267.9	3892.4	3623.2	265.8	3889	3626.4	266.8	3893.2
(Women, All tumours)	3333.8	173.5	3507.3	3338.9	175.1	3513.9	3335.5	175.4	3510.9	3333.8	173.7	3507.5	3334.7	172.5	3507.2
(Men, Mouth)	1575	70.4	1645.5	1605.8	58.2	1664	1578.2	66.4	1644.5	1594.2	58.1	1652.3	1583.6	69.2	1652.8
(Men, Stomach)	2112.5	100.7	2213.2	2126.9	96.7	2223.7	2116	98.1	2214.1	2120.2	95.4	2215.6	2114.5	100	2214.5
(Women, Stomach)	1854.2	54.1	1908.3	1867.3	52.1	1919.3	1854.9	54	1908.8	1858.7	50.8	1909.5	1856.9	53.9	1910.8
(Men, Colorectal)	2343.3	87.9	2431.1	2355.6	85.3	2440.9	2342.3	84.7	2427	2344.1	83.9	2428	2344.7	86.6	2431.3
(Women, Colorectal)	2260.9	80.4	2341.3	2269.7	81.2	2350.8	2261.8	81.2	2343	2263.8	79.9	2343.6	2263.8	81.2	2345
(Men, Colon)	2200.5	73.6	2274.1	2208.8	68.2	2277	2199.8	72.9	2272.7	2200.2	67	2267.2	2205.1	73.5	2278.6
(Women, Colon)	2110.6	63.6	2174.3	2114.8	66	2180.8	2114.6	61	2175.5	2111.8	63.2	2175	2114.2	62.5	2176.7
(Men, Rectum)	1580.5	60.1	1640.6	1584.6	49.8	1634.4	1585.4	53.4	1638.8	1589	49	1637.9	1578.3	54.7	1633
(Women, Rectum)	1458.2	52.5	1510.6	1481.6	43.6	1525.2	1461.8	50.6	1512.4	1470.2	42.8	1513	1465.1	52.6	1517.8
(Men, Liver)	1843.7	123.6	1967.3	1865.4	111.8	1977.3	1848.6	116.5	1965.1	1862.1	109.6	1971.7	1848.3	117.4	1965.7
(Women, Liver)	1649	67.8	1716.7	1670.3	61.9	1732.2	1651.6	66	1717.6	1659.4	59.4	1718.9	1656.7	66.9	1723.5
(Women, Vesicle)	1376.7	45.2	1421.8	1385.9	41.2	1427.1	1378.1	45.9	1424	1383.3	39.3	1422.6	1378.2	45.3	1423.5
(Men, Pancreas)	1701.1	56.6	1757.8	1702.9	51.9	1754.8	1704.2	53.1	1757.2	1704.9	50.6	1755.5	1699.1	54.1	1753.2
(Women, Pancreas)	1644.4	33.4	1677.7	1656.5	24.9	1681.4	1647.4	30.9	1678.3	1654.5	24.4	1678.9	1645.8	32.4	1678.2
(Men, Larynx)	1596.3	90.9	1687.2	1616.5	79	1695.5	1598.3	80.6	1678.8	1612.7	78.5	1691.2	1596.5	83.3	1679.8
(Men, Lung)	2869.3	190.1	3059.3	2889.2	186.2	3075.4	2871.5	188.7	3060.2	2884.4	185	3069.4	2871.8	188.7	3060.5

	BYM			FE			NFE			HGeo			ZGeo		
Sex & Cause	D	pD	DIC	D	pD	DIC	D	pD	DIC	D	pD	DIC	D	pD	DIC
(Women, Lung)	1590.2	66	1656.1	1613	56.1	1669.1	1595.3	62.2	1657.4	1605.9	55.7	1661.6	1594.5	63.7	1658.2
(Women, Breast)	2333.8	95.9	2429.7	2341	93.8	2434.9	2336.3	95.5	2431.8	2341.5	92.3	2433.7	2333.9	95.6	2429.5
(Women, Uterus)	1645.2	66.2	1711.4	1662.8	59.2	1722	1645	68.6	1713.5	1656.3	58.3	1714.6	1650.4	68	1718.4
(Women, Ovary)	1554.9	27.9	1582.8	1559.7	27.3	1586.9	1556.5	27	1583.5	1560.7	26.1	1586.9	1552.1	27.3	1579.5
(Men, Prostate)	2489.2	106.8	2596.1	2500	103.3	2603.3	2487.2	108	2595.3	2494	101.7	2595.7	2492.2	108.1	2600.2
(Men, Bladder)	2041.9	105.8	2147.8	2066.6	94.9	2161.5	2045.4	101.2	2146.6	2060.4	93.7	2154.1	2045.9	102.1	2148.1
(Men, Lymphatic)	1668.9	47.2	1716.1	1678.9	43.4	1722.3	1672.2	45.3	1717.5	1679	41.7	1720.8	1668.9	45.7	1714.6
(Women, Lymphatic)	1548.3	31.4	1579.7	1548.9	24.1	1573	1544.6	26.5	1571.1	1548.5	23.7	1572.2	1544.3	28.9	1573.2
(Men, Leukemia)	1608.8	25.1	1633.9	1619.4	22.7	1642.1	1607.8	23.5	1631.3	1611.3	21.3	1632.7	1611.6	25	1636.6
(Women, Leukemia)	1447.5	18.9	1466.5	1446.8	21.7	1468.5	1444.4	21.4	1465.8	1445	20.4	1465.4	1455.6	19.5	1475.1
(Men, Diabetes)	2165.2	115.7	2280.9	2181.9	109.2	2291.1	2168.9	112.5	2281.4	2173.4	109.9	2283.3	2172.9	113	2286
(Women, Diabetes)	2545.6	167.6	2713.1	2565.9	159.1	2725	2547.9	166.8	2714.7	2559.8	158.2	2718	2549.7	165.2	2714.9
(Men, Hypertensive)	1748.9	76.9	1825.7	1774	66.6	1840.6	1749.3	75.1	1824.4	1762.8	65.7	1828.5	1754.4	76.7	1831.1
(Women, Hypertensive)	2064.8	162.5	2227.3	2099.3	146.4	2245.7	2073.7	159.9	2233.5	2094.7	145.1	2239.7	2072.3	161.2	2233.5
(Men, Ischemic)	3217.4	250.9	3468.3	3223.8	253.6	3477.4	3218.4	252.4	3470.8	3219.1	251.7	3470.8	3218.6	251.4	3470
(Women, Ischemic)	2999	272.8	3271.9	3006.8	271	3277.9	3000.5	272.3	3272.8	3004.3	268.4	3272.7	3000.4	271	3271.4
(Men, Cerebrovascular)	3208	274.9	3482.9	3215.7	276.2	3491.9	3206.1	276.3	3482.4	3210.9	273.7	3484.6	3207.1	274.9	3482
(Women, Cerebrovascular)	3323.8	310.3	3634	3331.5	315.1	3646.6	3325.9	312.5	3638.4	3326	312.9	3638.9	3326.4	310.6	3637
(Men, Atherosclerosis)	1945.2	258.6	2203.8	2006.2	220.8	2227.1	1948.1	253.2	2201.4	2007.8	220	2227.7	1959.2	254.3	2213.5
(Women, Atherosclerosis)	2184.2	299.3	2483.5	2256.3	260.6	2516.9	2189.7	296.5	2486.2	2270	261	2531	2188.1	296	2484.1
(Men, Other Cardiovascular)	3132.1	251.4	3383.5	3143.1	250.2	3393.3	3130.6	252.2	3382.8	3137.6	247.5	3385.2	3132.7	250.7	3383.4
(Women, Other Cardiovascular)	3299.6	304.5	3604.1	3298.3	306.5	3604.8	3301.6	306.2	3607.7	3298	304.6	3602.6	3298	304.6	3602.6
(Men, Pneumonia)	2187.6	117.9	2305.5	2195.1	114.3	2309.4	2193.1	116.2	2309.3	2195.4	114.2	2309.6	2189.4	116.8	2306.2

	BYM			FE			NFE			HGeo			ZGeo		
Sex & Cause	D	pD	DIC	D	pD	DIC	D	pD	DIC	D	pD	DIC	D	pD	DIC
(Women, Pneumonia)	2172.1	134.6	2306.7	2186.4	127.6	2314.1	2172.9	135.2	2308.1	2185.2	126.4	2311.6	2173.8	134.8	2308.6
(Men, COPD)	2857.5	189.3	3046.8	2871.4	187.2	3058.6	2860.2	188.9	3049.1	2866.2	185.8	3052	2859.2	188.2	3047.4
(Women, COPD)	2177.5	159.6	2337.1	2205.2	142.1	2347.4	2183.8	157.5	2341.2	2198.7	141.2	2339.9	2188.9	153.9	2342.8
(Men, Cirrhosis)	2124	157.7	2281.7	2150.7	148.4	2299.1	2131.6	156.3	2287.8	2141.5	147.9	2289.4	2128.4	156.7	2285.1
(Women, Cirrhosis)	1675.5	150.7	1826.2	1709.9	127.5	1837.5	1683.2	144.6	1827.8	1701.8	128	1829.8	1686	146.4	1832.4

Mortality causes with bold font stand for those causes with zero excesses according to BYM.

A.2.4. Estimates of γ parameters

Table A.4.: Posterior means and 95% credible intervals for parameter γ in the model NFE for each data set.

Sex & Cause	γ
(Men, All tumours)	-0.57 [-1.78 - 0.8]
(Women, All tumours)	-0.08 [-0.98 - 0.97]
(Men, Mouth)	-0.38 [-0.64 - -0.11]
(Men, Stomach)	-0.34 [-0.61 - -0.01]
(Women, Stomach)	-0.23 [-0.51 - 0.02]
(Men, Colorectal)	-0.44 [-0.77 - -0.1]
(Women, Colorectal)	-0.05 [-0.36 - 0.28]
(Men, Colon)	-0.4 [-0.71 - -0.11]
(Women, Colon)	-0.16 [-0.45 - 0.13]
(Men, Rectum)	-0.37 [-0.63 - -0.13]
(Women, Rectum)	-0.19 [-0.44 - 0.06]
(Men, Liver)	-0.4 [-0.67 - -0.12]
(Women, Liver)	-0.2 [-0.46 - 0.05]
(Women, Vesicle)	-0.07 [-0.32 - 0.2]
(Men, Pancreas)	-0.28 [-0.55 - -0.01]
(Women, Pancreas)	-0.23 [-0.49 - 0.02]
(Men, Larynx)	-0.55 [-0.82 - -0.29]
(Men, Lung)	-0.56 [-1.02 - -0.1]
(Women, Lung)	-0.23 [-0.49 - 0.05]
(Women, Breast)	-0.19 [-0.51 - 0.11]
(Women, Uterus)	-0.03 [-0.26 - 0.22]
(Women, Ovary)	-0.09 [-0.34 - 0.14]
(Men, Prostate)	-0.35 [-0.72 - -0.01]
(Men, Bladder)	-0.45 [-0.76 - -0.16]
(Men, Lymphatic)	-0.14 [-0.38 - 0.12]
(Women, Lymphatic)	-0.47 [-0.74 - -0.22]
(Men, Leukemia)	-0.28 [-0.51 - -0.02]
(Women, Leukemia)	0.22 [-0.01 - 0.46]

Sex & Cause	γ
(Men, Diabetes)	-0.38 [-0.69 - -0.06]
(Women, Diabetes)	-0.36 [-0.73 - 0.01]
(Men, Hypertensive)	-0.25 [-0.51 - 0.01]
(Women, Hypertensive)	-0.27 [-0.56 - 0.05]
(Men, Ischemic)	0.1 [-0.67 - 1.04]
(Women, Ischemic)	-0.43 [-0.98 - 0.15]
(Men, Cerebrovascular)	0 [-0.77 - 0.86]
(Women, Cerebrovascular)	-0.02 [-0.92 - 0.94]
(Men, Atherosclerosis)	-0.08 [-0.4 - 0.27]
(Women, Atherosclerosis)	-0.23 [-0.57 - 0.1]
(Men, Other Cardiovascular)	-0.49 [-1.09 - 0.16]
(Women, Other Cardiovascular)	0.06 [-0.86 - 1.12]
(Men, Pneumonia)	-0.11 [-0.44 - 0.21]
(Women, Pneumonia)	0.05 [-0.29 - 0.37]
(Men, COPD)	-0.45 [-0.95 - 0.06]
(Women, COPD)	-0.43 [-0.71 - -0.13]
(Men, Cirrhosis)	-0.28 [-0.57 - 0.04]
(Women, Cirrhosis)	-0.33 [-0.61 - -0.04]

Mortality causes with bold font stand for those causes with zero excesses according to BYM.

A.2.5. Choropleth maps for all models in Section 4.4 for all causes

Choropleth maps for all models in Section 4.4 for all causes can be viewed online at https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7819.

A.3. Markdown document with all the analysis carried out

A.3.1. Execution of models in WinBUGS using the library R2WinBUGS

Load libraries, data and cartography

```
# Working directory
DirMain = " "  # Set an appropriate directory
setwd(DirMain)
# Load library and data
library(R2WinBUGS)
load("datos/OE.rdata")
load("VR.rdata")
# For running the models in parallel
# calls to WinBUGS
source("Pbugs.0.4.4.r")
# Load cartography
Cvalenciana <- dget("datos/Cvalenciana.txt")
# Total number of diseases
ndiseases <- 46
# Total number of municipalities
nareas <- 540
# Total number of observed and expected
# cases during the whole period of study
Obs <- list()
Exp <- list()
for (i in 1:ndiseases) {
    Obs[[i]] <- apply(Obs2[[i]], 1, sum)
    Exp[[i]] <- apply(Esp[[i]], 1, sum)
}
```

BYM model

```
# BYM model, WinBUGS code
model.BYM <- function() {
    for (i in 1:n) {
        O[i] ~ dpois(mu[i])
        # Modeling of the mean for each
        # municipality
        log(mu[i]) <- log(E[i]) + m + sd.phi *
            phi[i] + sd.theta * theta[i]
        # SMR for each municipality
        SMR[i] <- exp(m + sd.phi * phi[i] +
            sd.theta * theta[i])
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dpois(mu[i])
        pred.equal.0[i] <- equals(O.pred[i],  0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
    # Prior distribution for the mean risk
    # for all municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
}
```

```
# Run BYM model for each disease
for (i in 1:ndiseases) {
    # Working directory
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
    # Initial values
    initials <- function() {
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas))
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "mu", "zero.pred")
    # Calls to WinBUGS
    t.ResulBYM <- system.time(ResulBYM <- Pbugs(data =
        data,  inits = initials, parameters.to.save =
        param,  model.file = model.BYM, n.chains = 3,
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulBYM, t.ResulBYM, file = "ResulBYM.Rdata")
    setwd <- paste(DirMain)
}
```

Naive ZIP model

```
# Naive ZIP model, WinBUGS code
model.ZIP <- function() {
```

```
    for (i in 1:n) {
        O[i] ~ dpois(mu[i])
        Z[i] ~ dbern(p)
        # Modeling of the mean for each
        # municipality
        log(mu[i]) <- log(E[i]) + m + sd.phi *
            phi[i] + sd.theta * theta[i] -
            1000 * (1 - Z[i])
        # SMR for each municipality
        SMR[i] <- exp(m + sd.phi * phi[i] +
            sd.theta * theta[i] - 1000 *
            (1 - Z[i]))
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dpois(mu[i])
        pred.equal.0[i] <- equals(O.pred[i], 0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
    # Prior distribution for the mean risk
    # for all municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
    # Prior distribution for p
```

```
    p ~ dunif(0, 1)
}
# Run ZIP model for each disease
for (i in 1:ndiseases) {
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
    # Initial values
    initials <- function() {
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas),
            Z = as.numeric(Obs[[i]] > 0))
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "mu", "p", "zero.pred")
    # Calls to WinBUGS
    t.ResulZIP <- system.time(ResulZIP <- Pbugs(data =
        data,  inits = initials, parameters.to.save =
        param,  model.file = model.ZIP, n.chains = 3,
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulZIP, t.ResulZIP, file = "ResulZIP.Rdata")
    setwd <- paste(DirMain)
}
```

Naive Hurdle model

```
# Naive Hurdle model, WinBUGS code
model.Hurdle <- function() {
    # Modeling using the zero trick
    c <- 10000
    for (i in 1:n) {
        zeros[i] <- 0
        zeros[i] ~ dpois(zeros.mean[i])
        zeros.mean[i] <- -L[i] + c
        Z[i] <- step(O[i] - 1)
        # Expression of the log-likelihood por i
        L[i] <- (1 - Z[i]) * log(1 - p) +
            Z[i] * (log(p) + O[i] * log(lambda[i]) -
                lambda[i] - logfact(O[i]) -
                log(1 - exp(-lambda[i])))
        # Modeling of the mean Poisson for each
        # municipality
        log(lambda[i]) <- log(E[i]) + m +
            sd.phi * phi[i] + sd.theta *
            theta[i]
        # SMR for each municipality
        SMR[i] <- (p * lambda[i]/(1 - exp(-lambda[i])))
        /E[i]
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dbern(p)
        pred.equal.0[i] <- equals(O.pred[i], 0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
```

```
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
    # Prior distribution for the mean risk
    # all every municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
    # Prior distribution for p
    p ~ dunif(0, 1)
}
# Run Hurdle model for each disease
for (i in 1:ndiseases) {
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
    # Initial values
    initials <- function() {
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas))
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "lambda", "p", "zero.pred")
    # Calls to WinBUGS
    t.ResulHurdle <- system.time(ResulHurdle <- Pbugs(
data =  data,  inits = initials, parameters.to.save =
```

```
        param,  model.file = model.Hurdle, n.chains = 3,
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulHurdle, t.ResulHurdle, file =
        "ResulHurdle.Rdata")
    setwd <- paste(DirMain)
}
```

FE Hurdle model

```
# FE Hurdle model, WinBUGS code
model.HFE <- function() {
    # Modeling using the zero trick
    c <- 10000
    for (i in 1:n) {
        zeros[i] <- 0
        zeros[i] ~ dpois(zeros.mean[i])
        zeros.mean[i] <- (-L[i] + c)
        Z[i] <- step(O[i] - 1)
        # Expression of the log-likelihood por i
        L[i] <- (1 - Z[i]) * log(1 - p[i]) +
            Z[i] * (log(p[i]) + O[i] * log(lambda[i]) -
                lambda[i] - logfact(O[i]) -
                log(1 - exp(-lambda[i])))
        # Modeling of the mean Poisson for each
        # municipality
        log(lambda[i]) <- log(E[i]) + m +
            sd.phi * phi[i] + sd.theta *
            theta[i]
        # Modeling p for each municipality
        logit(p[i]) <- alpha + beta * LE[i]
        # SMR for each municipality
```

```
        SMR[i] <- (p[i] * lambda[i]/(1 -
            exp(-lambda[i])))/E[i]
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dbern(p[i])
        pred.equal.0[i] <- equals(O.pred[i], 0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
    # Prior distribution for the mean risk
    # for all municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
    # Prior distribution for the parameters
    # logistic regression
    alpha ~ dflat()
    beta ~ dflat()
}
# Run FE Hurdle model for each disease
for (i in 1:ndiseases) {
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
    # Initial values
    initials <- function() {
```

```
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas),
            alpha = rnorm(1, 0, 0.1), beta = rnorm(1,
                0, 0.1))
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], LE = log(Exp[[i]]) -
            mean(log(Exp[[i]])), map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "lambda", "p", "alpha", "beta", "zero.pred")
    # Calls to WinBUGS
    t.ResulHFE <- system.time(ResulHFE <- Pbugs(data =
        data,  inits = initials, parameters.to.save =
        param,  model.file = model.HFE, n.chains = 3,
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulHFE, t.ResulHFE, file =
        "ResulHFE.Rdata")
    setwd <- paste(DirMain)
}
```

NFE Hurdle model

```
# NFE Hurdle model, WinBUGS code
model.HNFE <- function() {
    # Modeling using the zero trick
    c <- 10000
    for (i in 1:n) {
```

```
        zeros[i] <- 0
        zeros[i] ~ dpois(zeros.mean[i])
        zeros.mean[i] <- (-L[i] + c)
        Z[i] <- step(O[i] - 1)
        # Expression of the log-likelihood por i
        L[i] <- (1 - Z[i]) * log(1 - p[i]) +
            Z[i] * (log(p[i]) + O[i] * log(lambda[i]) -
                lambda[i] - logfact(O[i]) -
                log(1 - exp(-lambda[i])))
        # Modeling of the mean Poisson for each
        # municipality
        log(lambda[i]) <- log(E[i]) + m +
            sd.phi * phi[i] + sd.theta *
            theta[i]
        # Modeling p for each municipality
        logit(p[i]) <- logit(1 - exp(-lambda[i])) +
            gamma
        # SMR for each municipality
        SMR[i] <- (p[i] * lambda[i]/(1 -
            exp(-lambda[i])))/E[i]
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dbern(p[i])
        pred.equal.0[i] <- equals(O.pred[i], 0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
```

```
    # Prior distribution for the mean risk
    # for all municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
    # Prior distribution for the parameters
    # logistic regression
    gamma ~ dflat()
}
# Run NFE Hurdle model for each disease
for (i in 1:ndiseases) {
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
    # Initial values
    initials <- function() {
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas),
            gamma = rnorm(1, 0, 0.1))
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "lambda", "p", "gamma", "zero.pred")
    # Calls to WinBUGS
    t.ResulHNFE <- system.time(ResulHNFE <- Pbugs(data =
        data,  inits = initials, parameters.to.save =
        param,  model.file = model.HNFE, n.chains = 3,
```

```
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulHNFE, t.ResulHNFE, file = "ResulHNFE.Rdata")
    setwd <- paste(DirMain)
}
```

HGeo model

```
# HGeo model, WinBUGS code
model.HGeo <- function() {
    # Modeling using the zero trick
    c <- 10000
    for (i in 1:n) {
        zeros[i] <- 0
        zeros[i] ~ dpois(zeros.mean[i])
        zeros.mean[i] <- (-L[i] + c)
        Z[i] <- step(O[i] - 1)
        # Expression of the log-likelihood por i
        L[i] <- (1 - Z[i]) * log(1 - p[i]) +
            Z[i] * (log(p[i]) + O[i] * log(lambda[i]) -
                lambda[i] - logfact(O[i]) -
                log(1 - exp(-lambda[i])))
        # Modeling of the mean Poisson for each
        # municipality
        log(lambda[i]) <- log(E[i]) + m +
            sd.phi * phi[i] + sd.theta *
            theta[i]
        # Modeling p for each municipality
        p[i] <- 1 - pow((1 - pi), E[i])
        # SMR for each municipality
        SMR[i] <- (p[i] * lambda[i]/(1 -
            exp(-lambda[i])))/E[i]
```

```
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dbern(p[i])
        pred.equal.0[i] <- equals(O.pred[i], 0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
    # Prior distribution for the mean risk
    # for all municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
    # Prior distribution for pi
    pi ~ dunif(0, 1)
}
# Run HGeo model for each disease
for (i in 2:ndiseases) {
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
    # Initial values
    initials <- function() {
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas),
            pi = runif(1, 0, 1))
```

```
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "lambda", "p", "pi", "zero.pred")
    # Calls to WinBUGS
    t.ResulHGeo <- system.time(ResulHGeo <- Pbugs(data =
        data,  inits = initials, parameters.to.save =
        param,  model.file = model.HGeo, n.chains = 3,
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulHGeo, t.ResulHGeo, file = "ResulHGeo.Rdata")
    setwd <- paste(DirMain)
}
```

ZGeo model

```
# ZIP model, WinBUGS code
model.ZGeo <- function() {
    for (i in 1:n) {
        O[i] ~ dpois(mu[i])
        Z[i] ~ dbern(p[i])
        # Modeling p for each municipality
        p[i] <- 1 - pow((1 - pi), E[i])
        # Modeling of the mean for each
        # municipality
        log(mu[i]) <- log(E[i]) + m + sd.phi *
            phi[i] + sd.theta * theta[i] -
            1000 * (1 - Z[i])
```

```
        lambda[i] <- E[i] * exp(m + sd.phi *
            phi[i] + sd.theta * theta[i])
        # SMR for each municipality
        SMR[i] <- exp(m + sd.phi * phi[i] +
            sd.theta * theta[i] - 1000 *
            (1 - Z[i]))
        # Prior distribution for the non-spatial
        # effect
        theta[i] ~ dnorm(0, 1)
        # Predictive distribution
        O.pred[i] ~ dpois(mu[i])
        pred.equal.0[i] <- equals(O.pred[i], 0)
    }
    # Predictive distribution for the number
    # of zeroes
    zero.pred <- sum(pred.equal.0[])
    # Prior distribution for the spatial
    # effect
    phi[1:n] ~ car.normal(map[], w[], nvec[], 1)
    # Prior distribution for the mean risk
    # for all municipalities
    m ~ dflat()
    # Prior distribution for the standard
    # deviations of the random effects
    sd.theta ~ dunif(0, 5)
    sd.phi ~ dunif(0, 5)
    # Prior distribution for pi
    pi ~ dunif(0, 1)
}
# Run ZGeo model for each disease
for (i in 1:ndiseases) {
    setwd(paste(DirMain, "/resul/", LabelsCausas[i],
        sep = ""))
```

```
    # Initial values
    initials <- function() {
        list(m = rnorm(1, 0, 0.1), sd.theta = runif(1,
            0, 1), sd.phi = runif(1, 0, 1),
            theta = rnorm(nareas), phi = rnorm(nareas),
            Z = as.numeric(Obs[[i]] > 0),
            pi = runif(1, 0, 1))
    }
    # Data
    data <- list(n = nareas, O = Obs[[i]],
        E = Exp[[i]], map = Cvalenciana$map,
        w = Cvalenciana$w, nvec = Cvalenciana$nvec)
    # Variables to retrieve
    param <- c("sd.phi", "sd.theta", "SMR",
        "mu", "lambda", "p", "pi", "zero.pred")
    # Calls to WinBUGS
    t.ResulZGeo <- system.time(ResulZGeo <- Pbugs(data =
        data,  inits = initials, parameters.to.save =
        param,  model.file = model.ZGeo, n.chains = 3,
        n.iter = 50000, n.burnin = 5000,
        DIC = F, working.directory = getwd()))
    # Save results
    save(ResulZGeo, t.ResulZGeo, file = "ResulZGeo.Rdata")
    setwd <- paste(DirMain)
}
```

A.3.2. Comparison: observed zeroes for each data set vs. posterior predicted zeroes for each model (Tables A.1 and A.2)

```
# Load libraries
library(xtable)
```

```
library(pander)
library(rmarkdown)
library(knitr)
# Posterior predicted zeroes for each
# model
zeros_BYM <- character()
zeros_ZIP <- character()
zeros_Hurdle <- character()
zeros_HFE <- character()
zeros_HNFE <- character()
zeros_HGeo <- character()
zeros_ZGeo <- character()
for (i in 1:ndiseases) {
    # Load WinBUGS results
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulBYM.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulZIP.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHurdle.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHFE.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHNFE.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHGeo.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulZGeo.Rdata", sep = ""))
    # Posterior predicted medians for zeroes
    # for each model run and corresponding
    # unilateral 95% posterior predictive
    # intervals
    zeros_BYM[i] <- paste0(round(summary
```

```
        (ResulBYM$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulBYM$sims.list$zero.pred,
            p = 0.95)), "]")
    zeros_ZIP[i] <- paste0(round(summary
        (ResulZIP$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulZIP$sims.list$zero.pred,
            p = 0.95)), "]")
    zeros_Hurdle[i] <- paste0(round(summary
        (ResulHurdle$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulHurdle$sims.list$zero.pred,
            p = 0.95)), "]")
    zeros_HFE[i] <- paste0(round(summary
        (ResulHFE$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulHFE$sims.list$zero.pred,
            p = 0.95)), "]")
    zeros_HNFE[i] <- paste0(round(summary
        (ResulHNFE$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulHNFE$sims.list$zero.pred,
            p = 0.95)), "]")
    zeros_HGeo[i] <- paste0(round(summary
        (ResulHGeo$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulHGeo$sims.list$zero.pred,
            p = 0.95)), "]")
    zeros_ZGeo[i] <- paste0(round(summary
        (ResulZGeo$sims.list$zero.pred)[3]),
        " [0,", round(quantile
        (ResulZGeo$sims.list$zero.pred,
```

```
            p = 0.95)), "]")
}
Disease <- c("(Men, All tumours)", "(Women, All tumours)",
    "(Men, Mouth)", "(Men, Stomach)", "(Women, Stomach)",
    "(Men, Colorectal)", "(Women, Colorectal)",
    "(Men, Colon)", "(Women, Colon)", "(Men, Rectum)",
    "(Women, Rectum)", "(Men, Liver)", "(Women, Liver)",
    "(Women, Vesicle)", "(Men, Pancreas)",
    "(Women, Pancreas)", "(Men, Larynx)",
    "(Men, Lung)", "(Women, Lung)", "(Women, Breast)",
    "(Women, Uterus)", "(Women, Ovary)",
    "(Men, Prostate)", "(Men, Bladder)",
    "(Men, Lymphatic)", "(Women, Lymphatic)",
    "(Men, Leukemia)", "(Women, Leukemia)",
    "(Men, Diabetes)", "(Women, Diabetes)",
    "(Men, Hypertensive)", "(Women, Hypertensive)",
    "(Men, Ischemic)", "(Women, Ischemic)",
    "(Men, Cerebrovascular)", "(Women, Cerebrovascular)",
    "(Men, Atherosclerosis)", "(Women, Atherosclerosis)",
    "(Men, Other Cardiovascular)",
    "(Women, Other Cardiovascular)",
    "(Men, Pneumonia)", "(Women, Pneumonia)",
    "(Men, COPD)", "(Women, COPD)", "(Men, Cirrhosis)",
    "(Women, Cirrhosis)")
Table <- cbind(Disease, unlist(lapply(Obs,
    function(x) {
        sum(x == 0)
    })), zeros_BYM, zeros_ZIP, zeros_Hurdle,
    zeros_HFE, zeros_HNFE, zeros_HGeo, zeros_ZGeo)
colnames(Table) <- c("Sex & Cause", "Obs. zeroes",
    "BYM", "ZIP", "Hurdle", "HFE", "HNFE",
    "HGeo", "ZGeo")
kable(Table, split.table = Inf, row.names = FALSE,
```

```
    align = "c", caption = "Observed zeroes for each
      data set and posterior predicted zeroes for each
      model. Values in the Obs.zeroes column correspond to
      the real observed zeroes for each data set. For the
      5 columns on the right, numbers correspond to the
      posterior predictive median for this same quantity
      for each model run and the corresponding unilateral
      95% posterior predictive interval.")
```

A.3.3. DIC for each model (Table A.3)

```
# DIC BYM model
CalculaDIC_BYM <- function(Simu, O, E, save = FALSE) {
    mu <- t(apply(Simu$sims.list$SMR, 1,
        function(x) {
            x * E
        }))
    D <- apply(mu, 1, function(x) {
        -2 * sum(O * log(x) - x - lfactorial(O))
    })
    Dmedia <- mean(D)
    mumedia <- apply(Simu$sims.list$SMR,
        2, mean) * E
    DenMedia <- -2 * sum(O * log(mumedia) -
        mumedia - lfactorial(O))
    if (save == TRUE) {
        return(c(Dmedia, Dmedia - DenMedia,
            2 * Dmedia - DenMedia))
    }
    cat("D=", Dmedia, "pD=", Dmedia - DenMedia,
        "DIC=", 2 * Dmedia - DenMedia, "\n")
}
```

```
# DIC Hurdle FE, Hurdle NFE and HGeo
# models
CalculaDIC_Hurdle <- function(Simu, O, E,
    save = FALSE) {
    log.verosim <- matrix(nrow = Simu$n.sims,
        ncol = length(O))
    Z <- as.numeric(O > 0)
    for (j in 1:Simu$n.sims) {
        for (k in 1:length(O)) {
            if (Z[k] == 0) {
                log.verosim[j, k] <- log(1 -
                  Simu$sims.list$p[j, k])
            }
            if (Z[k] == 1) {
                log.verosim[j, k] <- log(Simu$sims.list$
                  p[j,  k]) + O[k] * log(Simu$sims.list$
                  lambda[j,  k]) - Simu$sims.list$lambda[j,
                  k] - lfactorial(O[k]) -
                  log(1 - exp(-Simu$sims.list$lambda[j,
                    k]))
            }
        }
    }
    D <- -2 * apply(log.verosim, 1, sum)
    Dmedia <- mean(D)
    log.verosimMedia <- c()
    for (k in 1:length(O)) {
        if (Z[k] == 0) {
            log.verosimMedia[k] <- log(1 -
                Simu$mean$p[k])
        }
        if (Z[k] == 1) {
            log.verosimMedia[k] <- log(Simu$mean$p[k]) +
```

```
                O[k] * log(Simu$mean$lambda[k]) -
                Simu$mean$lambda[k] - lfactorial(O[k]) -
                log(1 - exp(-
                Simu$mean$lambda[k]))
        }
    }
    DenMedia <- -2 * sum(log.verosimMedia)
    if (save == TRUE) {
        return(c(Dmedia, Dmedia - DenMedia,
            2 * Dmedia - DenMedia))
    }
    cat("D=", Dmedia, "pD=", Dmedia - DenMedia,
        "DIC=", 2 * Dmedia - DenMedia, "\n")
}
# DIC ZGeo model
CalculaDIC_ZIP <- function(Simu, O, E, save = FALSE) {
    log.verosim <- matrix(nrow = Simu$n.sims,
        ncol = length(O))
    Z <- as.numeric(O > 0)
    for (j in 1:Simu$n.sims) {
        for (k in 1:length(O)) {
            if (Z[k] == 0) {
                log.verosim[j, k] <- log((1 -
                  Simu$sims.list$p[j, k]) +
                  Simu$sims.list$p[j, k] *
                    dpois(x = O[k], lambda =
                     Simu$sims.list$lambda[j,
                      k]))
            }
            if (Z[k] == 1) {
                log.verosim[j, k] <- log(Simu$sims.list$
                  p[j,  k] * dpois(x = O[k], lambda =
                  Simu$sims.list$lambda[j,
```

```
                    k]))
            }
        }
    }
    D <- -2 * apply(log.verosim, 1, sum)
    Dmedia <- mean(D)
    log.verosimMedia <- c()
    for (k in 1:length(O)) {
        if (Z[k] == 0) {
            log.verosimMedia[k] <- log((1 -
                Simu$mean$p[k]) + Simu$mean$p[k] *
                dpois(x = O[k], lambda =
                Simu$mean$lambda[k]))
        }
        if (Z[k] == 1) {
            log.verosimMedia[k] <- log(Simu$mean$p[k] *
                dpois(x = O[k], lambda =
                Simu$mean$lambda[k]))
        }
    }
    DenMedia <- -2 * sum(log.verosimMedia)
    if (save == TRUE) {
        return(c(Dmedia, Dmedia - DenMedia,
            2 * Dmedia - DenMedia))
    }
    cat("D=", Dmedia, "pD=", Dmedia - DenMedia,
        "DIC=", 2 * Dmedia - DenMedia, "\n")
}
DIC_BYM <- matrix(nrow = ndiseases, ncol = 3)
DIC_HFE <- matrix(nrow = ndiseases, ncol = 3)
DIC_HNFE <- matrix(nrow = ndiseases, ncol = 3)
DIC_HGeo <- matrix(nrow = ndiseases, ncol = 3)
DIC_ZGeo <- matrix(nrow = ndiseases, ncol = 3)
```

```
for (i in 1:ndiseases) {
    # Load WinBUGS results
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulBYM.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHFE.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHNFE.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHGeo.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulZGeo.Rdata", sep = ""))
    # DIC for each model and cause
    DIC_BYM[i, ] <- CalculaDIC_BYM(ResulBYM,
        Obs[[i]], Exp[[i]], save = TRUE)
    DIC_HFE[i, ] <- CalculaDIC_Hurdle(ResulHFE,
        Obs[[i]], Exp[[i]], save = TRUE)
    DIC_HNFE[i, ] <- CalculaDIC_Hurdle(ResulHNFE,
        Obs[[i]], Exp[[i]], save = TRUE)
    DIC_HGeo[i, ] <- CalculaDIC_Hurdle(ResulHGeo,
        Obs[[i]], Exp[[i]], save = TRUE)
    DIC_ZGeo[i, ] <- CalculaDIC_ZIP(ResulZGeo,
        Obs[[i]], Exp[[i]], save = TRUE)
}
Table <- cbind(Disease, round(DIC_BYM, 1),
    round(DIC_HFE, 1), round(DIC_HNFE, 1),
    round(DIC_HGeo, 1), round(DIC_ZGeo, 1))
colnames(Table) <- c("Disease", rep(c("D",
    "pD", "DIC"), 5))
rownames(Table) <- as.character(1:46)
cab <- c("Disease", rep(c("D", "pD", "DIC"),
    5))
Table2 <- rbind(cab, Table)
```

```
rownames(Table2) <- c("", rownames(Table))
addtorow <- list()
addtorow$pos <- list(0)
addtorow$command <- paste0("\\multicolumn{1}{c}{}",
    paste0(" & \\multicolumn{3}{c}{", c("BYM",
        "FE", "NFE", "HGeo", "ZGeo"), "}",
        collapse = ""), "\\\\")
print(xtable(Table2, caption = "DIC for each model.",
    align = rep("c", 17)), add.to.row = addtorow,
    include.colnames = F, hline.after = c(-1,
        0, 1, nrow(tabla2)), include.rownames = F,
    comment = FALSE)
```

A.3.4. Posterior distribution of γ in the Hurdle NFE model (Table A.4)

```
gamma <- character()
for (i in 1:ndiseases) {
    # Load WinBUGS NFE results
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHNFE.Rdata", sep = ""))
    # Posterior mean for gamma in the NFE
    # model and the corresponding 95%
    # posterior interval.
    gamma[i] <- paste0(round(ResulHNFE$summary["gamma",
        1], 2), " [", round(ResulHNFE$summary["gamma",
        3], 2), " - ", round(ResulHNFE$summary["gamma",
        7], 2), "]")
}
Table <- cbind(Disease, gamma)
colnames(Table) <- c("Sex & Cause", "$\\gamma$")
kable(Table, split.table = Inf, row.names = FALSE,
```

```
    align = "c", caption = "Posterior distribution of
      $\\gamma$ in the NFE model")
```

A.3.5. Choropleth maps for all models (Figures in A.2.5)

```
# Load libraries
library(RColorBrewer)
cuts_SMR <- c(0, 0.67, 0.8, 0.91, 1.1, 1.25,
    1.5)
palette <- brewer.pal(7, "BrBG")[7:1]
for (i in 1:ndiseases) {
    # Load WinBUGS results
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulBYM.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulZIP.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHurdle.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHFE.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHNFE.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulHGeo.Rdata", sep = ""))
    load(paste(getwd(), "/resul/", LabelsCausas[i],
        "/ResulZGeo.Rdata", sep = ""))
    # SMR estimates, BYM model
    plot(VR.cart, col = palette[findInterval
        (ResulBYM$mean$SMR,
        cuts_SMR)], main = paste0("BYM - ",
        Disease[i]))
```

```
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
    # SMR estimates, naive ZIP model
    plot(VR.cart, col = palette[findInterval
        (ResulZIP$mean$SMR,
        cuts_SMR)], main = paste0("ZIP - ",
        Disease[i]))
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
    # SMR estimates, naive Hurdle model
    plot(VR.cart, col = palette[findInterval
        (ResulHurdle$mean$SMR,
        cuts_SMR)], main = paste0("Hurdle - ",
        Disease[i]))
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
    # SMR estimates, FE model
    plot(VR.cart, col = palette[findInterval
        (ResulHFE$mean$SMR,
        cuts_SMR)], main = paste0("HFE - ",
        Disease[i]))
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
    # SMR estimates, NFE model
    plot(VR.cart, col = palette[findInterval
```

```
        (ResulHNFE$mean$SMR,
        cuts_SMR)], main = paste0("HNFE - ",
        Disease[i]))
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
    # SMR estimates, HGeo model
    plot(VR.cart, col = palette[findInterval
        (ResulHGeo$mean$SMR,
        cuts_SMR)], main = paste0("HGeo - ",
        Disease[i]))
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
    # SMR estimates, ZGeo model
    plot(VR.cart, col = palette[findInterval
        (ResulZGeo$mean$SMR,
        cuts_SMR)], main = paste0("ZGeo - ",
        Disease[i]))
    legend("bottomright", c("< 0.67", "0.67 - 0.80",
        "0.80 - 0.91", "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette, bty = "n")
}
```

B. Supplementary material to the paper: *"On the convenience of heteroscedasticity in highly multivariate disease mapping"*

B.1. Additional results

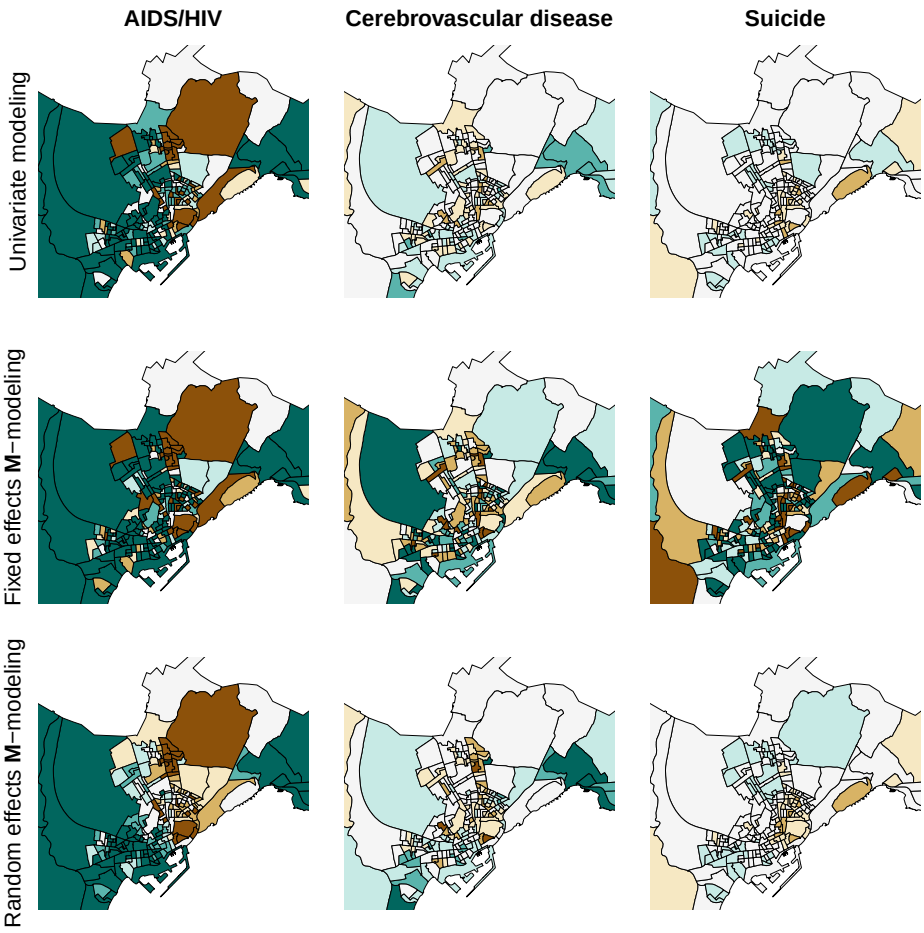


Figure B.1.: Graphical representation of the estimated risk in Alicante using traditional univariate modeling (BYM), the *fixed effects* $\boldsymbol{M}$*-modeling* and the *random effects* $\boldsymbol{M}$*-modeling* proposed in Botella-Rocamora et al. (2015).

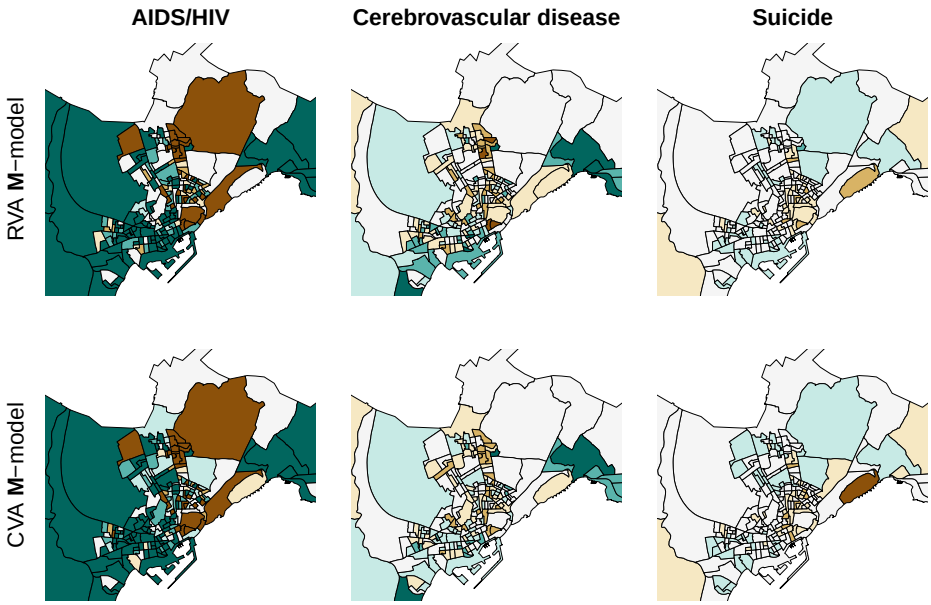


Figure B.2.: Graphical representation of the estimated risk in Alicante using the new variance-adaptive modeling proposals (RVA and CVA $\boldsymbol{M}$-modeling).

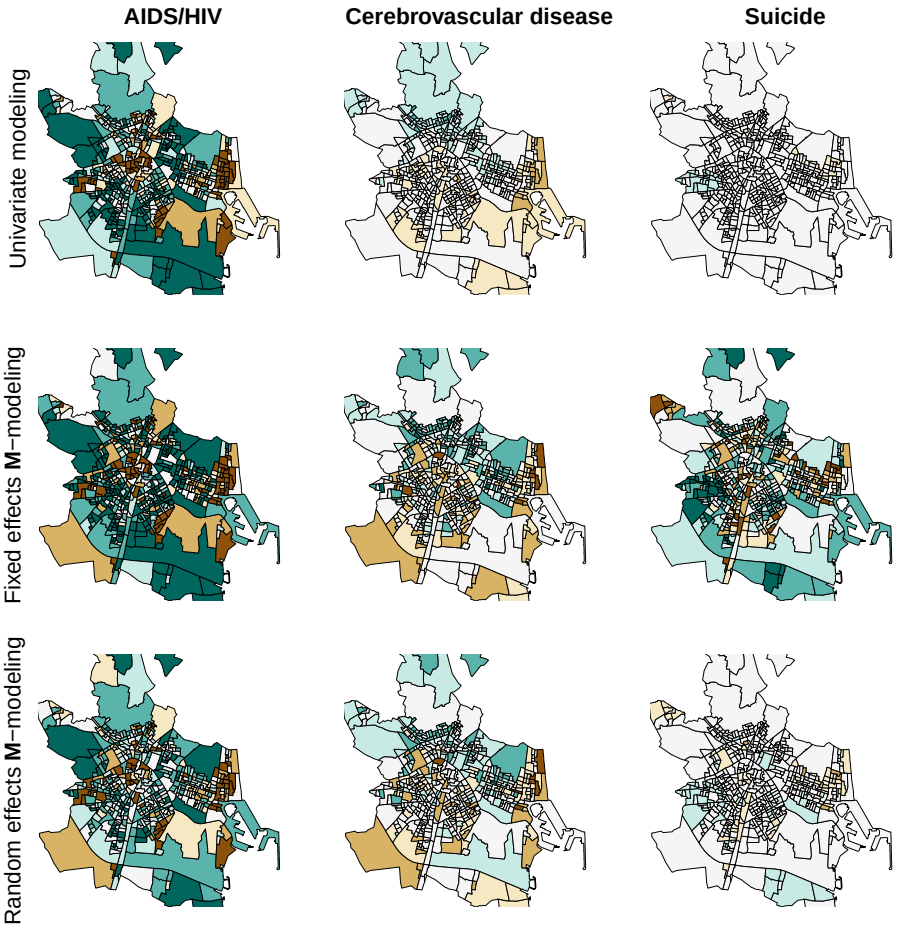


Figure B.3.: Graphical representation of the estimated risk in Valencia using traditional univariate modeling (BYM), the *fixed effects* $\boldsymbol{M}$*-modeling* and the *random effects* $\boldsymbol{M}$*-modeling* proposed in Botella-Rocamora et al. (2015).

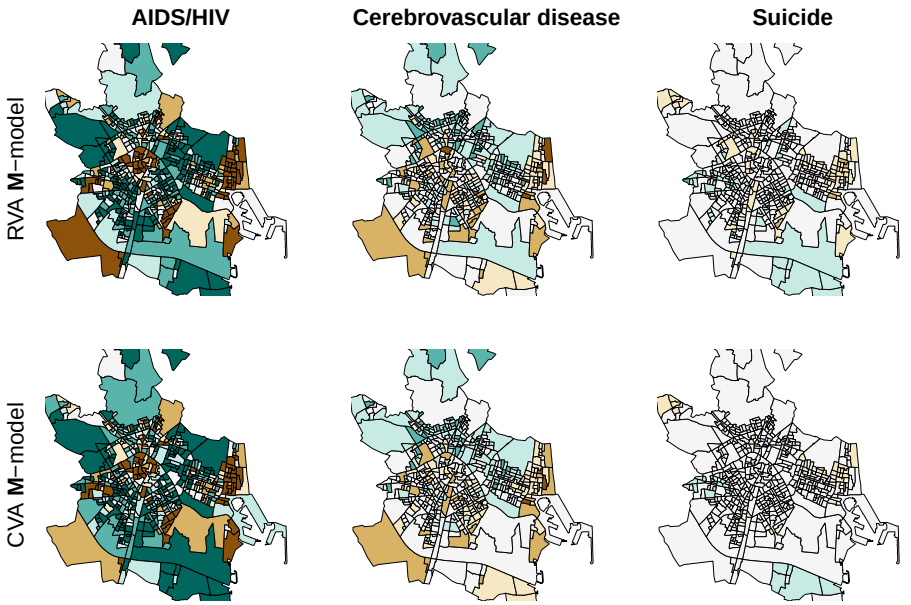


Figure B.4.: Graphical representation of the estimated risk in Valencia using the new variance-adaptive modeling proposals (RVA and CVA $\boldsymbol{M}$-modeling).

B.2. Code used to obtain results

B.2.1. Execution of models in WinBUGS using the library R2WinBUGS

Load libraries and data

```
# Working directory
DirMain = " "  # Set an appropriate directory
setwd(DirMain)
# Load library and data
library(R2WinBUGS)
library(knitr)
# For running the models in parallel calls to WinBUGS
library(pbugs)
load("datos.Rdata")
```

BYM model

```
# BYM model with independent diseases,
# WinBUGS code
BYM.indep <- function() {
    for (j in 1:Ndiseases) {
        for (i in 1:Nareas) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each
            # municipality
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + sdhet[j] *
                het[i, j] + sdsp[j] * sp[j,
                i]
            # SMR for each municipality
            SMR[i] <- exp(mu[j] + sdhet[j] *
```

```
                het[i, j] + sdsp[j] * sp[j,
                i])
            # Prior distribution for the non-spatial
            # effect
            het[i, j] ~ dnorm(0, 1)
        }
        # Prior distribution for the spatial
        # effect
        sp[j, 1:Nareas] ~ car.normal(adj[], weights[],
            num[], 1)
        # Prior distribution for the mean risk
        # for all municipalities
        mu[j] ~ dflat()
        # Prior distribution for the standard
        # deviations of the random effects
        sdhet[j] ~ dunif(0, 5)
        sdsp[j] ~ dunif(0, 5)
    }
}

# Run BYM model for each city, sex and
# disease

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    for (k in 1:2) {
        # Mortality cause
        l <- c(1:20)
```

```
# Matrix with observed and expected cases
O <- t(apply(Obs[[i]][[j]][k, , ,
    ], c(1, 2), sum)[1, ])
E <- t(apply(Esp[[i]][[j]][k, , ,
    ], c(1, 2), sum)[1, ])
Nareas <- dim(O)[1]
Ndiseases <- dim(O)[2]

# Data
data <- list(Nareas = Nareas, Ndiseases =
    Ndiseases,  O = O, E = E, adj =
    unlist(nb[[i]]),
    weights = rep(1, length(unlist(nb[[i]]))),
    num = sapply(nb[[i]], length))
# Initial values
initials <- function() {
    list(mu = rnorm(Ndiseases, 0,
        0.1), sdhet = runif(Ndiseases,
        0, 1), sdsp = runif(Ndiseases,
        0, 1), het = matrix(rnorm(Nareas *
        Ndiseases), nrow = Nareas,
        ncol = Ndiseases), sp = matrix(rnorm(Nareas *
        Ndiseases), nrow = Ndiseases,
        ncol = Nareas))
}
# Variables to retrive
param <- c("SMR", "lambda", "het",
    "sp", "mu", "sdsp", "sdhet")
# Calls to WinBUGS
t.result <- system.time(result <- pbugs(data =
    data,  model.file = BYM.indep, inits =
     initials,  parameters.to.save = param,
    n.chains = 3,  n.iter = 30000,
```

```
            n.burnin = 5000,  DIC = F))
        # Save results
        save(result, t.result, file = paste(getwd(),
            "/resul/resul.BYMIndep.", i,
            ".", j, ".", k, ".Rdata", sep = ""))
    }
}
```

Fixed effects M-model (Section 5.2 in paper)

```
# Fixed effects M-model, WinBUGS code

Mmodel.FE <- function() {
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each
            # municipality and disease
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + Theta[i, j]
            # SMR for each municipality and disease
            SMR[i, j] <- exp(mu[j] + Theta[i,
                j])
        }
    }

    # Definition of the random effects matrix
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            Theta[i, j] <- inprod2(tPhi[,
                i], M[, j])
        }
    }
```

```
    # Matrix of spatially correlated random
    # effects if M is a square matrix define
    # Nsp (Number of spatial underlying
    # patterns) as Ndiseases
    for (j in 1:Nsp) {
        # Prior distribution for the spatial
        # effect
        Spatial[j, 1:Nareas] ~ car.normal(adj[],
            weights[], num[], 1)
        for (i in 1:Nareas) {
            # Prior distribution for the non-spatial
            # effect
            Het[j, i] ~ dnorm(0, 1)
            tPhi[j, i] <- Spatial[j, i]
        }
    }

    for (j in (Nsp + 1):(2 * Nsp)) {
        for (i in 1:Nareas) {
            tPhi[j, i] <- Het[(j - Nsp),
                i]
        }
    }

    # M-matrix
    for (i in 1:(2 * Nsp)) {
        for (j in 1:Ndiseases) {
            M[i, j] ~ dflat()
        }
    }

    # Others prior distributions
    for (j in 1:Ndiseases) {
```

```
        # Prior distribution for the mean risk
        # for all municipalities
        mu[j] ~ dflat()
    }
}

# Run fixed effects M-model considering
# 20 causes of mortality

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    for (k in 1:2) {
        # Mortality causes
        l <- c(1:20)

        # Matrix with observed and expected cases
        O <- t(apply(Obs[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])
        E <- t(apply(Esp[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])
        Nareas <- dim(O)[1]
        Ndiseases <- dim(O)[2]
        # Data
        data <- list(Nareas = Nareas, Ndiseases =
            Ndiseases,  Nsp = Ndiseases, O = O, E = E,
            adj = unlist(nb[[i]]), weights = rep(1,
                length(unlist(nb[[i]]))),
            num = sapply(nb[[i]], length))
        # Initial values
```

```
        initials <- function() {
            list(mu = rnorm(Ndiseases, 0,
                0.1), M = matrix(rnorm(2 *
                Nsp * Ndiseases), nrow = 2 *
                Nsp, ncol = Ndiseases), Het = matrix(rnorm
               (Nareas *  Nsp), nrow = Nsp, ncol =Nareas),
                Spatial = matrix(rnorm(Nareas *
                  Nsp), nrow = Nsp, ncol = Nareas))
        }
        # Variables to retrieve
        param <- c("SMR", "lambda", "M",
            "Het", "Spatial", "mu", "Theta")
        # Calls to WinBUGS
        t.result <- system.time(result <- pbugs(data =
            data,  model.file = Mmodel.FE, inits =initials,
            parameters.to.save = param, n.chains = 3,
            n.iter = 30000, n.burnin = 5000,
            DIC = F))
        # Save results
        save(result, t.result, file = paste(getwd(),
            "/resul/resul.MmodelFE.", i,
            ".", j, ".", k, ".Rdata", sep = ""))

    }
}
```

Random effects M-model or NVA M-model (Section 5.2 in paper)

```
# Random effects M-model, WinBUGS code

Mmodel.RE <- function() {
    for (i in 1:Nareas) {
```

```
        for (j in 1:Ndiseases) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each
            # municipality and disease
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + Theta[i, j]
            # SMR for each municipality and disease
            SMR[i, j] <- exp(mu[j] + Theta[i,
                j])
        }
    }

    # Definition of the random effects matrix
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            Theta[i, j] <- inprod2(tPhi[,
                i], M[, j])
        }
    }

    # Matrix of spatially correlated random
    # effects: if M is a square matrix define
    # Nsp (Number of spatial underlying
    # patterns) as Ndiseases
    for (j in 1:Nsp) {
        # Prior distribution for the spatial
        # effect
        Spatial[j, 1:Nareas] ~ car.normal(adj[],
            weights[], num[], 1)
        for (i in 1:Nareas) {
            # Prior distribution for the non-spatial
            # effect
            Het[j, i] ~ dnorm(0, 1)
```

```
            tPhi[j, i] <- Spatial[j, i]
        }
    }

    for (j in (Nsp + 1):(2 * Nsp)) {
        for (i in 1:Nareas) {
            tPhi[j, i] <- Het[(j - Nsp),
                i]
        }
    }

    # M-matrix
    for (j in 1:Ndiseases) {
        for (i in 1:Nsp) {
            M[i, j] ~ dnorm(0, prec.sp)
        }

        for (i in (Nsp + 1):(2 * Nsp)) {
            M[i, j] ~ dnorm(0, prec.het)
        }
    }

    # Others prior distributions

    # Prior distribution for the mean risk
    # for all municipalities
    for (j in 1:Ndiseases) {
        mu[j] ~ dflat()
    }
    # Prior distribution for the standard
    # deviations of the random effects
    prec.sp <- pow(sd.sp, -2)
    sd.sp ~ dunif(0, 100)
```

```
    prec.het <- pow(sd.het, -2)
    sd.het ~ dunif(0, 100)
}

# Run random effects M-model considering
# 20 causes of mortality

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    for (k in 1:2) {
        # Mortality causes
        l <- c(1:20)

        # Matrix with observed and expected cases
        O <- t(apply(Obs[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])
        E <- t(apply(Esp[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])

        Nareas <- dim(O)[1]
        Ndiseases <- dim(O)[2]

        # Data
        data <- list(Nareas = Nareas, Ndiseases =
            Ndiseases,  Nsp = Ndiseases, O = O, E = E,
            adj = unlist(nb[[i]]), weights = rep(1,
                length(unlist(nb[[i]]))),
            num = sapply(nb[[i]], length))
```

```
        # Initial values
        initials <- function() {
            list(mu = rnorm(Ndiseases, 0,
                0.1), sd.sp = runif(1, 0,
                1), sd.het = runif(1, 0,
                1), Het = matrix(rnorm(Nareas *
                Nsp), nrow = Nsp, ncol = Nareas),
                Spatial = matrix(rnorm(Nareas *
                 Nsp), nrow = Nsp, ncol = Nareas))
        }

        # Variables to retrieve
        param <- c("sd.sp", "sd.het", "SMR",
            "lambda", "M", "Het", "Spatial",
            "mu")

        # Calls to WinBUGS
        t.result <- system.time(result <- pbugs(data =
            data,  model.file = Mmodel.RE, inits =initials,
            parameters.to.save = param, n.chains = 3,
            n.iter = 30000, n.burnin = 5000,
            DIC = F))

        # Save results
        save(result, t.result, file = paste(getwd(),
            "/resul/resul.MmodelRE.", i,
            ".", j, ".", k, ".Rdata", sep = ""))
    }
}
```

RVA M-model (Section 5.4 in paper)

```
# RVA M-model, WinBUGS code

Mmodel.RVA <- function() {
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each
            # municipality and disease
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + Theta[i, j]
            # SMR for each municipality and disease
            SMR[i, j] <- exp(mu[j] + Theta[i,
                j])
        }
    }

    # Definition of the random effects matrix
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            Theta[i, j] <- inprod2(tPhi[,
                i], M[, j])
        }
    }

    # Matrix of spatially correlated random
    # effects if M is a square matrix define
    # Nsp (Number of spatial underlying
    # patterns) as Ndiseases
    for (j in 1:Nsp) {
        # Prior distribution for the spatial
        # effect
        Spatial[j, 1:Nareas] ~ car.normal(adj[],
```

```
            weights[], num[], 1)
        for (i in 1:Nareas) {
            # Prior distribution for the non-spatial
            # effect
            Het[j, i] ~ dnorm(0, 1)
            tPhi[j, i] <- Spatial[j, i]
        }
    }

    for (j in (Nsp + 1):(2 * Nsp)) {
        for (i in 1:Nareas) {
            tPhi[j, i] <- Het[(j - Nsp),
                i]
        }
    }


    # M-matrix
    for (j in 1:Ndiseases) {

        for (i in 1:(2 * Nsp)) {
            M.aux[i, j] ~ dnorm(0, 1)
            M[i, j] <- sd[i] * M.aux[i, j]
        }
    }

    # Others prior distributions

    # Prior distribution for the mean risk
    # for all municipalities
    for (j in 1:Ndiseases) {
        mu[j] ~ dflat()
    }
```

```
    # Prior distribution for the standard
    # deviations of the random effects
    for (i in 1:(2 * Nsp)) {
        sd[i] ~ dunif(0, 5)
    }
}

# Run RVA M-model considering 20 causes
# of mortality

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    for (k in 1:2) {
        # Mortality causes
        l <- c(1:20)

        # Matrix with observed and expected cases
        O <- t(apply(Obs[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])
        E <- t(apply(Esp[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])

        Nareas <- dim(O)[1]
        Ndiseases <- dim(O)[2]

        # Data
        data <- list(Nareas = Nareas, Ndiseases =
            Ndiseases,  Nsp = Ndiseases, O = O, E = E,
            adj = unlist(nb[[i]]), weights = rep(1,
```

```
                length(unlist(nb[[i]]))),
            num = sapply(nb[[i]], length))

        # Initial values
        initials <- function() {
            list(mu = rnorm(Ndiseases, 0,
                0.1), sd = runif(2 * Nsp,
                0.1, 1), Het = matrix(rnorm(Nareas *
                Nsp), nrow = Nsp, ncol = Nareas),
                Spatial = matrix(rnorm(Nareas *
                Nsp), nrow = Nsp, ncol = Nareas))
        }

        # Variables to retrieve
        param <- c("sd", "SMR", "lambda",
            "M", "mu")

        # Calls to WinBUGS
        t.result <- system.time(result <- pbugs(data =
            data,  model.file = Mmodel.RVA, inits=initials,
            parameters.to.save = param, n.chains = 3,
            n.iter = 30000, n.burnin = 5000,
            DIC = F))

        # Save results
        save(result, t.result, file = paste(getwd(),
            "/resul/resul.MmodelRVA.", i,
            ".", j, ".", k, ".Rdata", sep = ""))
    }
}
```

CVA M-model (Section 5.4 in paper)

```
# CVA M-model, WinBUGS code

Mmodel.CVA <- function() {
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each
            # municipality and disease
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + Theta[i, j]
            # SMR for each municipality and disease
            SMR[i, j] <- exp(mu[j] + Theta[i,
                j])
        }
    }

    # Definition of the random effects matrix
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            Theta[i, j] <- inprod2(tPhi[,
                i], M[, j])
        }
    }

    # Matrix of spatially correlated random
    # effects if M is a square matrix define
    # Nsp (Number of spatial underlying
    # patterns) as Ndiseases
    for (j in 1:Nsp) {
        # Prior distribution for the spatial
        # effect
        Spatial[j, 1:Nareas] ~ car.normal(adj[],
```

```
            weights[], num[], 1)
        for (i in 1:Nareas) {
            # Prior distribution for the non-spatial
            # effect
            Het[j, i] ~ dnorm(0, 1)
            tPhi[j, i] <- Spatial[j, i]
        }
    }

    for (j in (Nsp + 1):(2 * Nsp)) {
        for (i in 1:Nareas) {
            tPhi[j, i] <- Het[(j - Nsp),
                i]
        }
    }

    # M-matrix
    for (j in 1:Ndiseases) {
        for (i in 1:Nsp) {
            M.aux[i, j] ~ dnorm(0, 1)
            M[i, j] <- sdstruct.sp[j] * M.aux[i,
                j]
        }
        for (i in (Nsp + 1):(2 * Nsp)) {
            M.aux[i, j] ~ dnorm(0, 1)
            M[i, j] <- sdstruct.het[j] *
                M.aux[i, j]
        }
    }

    # Others prior distributions Prior
    # distribution for the mean risk for all
    # municipalities
```

```
    for (j in 1:Ndiseases) {
        mu[j] ~ dflat()
    }

    # Prior distribution for the standard
    # deviations of the random effects
    for (j in 1:Ndiseases) {
        prec.sp[j] <- pow(sdstruct.sp[j],
            -2)
        sdstruct.sp[j] ~ dunif(0, 5)

        prec.het[j] <- pow(sdstruct.het[j],
            -2)
        sdstruct.het[j] ~ dunif(0, 5)
    }
}

# Run CVA M-model considering 20 causes
# of mortality

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    for (k in 1:2) {
        # Mortality causes
        l <- c(1:20)

        # Matrix with observed and expected cases
        O <- t(apply(Obs[[i]][[j]][k, , ,
            ], c(1, 2), sum)[l, ])
```

```
E <- t(apply(Esp[[i]][[j]][k, , ,
    ], c(1, 2), sum)[1, ])

Nareas <- dim(0)[1]
Ndiseases <- dim(O)[2]

# Data
data <- list(Nareas = Nareas, Ndiseases =
    Ndiseases,  Nsp = Ndiseases, O = O, E = E,
    adj = unlist(nb[[i]]), weights = rep(1,
        length(unlist(nb[[i]]))),
    num = sapply(nb[[i]], length))

# Initial values
initials <- function() {
    list(mu = rnorm(Ndiseases, 0,
        0.1), sdstruct.sp = runif(Ndiseases,
        0, 1), sdstruct.het = runif(Ndiseases,
        0, 1), Het = matrix(rnorm(Nareas *
      Nsp), nrow = Nsp, ncol = Nareas),
        Spatial = matrix(rnorm(Nareas *
          Nsp), nrow = Nsp, ncol = Nareas))
}

# Variables to retrieve
param <- c("sdstruct.sp", "sdstruct.het",
    "SMR", "lambda", "M", "mu")

# Calls to WinBUGS
t.result <- system.time(result <- pbugs(data =
    data,  model.file = Mmodel.CVA, inits=initials,
    parameters.to.save = param, n.chains = 3,
    n.iter = 30000, n.burnin = 5000,
```

```
            DIC = F))

        # Save results
        save(result, t.result, file = paste(getwd(),
            "/resul/resul.MmodelCVA.", i,
            ".", j, ".", k, ".Rdata", sep = ""))
    }
}
```

B.2.2. Choropleth maps for all models

```
# Load libraries
library(RColorBrewer)

cuts_SMR <- c(0, 0.67, 0.8, 0.91, 1.1, 1.25,
    1.5)
palette <- brewer.pal(7, "BrBG")[7:1]

# Name of mortality causes
Causes <- dimnames(Obs[[i]][[j]])[[2]]

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    k <- 1
    # Mortality causes
    for (l in 1:20) {

        # Load WinBUGS results, BYM model
```

```
load(paste(getwd(), "/resul.BYMIndep.",
    i, ".", j, ".", k, ".", l, ".Rdata",
    sep = ""))

# SMR estimates, BYM model
aux <- palette[findInterval(result$mean$SMR,
    cuts_SMR)]
plot(Carto[[i]], col = palette[aux],
    main = paste0("BYM - ", Causas[l]),
    lwd = 0.2)
legend("bottomright", c("< 0.67",
    "0.67 - 0.80", "0.80 - 0.91",
    "0.91 - 1.10", "1.10 - 1.25",
    "1.25 - 1.50", "> 1.50"), title = "SMR",
    border = NULL, fill = palette,
    bty = "n")

# Load WinBUGS results, fixed effects
# M-model
load(paste(getwd(), "/resul.MmodelFE.",
    i, ".", j, ".", k, ".Rdata",
    sep = ""))

# SMR estimates, fixed effects M-model
aux <- palette[findInterval(result$mean$SMR[,
    l], cuts_SMR)]
plot(Carto[[i]], col = palette[aux],
    main = paste0("MmodelFE - ",
        Causas[l]), lwd = 0.2)
legend("bottomright", c("< 0.67",
    "0.67 - 0.80", "0.80 - 0.91",
    "0.91 - 1.10", "1.10 - 1.25",
    "1.25 - 1.50", "> 1.50"), title = "SMR",
```

```
        border = NULL, fill = palette,
        bty = "n")

    # Load WinBUGS results, random effects
    # M-model
    load(paste(getwd(), "/resul.MmodelRE.",
        i, ".", j, ".", k, ".Rdata",
        sep = ""))

    # SMR estimates, random effects M-model
    aux <- palette[findInterval(result$mean$SMR[,
        l], cuts_SMR)]
    plot(Carto[[i]], col = palette[aux],
        main = paste0("MmodelRE - ",
            Causas[l]), lwd = 0.2)
    legend("bottomright", c("< 0.67",
        "0.67 - 0.80", "0.80 - 0.91",
        "0.91 - 1.10", "1.10 - 1.25",
        "1.25 - 1.50", "> 1.50"), title = "SMR",
        border = NULL, fill = palette,
        bty = "n")

    # Load WinBUGS results, RVA m-model
    load(paste(getwd(), "/resul.MmodelRVA.",
        i, ".", j, ".", k, ".Rdata",
        sep = ""))

    # SMR estimates, RVA M-model
    aux <- palette[findInterval(result$mean$SMR[,
        l], cuts_SMR)]
    plot(Carto[[i]], col = palette[aux],
        main = paste0("MmodelRVA - ",
            Causas[l]), lwd = 0.2)
```

```
        legend("bottomright", c("< 0.67",
            "0.67 - 0.80", "0.80 - 0.91",
            "0.91 - 1.10", "1.10 - 1.25",
            "1.25 - 1.50", "> 1.50"), title = "SMR",
            border = NULL, fill = palette,
            bty = "n")

        # Load WinBUGS results, CVA m-model
        load(paste(getwd(), "/resul.MmodelCVA.",
            i, ".", j, ".", k, ".Rdata",
            sep = ""))

        # SMR estimates, CVA M-model
        aux <- palette[findInterval(result$mean$SMR[,
            l], cuts_SMR)]
        plot(Carto[[i]], col = palette[aux],
            main = paste0("MmodelCVA - ",
                Causas[l]), lwd = 0.2)
        legend("bottomright", c("< 0.67",
            "0.67 - 0.80", "0.80 - 0.91",
            "0.91 - 1.10", "1.10 - 1.25",
            "1.25 - 1.50", "> 1.50"), title = "SMR",
            border = NULL, fill = palette,
            bty = "n")

    }
}
```

B.2.3. DIC for each model (Table 5.2 in paper)

```
# Function for DICs calculation
CalculaDIC <- function(Simu, O, save = FALSE) {
```

```
    mu <-
        Simu$sims.matrix[, which(substr(dimnames
        (Simu$sims.matrix)[[2]],  1, 2) == "la")]
    D <- apply(mu, 1, function(x) {
        -2 * sum(dpois(as.vector(t(O)), x,
            log = T))
    })
    Dmedia <- mean(D)
    mumedia <- apply(mu, 2, mean)
    DenMedia <- -2 * sum(dpois(as.vector(t(O)),
        mumedia, log = T))
    if (save == TRUE) {
        return(c(Dmedia, Dmedia - DenMedia,
            2 * Dmedia - DenMedia))
    }
    cat("D=", Dmedia, "pD=", Dmedia - DenMedia,
        "DIC=", 2 * Dmedia - DenMedia, "\n")
}

DIC.BYMIndep <- list()
DIC.MmodelFE <- list()
DIC.MmodelRE <- list()
DIC.MmodelRVA <- list()
DIC.MmodelCVA <- list()

# City = 1: Alicante, 2: Castellón, 3:
# Valencia
for (i in 1:3) {
    # Specific mortality causes
    j <- 2
    # Sex = 1: Mens, 2: Women
    k <- 1
    # Mortality causes
```

```
    l <- c(1:20)

    # Matrix with observed cases
    O <- t(apply(Obs[[i]][[j]][k, , , ],
        c(1, 2), sum)[l, ])

    # DIC M-model with independent diseases
    load(paste(getwd(), "/resul.BYMIndep.",
        i, ".", j, ".", k, ".Rdata", sep = ""))
    DIC.BYMIndep[[i]] <- CalculaDIC(Simu = result,
        O = O, save = TRUE)[3]
    # DIC fixed effects M-model
    load(paste(getwd(), "/resul.MmodelFE.",
        i, ".", j, ".", k, ".Rdata", sep = ""))
    DIC.MmodelFE[[i]] <- CalculaDIC(Simu = result,
        O = O, save = TRUE)[3]
    # DIC random effects M-model
    load(paste(getwd(), "/resul.MmodelRE.",
        i, ".", j, ".", k, ".Rdata", sep = ""))
    DIC.MmodelRE[[i]] <- CalculaDIC(Simu = result,
        O = O, save = TRUE)[3]
    # DIC RVA M-model
    load(paste(getwd(), "/resul.MmodelRVA.",
        i, ".", j, ".", k, ".Rdata", sep = ""))
    DIC.MmodelRVA[[i]] <- CalculaDIC(Simu = result,
        O = O, save = TRUE)[3]
    # DIC CVA M-model
    load(paste(getwd(), "/resul.MmodelCVA.",
        i, ".", j, ".", k, ".Rdata", sep = ""))
    DIC.MmodelCVA[[i]] <- CalculaDIC(Simu = result,
        O = O, save = TRUE)[3]

}
```

```
Table <- matrix(c(unlist(DIC.BYMIndep),
    unlist(DIC.MmodelFE),  unlist(DIC.MmodelRE),
    unlist(DIC.MmodelRVA),  unlist(DIC.MmodelCVA)),
    ncol = 3, byrow = TRUE)
rownames(Table) <- c("BYM with independent diseases",
    "Fixed effects $M$-model", "Random effects $M$-model",
    "RVA $M$-model", "CVA $M$-model")
colnames(Table) <- c("Alicante", "Castellón",
    "Valencia")
print(kable(Table, caption = "DICs for the adjusted models
    in each study city"))
```

B.2.4. Used code in the simulation study

Simulation of data for each setting and city

Load libraries and data

```
# Working directory
DirMain = " "  # Set an appropriate directory
setwd(DirMain)
# Load library and data
library(R2WinBUGS)
library(knitr)
# For running the models in parallel calls to WinBUGS
library(pbugs)
load("datos.Rdata")

# Function to generate values of a CAR
# distribution
Genera_CAR <- function(desv, nvec, adj, rho = 1) {
    n <- length(nvec)
    D.W <- matrix(0, n, n)
```

```
    diag(D.W) <- nvec
    indice_veci <- cbind(rep(1:n, nvec),
        adj)
    D.W[indice_veci] <- -rho
    UDUt <- eigen(D.W)
    rango <- sum(UDUt$values > 1e-10)
    Spat <- as.vector(UDUt$vectors[, 1:rango] %*%
        matrix(rnorm(rango, 0, UDUt$values[1:rango]^{
            -1/2
        }), ncol = 1)) * desv
    return(Spat)
}

# Seeds for each replica (1:5) and city
# (1: Alicante, 2:Castellón, 3:Valencia)
seeds <- list()
seeds[[1]] <- c(20, 54, 86, 92, 6)
seeds[[2]] <- c(89, 94, 102, 92, 6)
seeds[[3]] <- c(20, 54, 67, 92, 6)
```

Setting 1

```
for (City in 1:3) {
    # 1: Alicante, 2:Castellón, 3:Valencia
    for (Ndiseases in c(5, 10)) {
        # Expected cases
        E <- t(apply(Esp[[City]][[1]][1,
            , , ], c(1, 2), sum))[, 1:Ndiseases]
        for (Replica in 1:5) {
            # Matrix with spatial random effects for
            # each disease
            Y <- matrix(NA, nrow = dim(Carto[[City]])[1],
```

```
    ncol = Ndiseases)
# Matrix with simulated observed cases
# for each disease
Obs_simu <- matrix(NA, nrow = dim
    (Carto[[City]])[1],  ncol = Ndiseases)

# Common spatial pattern to all diseases
set.seed(79 * seeds[[City]][Replica])
patron_comun <- Genera_CAR(desv = 0,
    nvec = sapply(nb[[City]],
      length), adj = unlist(nb[[City]]),
    rho = 0.9)

# Specific spatial pattern for each
# disease First disease
i <- 1
set.seed(i * seeds[[City]][Replica])
Y[, i] <- patron_comun + Genera_CAR(desv = 1,
    nvec = sapply(nb[[City]],
      length), adj = unlist(nb[[City]]),
    rho = 0.9)

# Other diseases
for (i in 2:Ndiseases) {
    set.seed(i * seeds[[City]][Replica])
    Y[, i] <- patron_comun +
      Genera_CAR(desv = 0.2,
        nvec = sapply(nb[[City]],
          length), adj = unlist(nb[[City]]),
        rho = 0.9)
}

# Simulation of the observed cases for
```

```
            # each disease
            for (i in 1:Ndiseases) {
                mu_Obs <- exp(Y[, i]) * E[,
                  i]
                set.seed(i)
                Obs_simu[, i] <- rpois(
                  dim(Carto[[City]])[1],  mu_Obs)
            }

            save(Obs_simu, Y, E, file = paste0(City,
                "/", Ndiseases, "enfermedades/Escenario 1/
                datos_simulados",  Replica, ".RData"))
        }
    }
}

# Next, adjust the BYM model with
# independent diseases, NVA M-model, RVA
# M-model and CVA M-Model to the
# simulated observed cases for each data
# set (following the code specified in
# the Annex).
```

Setting 2

```
for (City in 1:3) {
    # 1: Alicante, 2:Castellón, 3:Valencia
    for (Ndiseases in c(5, 10)) {
        # Expected cases
        E <- t(apply(Esp[[City]][[1]][1,
            , , ], c(1, 2), sum))[, 1:Ndiseases]
        for (Replica in 1:5) {
            # Matrix with spatial random effects for
```

```
# each disease
Y <- matrix(NA, nrow = dim(Carto[[City]])[1],
    ncol = Ndiseases)
# Matrix with simulated observed cases
# for each disease
Obs_simu <- matrix(NA, nrow = dim
    (Carto[[City]])[1],  ncol = Ndiseases)

# Common spatial pattern to all diseases
set.seed(79 * seeds[[City]][Replica])
patron_comun <- Genera_CAR(desv = 0.5,
    nvec = sapply(nb[[City]],
      length), adj = unlist(nb[[City]]),
    rho = 0.9)

# Specific spatial pattern for each
# disease First disease
i <- 1
set.seed(i * seeds[[City]][Replica])
Y[, i] <- patron_comun + Genera_CAR(desv = 1,
    nvec = sapply(nb[[City]],
      length), adj = unlist(nb[[City]]),
    rho = 0.9)

# Other diseases
for (i in 2:Ndiseases) {
    set.seed(i * seeds[[City]][Replica])
    Y[, i] <- patron_comun +
      Genera_CAR(desv = 0.2,
        nvec = sapply(nb[[City]],
          length), adj = unlist(nb[[City]]),
        rho = 0.9)
}
```

```
            # Simulation of the observed cases for
            # each disease
            for (i in 1:Ndiseases) {
                mu_Obs <- exp(Y[, i]) * E[,
                  i]
                set.seed(i)
                Obs_simu[, i] <- rpois(
                  dim(Carto[[City]])[1],  mu_Obs)
            }

            save(Obs_simu, Y, E, file = paste0(City,
                "/", Ndiseases, "enfermedades/Escenario 2/
                datos_simulados",  Replica, ".RData"))
        }
    }
}

# Next, adjust the BYM model with
# independent diseases, NVA M-model, RVA
# M-model and CVA M-Model to the
# simulated observed cases for each data
# set (following the code specified in
# the Annex).
```

Setting 3

```
for (City in 1:3) {
    # 1: Alicante, 2:Castellón, 3:Valencia
    for (Ndiseases in c(5, 10)) {
        # Expected cases
        E <- t(apply(Esp[[City]][[1]][1,
            , , ], c(1, 2), sum))[, 1:Ndiseases]
```

```
    for (Replica in 1:5) {
        # Matrix with spatial random effects for
        # each disease
        Y <- matrix(NA, nrow = dim(Carto[[City]])[1],
            ncol = Ndiseases)
        # Matrix with simulated observed cases
        # for each disease
        Obs_simu <- matrix(NA, nrow = dim
            (Carto[[City]])[1],  ncol = Ndiseases)

        # Common spatial pattern to all diseases
        set.seed(79 * seeds[[City]][Replica])
        patron_comun <- Genera_CAR(desv = 1,
            nvec = sapply(nb[[City]],
              length), adj = unlist(nb[[City]]),
            rho = 0.9)

        # Specific spatial pattern for each
        # disease First disease
        i <- 1
        set.seed(i * seeds[[City]][Replica])
        Y[, i] <- patron_comun + Genera_CAR(desv = 1,
            nvec = sapply(nb[[City]],
              length), adj = unlist(nb[[City]]),
            rho = 0.9)

        # Other diseases
        for (i in 2:Ndiseases) {
            set.seed(i * seeds[[City]][Replica])
            Y[, i] <- patron_comun +
              Genera_CAR(desv = 0.2,
                nvec = sapply(nb[[City]],
                  length), adj = unlist(nb[[City]]),
```

```
                    rho = 0.9)
            }

            # Simulation of the observed cases for
            # each disease
            for (i in 1:Ndiseases) {
                mu_Obs <- exp(Y[, i]) * E[,
                  i]
                set.seed(i)
                Obs_simu[, i] <- rpois(
                  dim(Carto[[City]])[1],  mu_Obs)
            }

            save(Obs_simu, Y, E, file = paste0(City,
                "/", Ndiseases, "enfermedades/Escenario 3/
                datos_simulados",  Replica, ".RData"))
        }
    }
}

# Next, adjust the BYM model with
# independent diseases, NVA M-model, RVA
# M-model and CVA M-Model to the
# simulated observed cases for each data
# set (following the code specified in
# the Annex).
```

Mean standard deviation of the risks for the first disease and for the rest of diseases in each setting, city and model used in the adjustment of the data

```
# Specify number of diseases (in our
# study 5 and 10 diseases)
Ndiseases <- 5

# Object in which we save the results
Resul <- list()
for (City in 1:3) {
    Resul[[City]] <- list()
}
names(Resul) <- c("Alicante", "Castellon",
    "Valencia")
n_geographicalunits <- c(215, 95, 553)
names(n_geographicalunits) <- c("Alicante",
    "Castellon", "Valencia")

for (City in 1:3) {
    for (Setting in 1:3) {
        # Object in which we save the simulated
        # spatial patterns for each disease in
        # each replica
        Sim_data <- array(NA, dim = c(5,
            n_geographicalunits[City], Ndiseases))
        # Object in which we save the estimated
        # risks with the BYM model with
        # independent diseases for each replica
        BYM_indep <- array(NA, dim = c(5,
            n_geographicalunits[City], Ndiseases))
        # Object in which we save the estimated
        # risks with the NVA M-model for each
        # replica
        NVA <- array(NA, dim = c(5
            , n_geographicalunits[City],  Ndiseases))
        # Object in which we save the estimated
```

```
        # risks with the CVA M-model for each
        # replica
        CVA <- array(NA, dim = c(5
            , n_geographicalunits[City],  Ndiseases))
        # Object in which we save the estimated
        # risks with the RVA M-model for each
        # replica
        RVA <- array(NA, dim = c(5
            , n_geographicalunits[City],  Ndiseases))

        for (Replica in 1:5) {
            # Simulated observed cases
            load(paste0(City, "/", Ndiseases,
                " enfermedades/Escenario ",
                Setting, "/datos_simulados",
                Replica, ".RData"))
            Sim_data[Replica, , ] <- Y
            # Estimated risks with the BYM model with
            # independent diseases
            load(paste0(City, "/", Ndiseases,
                " enfermedades/Escenario ",
                Setting, "/Resultados/Replica ",
                Replica, "/resul.BYMIndep.Rdata"))
            BYM_indep[Replica, , ] <- result$mean$SMR
            # Estimated risks with NVA M-model
            load(paste0(City, "/", Ndiseases,
                " enfermedades/Escenario ",
                Setting, "/Resultados/Replica ",
                Replica, "/resul.MmodelRE.Rdata"))
            NVA[Replica, , ] <- result$mean$SMR
            # Estimated risks with CVA M-model
            load(paste0(City, "/", Ndiseases,
                " enfermedades/Escenario ",
```

```
                Setting, "/Resultados/Replica ",
                Replica, "/resul.MmodelCVA.Rdata"))
            CVA[Replica, , ] <- result$mean$SMR
            # Estimated risks with RVA M-model
            load(paste0(City, "/", Ndiseases,
                " enfermedades/Escenario ",
                Setting, "/Resultados/Replica ",
                Replica, "/resul.MmodelRVA.Rdata"))
            RVA[Replica, , ] <- result$mean$SMR
        }

        # Original standard deviation of the
        # simulated spatial patterns and standard
        # deviation of the estimated risks with
        # each model
        Resul[[City]][[Setting]] <- list()
        Resul[[City]][[Setting]]$sds <- cbind(apply(apply(
            exp(Sim_data),
            c(1, 3), sd), 2, mean), apply(apply(BYM_indep,
            c(1, 3), sd), 2, mean), apply(apply(NVA,
            c(1, 3), sd), 2, mean), apply(apply(CVA,
            c(1, 3), sd), 2, mean), apply(apply(RVA,
            c(1, 3), sd), 2, mean))
        dimnames(Resul[[City]][[Setting]]$sds)[[2]] <-
            c("Original",  "BYM", "NVA", "CVA", "RVA")
    }
}

# Mean standard deviation of the first
# spatial pattern and the rest of spatial
# patterns
for (City in 1:3) {
    for (Setting in 1:3) {
```

```
        print(paste0("# ", names(Resul)[City],
            ", Escenario ", Setting, ", ",
            Ndiseases, " enfermedades"))
        print(round(rbind(Resul[[City]][[Setting]]$sds[1,
            ], apply(Resul[[City]][[Setting]]$sds[2:
            Ndiseases,  ], 2, mean)), 2))
    }
}
```

C. Supplementary material to the paper: *"On the use of adaptive spatial weight matrices from disease mapping multivariate analyses"*

C.1. Additional results

C.1.1. Standardized Mortality Ratios for studied mortality causes in Valencia estimated with the BYM (upper row) and Leroux (lower row) models and with spatial weights matrices of either unitary weights (left) or using the values obtained from the multivariate analysis of 14 diseases (all mortality causes of study except the evaluated cause) (right)

Figures can be viewed online at: https://link.springer.com/article/10.1007/s00477-020-01781-5?.

C.1.2. Mean absolute difference for the risks of the adjacency and adaptive BYM models as a function of the magnitude of the corresponding spatial weights. The spatial weigths matrix for each disease is that estimated with 14 diseases, excluding that particular disease

	Low spatial weight (5%)	Medium spatial weight (90%)	High spatial weight (5%)
AIDS	0.218	0.087	0.099
Stomach cancer	0.007	0.006	0.008
Colorectal cancer	0.004	0.004	0.007
Lung cancer	0.013	0.006	0.008
Prostate cancer	0.006	0.004	0.003
Bladder cancer	0.007	0.008	0.011
Hematological cancer	0.005	0.005	0.007
Mellitus diabetes	0.009	0.008	0.017
Dementia	0.012	0.007	0.010
Ischemic heart disease	0.014	0.010	0.014
Ictus	0.011	0.006	0.016
COPD	0.037	0.011	0.014
Liver cirrhosis	0.031	0.016	0.021
Suicides	0.011	0.009	0.013
Traffic accidents	0.012	0.010	0.017

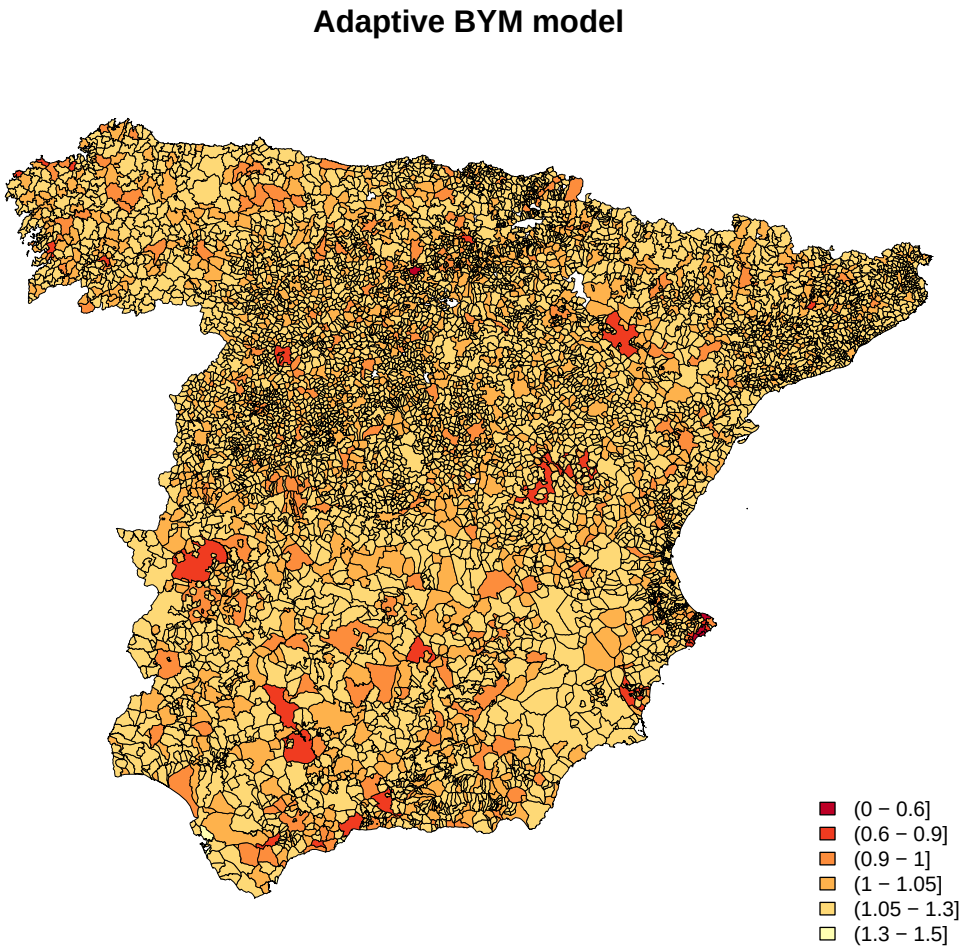
	Low spatial weight (5%)	Medium spatial weight (90%)	High spatial weight (5%)
Median	0.011	0.008	0.013
Mean	0.027	0.013	0.018

C.1.3. Mean absolute difference for the risks of each spatial unit and the mean risk for their neighbors, for the adjacency and adaptive BYM models, as a function of the magnitude of the corresponding spatial weights. The spatial weights matrix for each disease is that estimated with 14 diseases, excluding that particular disease

	Low spatial weight (5%)		High spatial weight (5%)	
	Adjacency BYM model	Adaptive BYM model	Adjacency BYM model	Adaptive BYM model
AIDS	1.641	1.692	0.433	0.368
Stomach cancer	0.019	0.021	0.016	0.016
Colorectal cancer	0.015	0.016	0.014	0.016
Lung cancer	0.037	0.048	0.025	0.024
Prostate cancer	0.012	0.015	0.013	0.012
Bladder cancer	0.061	0.063	0.052	0.051
Hematological cancer	0.022	0.024	0.020	0.020
Mellitus diabetes	0.021	0.026	0.020	0.021
Dementia	0.042	0.049	0.034	0.034
Ischemic heart disease	0.111	0.112	0.056	0.049
Ictus	0.032	0.039	0.027	0.027
COPD	0.096	0.133	0.060	0.060

	Low spatial weight (5%)		High spatial weight (5%)	
	Adjacency BYM model	Adaptive BYM model	Adjacency BYM model	Adaptive BYM model
Liver cirrhosis	0.110	0.134	0.084	0.079
Suicides	0.043	0.044	0.039	0.038
Traffic accidents	0.046	0.054	0.035	0.039
Median	0.042	0.048	0.034	0.034
Mean	0.154	0.165	0.062	0.057

C.1.4. Estimated spatial weights c_i for each municipality of Spain according to all 18 diseases in the data set. Choropleth map corresponds to either BYM model for the log-relative risks



C.2. R code to obtain results

C.2.1. Execution of models in WinBUGS using the R2WinBUGS and pbugs libraries

Load libraries and data

```
# Working directory
DirMain = " "  # Set an appropriate directory
setwd(DirMain)
# Load libraries and data
library(R2WinBUGS)  # For running WinBUGS from R
library(knitr)
# For running the models in parallel calls to WinBUGS
library(pbugs)
# For preparing information about spatial neighbors
# of each geographic unit to be used in WinBUGS
# (using the poly2nb and nb2WB functions)
library(spdep)
load("data.RData")

# Loaded data Obs: 4-dimensional array
# with the observed mortality cases for
# each year of the study, sex, geographic
# unit and disease Exp: 4-dimensional
# array with the expected mortality cases
# for each year of the study, sex,
# geographic unit and disease carto:
# SpatialPolygonsDataFrame of the study
# region
# Neighbours list of each geographic unit with class nb
carto.nb <- poly2nb(carto)
# List with the adjacency vector (carto.wb$adj) and the
# number of neighbors of each geographic unit
```

```
# (carto.wb$num) to use in WinBUGS
carto.wb <- nb2WB(carto.nb)
# Vector to identify the positions of the neighbors
# of each geographic unit
index <- c(1, cumsum(carto.wb$num))

# Studied mortality causes
causes <- c(1:2, 5, 7, 9:12, 15:21)
sex <- 1  # Mens
```

Multivariate adaptive BYM model

```
# Multivariate adaptive BYM model,
# WinBUGS code
AdaptiveBYM_model <- function() {
    # Likelihood
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each census
            # tract and disease
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + phi[i, j] +
                sd.theta[j] * theta[i, j]
            # SMR for each census tract and disease
            SMR[i, j] <- exp(mu[j] + phi[i,
                j] + sd.theta[j] * theta[i, j])
            # Prior distribution for spatial effects
            phi[i, j] ~ dnorm(mean.phi[i,
                j], prec.phi[i, j])
            # Prior distribution for non-spatial
            # effects
            theta[i, j] ~ dnorm(0, 1)
```

```
        }
    }

    for (i in 1:n.adj) {
        sqrt.c.adj[i] <- sqrt(c[adj[i]])
        for (j in 1:Ndiseases) {
            phi.adj[i, j] <- phi[adj[i],
                j]
        }
    }

    # Precision of the conditioned
    # distribution of spatial effects
    for (j in 1:Ndiseases) {
        prec.phi[1, j] <- pow(sd.phi[j],
            -2) * sqrt(c[1]) * sum(sqrt.c.adj[index[1]:
            index[2]])
        for (i in 2:Nareas) {
            prec.phi[i, j] <- pow(sd.phi[j],
               -2) * sqrt(c[i]) * sum(sqrt.c.adj[(index[i]
               +  1):index[i + 1]])
        }
    }

    # Mean of the conditioned distribution of
    # spatial effects
    for (j in 1:Ndiseases) {
        mean.phi[1, j] <- inprod2(sqrt.c.adj[index[1]:
            index[2]],  phi.adj[index[1]:index[2], j])
            /sum(sqrt.c.adj[index[1]:index[2]])
        for (i in 2:Nareas) {
            mean.phi[i, j] <- inprod2(sqrt.c.adj[(index[i]
                +  1):index[i + 1]], phi.adj[(index[i] +
```

```
                1):index[i + 1], j])/
                sum(sqrt.c.adj[(index[i] +
                1):index[i + 1]])
        }

        # Sum-to-zero restriction for spatial
        # effects
        ceros[j] <- 0
        ceros[j] ~ dnorm(sum.phi[j], 10)
        sum.phi[j] <- sum(phi[, j])
    }

    # Prior distributions for c
    for (i in 1:Nareas) {
        c[i] ~ dgamma(tau, tau) %_% I(0.001,  )
    }
    tau <- pow(sd.c, -2)
    sd.c ~ dunif(0, 5)

    # Other prior distributions
    for (j in 1:Ndiseases) {
        sd.phi[j] ~ dunif(0, 5)
        sd.theta[j] ~ dunif(0, 5)
        mu[j] ~ dflat()
    }
}

# Object where the results for each set
# of diseases will be saved
results.AdaptiveBYM <- list()

# Run multivariate adaptive BYM model for
# each set of diseases
```

```
for (i in 1:15) {
    # Selection of mortality causes
    causes.id <- causes[-c(i)]

    # Data
    data <- list(O = apply(Obs[, sex, , causes.id],
        c(2, 3), sum), E = apply(Exp[, sex,
        , causes.id], c(2, 3), sum), Nareas = dim(Obs)[3],
        Ndiseases = length(causes.id), n.adj =
        length(carto.wb$adj),
        adj = carto.wb$adj, index = index)
    # Initial values
    initials <- function() {
        list(mu = rnorm(data$Ndiseases, 0,
            1), sd.phi = runif(data$Ndiseases,
            0, 1), sd.theta = runif(data$Ndiseases,
            0, 1), phi = matrix(rnorm(data$Nareas *
            data$Ndiseases), nrow = data$Nareas,
            ncol = data$Ndiseases), theta =
            matrix(rnorm(data$Nareas *
            data$Ndiseases), nrow = data$Nareas,
            ncol = data$Ndiseases), c = runif(data$Nareas,
            0.9, 1.1), sd.c = runif(1, 0.5,
            0.6))
    }
    # Variables to retrive
    param <- c("mu", "lambda", "sd.phi",
        "phi", "sd.theta", "theta", "SMR",
        "c", "sd.c", "tau")
    # Calls to WinBUGS
    results.AdaptiveBYM[[i]] <- pbugs(data = data,
        inits = initials, parameters.to.save = param,
        model = AdaptiveBYM_model, n.iter = 2e+05,
```

```
        n.burnin = 50000, n.chains = 3, DIC = F)
}

# Save results
save(results.AdaptiveBYM, file =
        "Results/results.AdaptiveBYM.RData")
```

Multivariate adaptive Leroux model

```
# Multivariate adaptive Leroux model,
# WinBUGS code
AdaptiveLeroux_model <- function() {
    # Likelihood
    for (i in 1:Nareas) {
        for (j in 1:Ndiseases) {
            O[i, j] ~ dpois(lambda[i, j])
            # Modeling of the mean for each census
            # tract and disease
            log(lambda[i, j]) <- log(E[i,
                j]) + mu[j] + eta[i, j]
            # SMR for each census tract and disease
            SMR[i, j] <- exp(mu[j] + eta[i,
                j])
            # Prior distribution for spatial effects
            eta[i, j] ~ dnorm(mean.eta[i,
                j], prec.eta[i, j])
        }
    }

    for (i in 1:n.adj) {
        sqrt.c.adj[i] <- sqrt(c[adj[i]])
        for (j in 1:Ndiseases) {
```

```
            eta.adj[i, j] <- eta[adj[i],
                j]
        }
    }

    # Precision of the conditioned
    # distribution of spatial effects
    for (j in 1:Ndiseases) {
        prec.eta[1, j] <- pow(sd.eta[j],
            -2) * sqrt(c[1]) * (rho[j] *
            sum(sqrt.c.adj[index[1]:index[2]]) +
            1 - rho[j])
        for (i in 2:Nareas) {
            prec.eta[i, j] <- pow(sd.eta[j],
                -2) * sqrt(c[i]) * (rho[j] *
                sum(sqrt.c.adj[(index[i] +
                  1):index[i + 1]]) + 1 -
                rho[j])
        }
    }

    # Mean of the conditioned distribution of
    # spatial effects
    for (j in 1:Ndiseases) {
        mean.eta[1, j] <- (rho[j] * inprod2
            (sqrt.c.adj[index[1]:index[2]],
            eta.adj[index[1]:index[2], j]))/(rho[j] *
            sum(sqrt.c.adj[index[1]:index[2]]) +
            1 - rho[j])
        for (i in 2:Nareas) {
            mean.eta[i, j] <- (rho[j] * inprod2
                (sqrt.c.adj[(index[i] +
                1):index[i + 1]], eta.adj[(index[i] +
```

```
                1):index[i + 1], j]))/(rho[j] *
                sum(sqrt.c.adj[(index[i] +
                  1):index[i + 1]]) + 1 -
                rho[j])
        }

        # Sum-to-zero restriction for spatial
        # effects
        ceros[j] <- 0
        ceros[j] ~ dnorm(sum.eta[j], 10)
        sum.eta[j] <- sum(eta[, j])
    }

    # Prior distributions for c
    for (i in 1:Nareas) {
        c[i] ~ dgamma(tau, tau) %_% I(0.001,
            )
    }
    tau <- pow(sd.c, -2)
    sd.c ~ dunif(0, 5)

    # Other prior distributions
    for (j in 1:Ndiseases) {
        mu[j] ~ dflat()
        sd.eta[j] ~ dunif(0, 5)
        rho[j] ~ dunif(0, 1)
    }
}

# Object where the results for each set
# of diseases will be saved
results.AdaptiveLeroux <- list()
```

```
# Run multivariate adaptive Leroux model
# for each set of diseases
for (i in 1:15) {
    # Selection of mortality causes
    causes.id <- causes[-c(i)]

    # Data
    data <- list(O = apply(Obs[, sex, , causes.id],
        c(2, 3), sum), E = apply(Exp[, sex,
        , causes.id], c(2, 3), sum), Nareas = dim(Obs)[3],
        Ndiseases = length(causes.id), n.adj =
        length(carto.wb$adj),
        adj = carto.wb$adj, index = index)
    # Initial values
    initials <- function() {
        list(mu = rnorm(data$Ndiseases, 0,
            1), sd.eta = runif(data$Ndiseases,
            0, 1), rho = runif(data$Ndiseases,
            0, 1), eta = matrix(rnorm(data$Nareas *
            data$Ndiseases), nrow = data$Nareas,
            ncol = data$Ndiseases), c = runif(data$Nareas,
            0.9, 1.1), sd.c = runif(1, 0.5,
            1.5))
    }
    # Variables to retrive
    param <- c("mu", "lambda", "sd.eta",
        "SMR", "c", "sd.c", "rho")
    # Calls to WinBUGS
    results.AdaptiveLeroux[[i]] <- pbugs(data = data,
        inits = initials, parameters.to.save = param,
        model = AdaptiveLeroux_model, n.iter = 2e+05,
        n.burnin = 50000, n.chains = 3, DIC = F)
}
```

```
# Save results
save(results.AdaptiveLeroux, file =
        "Results/results.AdaptiveLeroux.RData")
```

Univariate BYM model with spatial weights matrices of either unitary weights or using the values obtained from the multivariate analysis of 14 diseases

```
# Univariate BYM model, WinBUGS code
BYM_model <- function() {
    # Likelihood
    for (i in 1:Nareas) {
        O[i] ~ dpois(lambda[i])
        # Modeling of the mean for each census
        # tract
        log(lambda[i]) <- log(E[i]) + mu +
            sd.phi * phi[i] + sd.theta *
            theta[i]
        # SMR for each census tract
        SMR[i] <- exp(mu + sd.phi * phi[i] +
            sd.theta * theta[i])
        # Prior distribution for non-spatial
        # effects
        theta[i] ~ dnorm(0, 1)
    }

    # Prior distribution for spatial effects
    phi[1:Nareas] ~ car.normal(adj[], w[],
        num[], 1)

    # Other prior distributions
    sd.phi ~ dunif(0, 5)
```

```
    sd.theta ~ dunif(0, 5)
    mu ~ dflat()
}

# Object where the results for each
# disease will be saved
results.BYM <- list()

# Run BYM model for each disease
for (i in 1:15) {
    # ATTENTION: Specify spatial weights

    # For unitary weights
    w <- rep(1, length(carto.wb$adj))

    # For adaptive weights:
    index_neighbors <- cbind(rep(1:dim(carto)[1],
        carto.wb$num), carto.wb$adj)
    w <- c(sqrt(results.AdaptiveBYM[[i]]$mean$
        c[index_neighbors[,  1]]) *
         sqrt(results.AdaptiveBYM[[i]]$mean$
        c[index_neighbors[,  2]]))

    # Data
    data <- list(O = apply(Obs[, sex, , causes[i]],
        2, sum), E = apply(Exp[, sex, , causes[i]],
        2, sum), Nareas = dim(Obs)[3], adj = carto.wb$adj,
        w = w, num = carto.wb$num)
    # Initial values
    initials <- function() {
        list(mu = rnorm(1, 0, 1), sd.phi = runif(1,
            0, 1), sd.theta = runif(1, 0,
            1), phi = rnorm(data$Nareas),
```

```
            theta = rnorm(data$Nareas))
    }
    # Variables to retrive
    param <- c("mu", "lambda", "sd.phi",
        "phi", "sd.theta", "theta", "SMR")

    results.BYM[[i]] <- pbugs(data = data,
        inits = initials, parameters.to.save = param,
        model = BYM_model, n.iter = 1e+05,
        n.burnin = 30000, n.chains = 3, DIC = F)
}

# For the model with unitary weights
save(results.BYM, file =
        "Results/results.BYM.unitaryw.RData")
# For the model with adaptive weights
save(results.BYM, file =
        "Results/results.BYM.adaptw.RData")
```

Univariate Leroux model with spatial weights matrices of either unitary weights or using the values obtained from the multivariate analysis of 14 diseases

```
# Univariate Leroux model, WinBUGS code
Leroux_model <- function() {
    # Likelihood
    for (i in 1:Nareas) {
        O[i] ~ dpois(lambda[i])
        # Modeling of the mean for each census
        # tract
        log(lambda[i]) <- log(E[i]) + mu +
            sd.eta * eta[i]
```

```
        # SMR for each census tract
        SMR[i] <- exp(mu + sd.eta * eta[i])
        # Prior distribution for spatial effects
        eta[i] ~ dnorm(mean.eta[i], prec.eta[i])
    }

    for (i in 1:n.adj) {
        sqrt.c.adj[i] <- sqrt(c[adj[i]])
        eta.adj[i] <- eta[adj[i]]
    }

    # Precision of conditioned distribution
    # eta[i]
    prec.eta[1] <- (rho * sqrt(c[1]) * sum
        (sqrt.c.adj[index[1]:index[2]]) +  1 - rho)
    for (i in 2:Nareas) {
        prec.eta[i] <- (rho * sqrt(c[i]) *
            sum(sqrt.c.adj[(index[i] + 1):index[i +
                1]]) + 1 - rho)
    }

    # Mean of conditioned distribution eta[i]
    mean.eta[1] <- (rho * inprod2
        (sqrt.c.adj[index[1]:index[2]],
        eta.adj[index[1]:index[2]]))/(rho *
        sum(sqrt.c.adj[index[1]:index[2]]) +
        1 - rho)
    for (i in 2:Nareas) {
        mean.eta[i] <- (rho *
            inprod2(sqrt.c.adj[(index[i] +
            1):index[i + 1]], eta.adj[(index[i] +
            1):index[i + 1]]))/(rho *
            sum(sqrt.c.adj[(index[i] +
```

```
            1):index[i + 1]]) + 1 - rho)
    }

    # Sum-to-zero restriction for spatial
    # effects
    ceros <- 0
    ceros ~ dnorm(sum.eta, 10)
    sum.eta <- sum(eta[])

    # Other prior distributions
    mu ~ dflat()
    sd.eta ~ dunif(0, 5)
    rho ~ dunif(0, 1)
}

# Object where the results for each
# disease will be saved
results.Leroux <- list()

# Run Leroux model for each disease
for (i in 1:15) {
    # ATTENTION: Specify spatial weights

    # For unitary weights:
    w <- rep(1, length(carto.wb$adj))

    # For adaptive weights:
    w <- c(results.AdaptiveLeroux[[i]]$mean$c)

    # Data
    data <- list(O = apply(Obs[, sex, , causes[i]],
        2, sum), E = apply(Exp[, sex, , causes[i]],
        2, sum), Nareas = dim(Obs)[3], adj = carto.wb$adj,
```

```
        n.adj = length(carto.wb$adj), num = carto.wb$num,
        index = index, c = w)
    # Initial values
    initials <- function() {
        list(mu = rnorm(1, 0, 1), sd.eta = runif(1,
            0, 1), eta = rnorm(data$Nareas))
    }
    # Variables to retrive
    param <- c("mu", "lambda", "sd.eta",
        "eta", "SMR", "rho")

    results.Leroux[[i]] <- pbugs(data = data,
        inits = initials, parameters.to.save = param,
        model = Leroux_model, n.iter = 1e+05,
        n.burnin = 30000, n.chains = 3, DIC = F)

}

# For the model with unitary weights
save(results.Leroux, file =
        "Results/results.Leroux.unitaryw.RData")
# For the model with adaptive weights
save(results.Leroux, file =
        "Results/results.Leroux.adaptw.RData")
```

C.2.2. Estimated spatial weights c_i with multivariate adaptive BYM and Leroux models for each census tract of Valencia according to all 15 diseases in the data set (Figure 6.1 in paper)

```
library(RColorBrewer)
library(sp)

# Results of the multivariate adaptive
# BYM and Leroux models
load("Results/results.AdaptiveBYM.15diseases.RData")
load("Results/results.AdaptiveLeroux.15diseases.RData")

palette <- brewer.pal(7, "YlOrRd")[7:1]
intervals_c <- c(0, 0.3, 0.7, 1, 1.2, 1.4,
    1.6, 2.05)

par(oma = c(2, 0, 2, 0), mar = c(1, 0, 0,
    0), mfrow = c(1, 2), xpd = NA)

# Estimated spatial weights with the
# multivariate adaptive BYM model
# according to all 15 diseases in the
# data set
plot(carto, col = palette[cut(results.AdaptiveBYM$mean$c,
    intervals_c)], xlim = c(-0.4430475, -0.2739941),
    ylim = c(39.45547, 39.55039), main =
    "Adaptive BYM model")
# Estimated spatial weights with the
# multivariate adaptive Leroux model
# according to all 15 diseases in the
# data set
plot(carto, col =palette[cut(results.AdaptiveLeroux$mean$c,
    intervals_c)], xlim = c(-0.4430475, -0.2739941),
    ylim = c(39.45547, 39.55039), main =
    "Adaptive Leroux model")
legend(-0.4870006, 39.49428, levels
    (cut(results.AdaptiveBYM$mean$c,
```

```
    intervals_c)), title = " ", border = NULL,
    fill = paleta, bty = "n")
```

C.2.3. Standardized Mortality Ratios for studied mortality causes in Valencia estimated with the BYM (upper row) and Leroux (lower row) models and with spatial weights matrices of either unitary weights (left) or using the values obtained from the multivariate analysis of 14 diseases (all mortality causes of study except the evaluated cause) (Section C.1, supplementary material in paper)

```
# Results of the BYM model with unitary
# weights
load("Results/results.BYM.unitaryw.RData")
BYM <- results.BYM
# Results of the Leroux model with
# unitary weights
load("Results/results.Leroux.unitaryw.RData")
Leroux <- results.Leroux
# Results of the BYM model with adaptive
# weights
load("Results/results.BYM.adaptw.RData")
BYM_adapt <- results.BYM
# Results of the Leroux model with
# adaptive weights
load("Results/results.Leroux.adaptw.RData")
Leroux_adapt <- results.Leroux
```

```r
# Studied mortality causes
causes_name <- c("AIDS", "Stomach cancer",
    "Colorectal cancer", "Lung cancer", "Prostate cancer",
    "Bladder cancer", "Hematological cancer",
    "Mellitus diabetes", "Dementia",
    "Ischemic heart disease",  "Ictus", "COPD",
    "Liver cirrhosis", "Suicides",
    "Traffic accidents")

Palette.RR <- brewer.pal(7, "BrBG")[7:1]

par(mfrow = c(2, 2), xpd = TRUE)

for (i in 1:15) {
    par(mfrow = c(2, 2), xpd = TRUE)
    aux <- cut(BYM[[i]]$mean$SMR, c(-100,
        0.67, 0.8, 0.91, 1.1, 1.25, 1.5,
        100))
    plot(carto, col = Palette.RR[aux], main =
        paste0("BYM model"),
        xlim = c(-0.4430475, -0.2739941),
        ylim = c(39.45547, 39.55039), cex.main = 1.5)

    aux <- cut(BYM_adapt[[i]]$mean$SMR, c(-100,
        0.67, 0.8, 0.91, 1.1, 1.25, 1.5,
        100))
    plot(carto, col = Palette.RR[aux], main =
        paste0("Adaptive BYM model"),
        xlim = c(-0.4430475, -0.2739941),
        ylim = c(39.45547, 39.55039), cex.main = 1.5)

    aux <- cut(Leroux[[i]]$mean$SMR, c(-100,
        0.67, 0.8, 0.91, 1.1, 1.25, 1.5,
```

```
        100))
    plot(carto, col = Palette.RR[aux], main =
        paste0("Leroux model"),
        xlim = c(-0.4430475, -0.2739941),
        ylim = c(39.45547, 39.55039), cex.main = 1.5)

    aux <- cut(Leroux_adapt[[i]]$mean$SMR,
        c(-100, 0.67, 0.8, 0.91, 1.1, 1.25,
            1.5, 100))
    plot(carto, col = Palette.RR[aux], main =
        paste0("Adaptive Leroux model"),
        xlim = c(-0.4430475, -0.2739941),
        ylim = c(39.45547, 39.55039), cex.main = 1.5)

    par(xpd = NA)
    legend(-0.5253527, 39.61729, c("< 0.67",
        "0.67 - 0.80", "0.80 - 0.91", "0.91 - 1.10",
        "1.10 - 1.25", "1.25 - 1.50", "> 1.50"),
        title = "SMR", border = NULL, fill = Palette.RR,
        bty = "n")

    mtext(causes_name[i], side = 3, cex = 2,
        line = 0, outer = TRUE)
}
```

C.2.4. DIC for the BYM and Leroux models with adaptive and unweighed spatial weights matrices (Table 6.1 in paper)

```
# Studied mortality causes
causes <- c(1:2, 5, 7, 9:12, 15:21)
# Observed mortality cases
```

```
Observados <- apply(Obs[, , , causes], c(2,
    3, 4), sum)

# Function for DICs calculation
CalculaDIC <- function(mu, O, save = FALSE) {
    D <- apply(mu, 1, function(x) {
        -2 * sum(dpois(O, x, log = T))
    })
    Dmedia <- mean(D)
    mumedia <- apply(mu, 2, mean)
    DenMedia <- -2 * sum(dpois(O, mumedia,
        log = T))
    if (save == TRUE) {
        return(c(Dmedia, Dmedia - DenMedia,
            2 * Dmedia - DenMedia))
    }
    cat("D = ", Dmedia, "pD = ", Dmedia -
        DenMedia, "DIC = ", 2 * Dmedia -
        DenMedia, " \n")
}

# Objects where the DIC of the models for
# each disease will be saved
DIC_BYM <- c()
DIC_BYMadapt <- c()
DIC_Leroux <- c()
DIC_Lerouxadapt <- c()

for (j in 1:15) {
    # DIC BYM model with unitary weights
    DIC_BYM[j] <- CalculaDIC(mu=BYM[[j]]$sims.list$lambda,
        O = Observados[sex, , j], save = TRUE)[3]
    # DIC BYM model with adaptive weights
```

```
    DIC_BYMadapt[j] <- CalculaDIC(mu = BYM_adapt[[j]]$
        sims.list$lambda,
        O = Observados[sex, , j], save = TRUE)[3]
    # DIC Leroux model with unitary weights
    DIC_Leroux[j] <- CalculaDIC(mu = Leroux[[j]]$
        sims.list$lambda,
        O = Observados[sex, , j], save = TRUE)[3]
    # DIC Leroux model with adaptive weights
    DIC_Lerouxadapt[j] <- CalculaDIC(mu=Leroux_adapt[[j]]
        $sims.list$lambda,
        O = Observados[sex, , j], save = TRUE)[3]
}

kable(data.frame(causes_name, DIC_BYM, DIC_BYMadapt,
    DIC_Leroux, DIC_Lerouxadapt), digits = 2,
    col.names = c("Causes", "BYM model Adjacency",
        "BYM model Adaptive", "Leroux model Adjacency",
        "Leroux model Adaptive"))
```

C.2.5. CPO for the BYM and Leroux models with adaptive and unweighed spatial weights matrices (Table 6.1 in paper)

```
# Objects where the likelihood of the
# models for each simulation, geographic
# unit and disease will be saved
likelihood_BYM <- array(NA, dim = c(1002,
    531, 15))
likelihood_BYMadapt <- array(NA, dim = c(1002,
    531, 15))
likelihood_Leroux <- array(NA, dim = c(1002,
    531, 15))
```

```
likelihood_Lerouxadapt <- array(NA, dim = c(1002,
    531, 15))

# Objects where the CPO of the models for
# each geographic unit and disease will
# be saved
CPO_BYM <- array(NA, dim = c(531, 15))
CPO_BYMadapt <- array(NA, dim = c(531, 15))
CPO_Leroux <- array(NA, dim = c(531, 15))
CPO_Lerouxadapt <- array(NA, dim = c(531,
    15))

# Likelihood of the models for each
# disease, geographic unit and MCMC
# simulation
for (i in 1:15) {
    for (j in 1:531) {
        for (k in 1:1002) {
            likelihood_BYM[k, j, i] <- dpois
                (Observados[sex,
                j, i], BYM[[i]]$sims.list$lambda[k,
                j])
            likelihood_BYMadapt[k, j, i] <- dpois
                (Observados[sex,
                j, i], BYM_adapt[[i]]$sims.list$lambda[k,
                j])
            likelihood_Leroux[k, j, i] <- dpois
                (Observados[sex,
                j, i], Leroux[[i]]$sims.list$lambda[k,
                j])
            likelihood_Lerouxadapt[k, j,
                i] <- dpois
                (Observados[sex,
```

```
                j, i], Leroux_adapt[[i]]$sims.list$
                lambda[k,  j])
        }
    }
}

# CPO of the models for each disease and
# geographic unit
for (i in 1:15) {
    for (j in 1:531) {
        CPO_BYM[j, i] <- 1/(mean(1/likelihood_BYM[,
            j, i]))
        CPO_BYMadapt[j, i] <- 1/(mean(1/
            likelihood_BYMadapt[,  j, i]))
        CPO_Leroux[j, i] <- 1/(mean(1/
            likelihood_Leroux[,  j, i]))
        CPO_Lerouxadapt[j, i] <- 1/(mean(1/
            likelihood_Lerouxadapt[,  j, i]))
    }
}

# Total CPO of the models
CPO_TOTAL_BYM <- apply(apply(CPO_BYM, 2,
    function(x) {
        log(x)
    }), 2, sum)
CPO_TOTAL_Leroux <- apply(apply(CPO_Leroux,
    2, function(x) {
        log(x)
    }), 2, sum)
CPO_TOTAL_BYMadapt <- apply(apply(CPO_BYMadapt,
    2, function(x) {
        log(x)
```

```
    }), 2, sum)
CPO_TOTAL_Lerouxadapt <- apply(apply(CPO_Lerouxadapt,
    2, function(x) {
        log(x)
    }), 2, sum)

kable(data.frame(causes_name, CPO_TOTAL_BYM,
    CPO_TOTAL_BYMadapt, CPO_TOTAL_Leroux,
    CPO_TOTAL_Lerouxadapt), col.names = c("Causes",
    "BYM model Adjacency", "BYM model Adaptive",
    "Leroux model Adjacency", "Leroux model Adaptive"),
    digits = 2)
```

Bibliography

Adin, A., Goicoa, T., and Ugarte, M. D. (2019a). Online relative risks/rates estimation in spatial and spatio-temporal disease mapping. *Computer Methods and Programs in Biomedicine*, 172:103–116.

Adin, A., Lee, D., Goicoa, T., and Ugarte, M. D. (2019b). A two-stage approach to estimate spatial and spatio-temporal disease risks in the presence of local discontinuities and clusters. *Statistical Methods in Medical Research*, 28(9):2595–2613. PMID: 29651927.

Adin, A., Martinez-Beneito, M. A., Botella-Rocamora, P., Goicoa, T., and Ugarte, M. D. (2017). Smoothing and high risk areas detection in space-time disease mapping: a comparison of P-splines, autoregressive and moving average models. *Stochastic and Environmental Research and Risk Assessment*, 31:403–415.

Agarwal, D. K., Gelfand, A. E., and Citron-Pousty, S. (2002). Zero-inflated models with application to spatial count data. *Environmental and Ecological Statistics*, 9:341–355.

Aguilar-Palacio, I., Martinez-Beneito, M., Rabanaque, M., Borrell, C., Cirera, L., Daponte, A., Domínguez-Berjón, M., Gandarillas, A., Gotsens, M., Lorenzo-Ruano, P., Marí-Dell'Olmo, M., Nolasco, A., Saez, M., Sánchez-Villegas, P., Saurina, C., and Martos, C. (2017). Diabetes mellitus mortality in spanish cities: Trends and geographical inequalities. *Primary Care Diabetes*, 11(5):453 – 460.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10.

Arab, A. (2015). Spatial and spatio-temporal models for modeling epidemiological data with excess zeros. *International Journal of Environmental Research and Public Health*, 12(9):10536–10548.

Assunçao, R. M. (2003). Space varying coefficient models for small area data. *Environmetrics*, 14:453–473.

Barnard, J., McCulloch, R., and Meng, X.-L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with applications to shrinkage. *Statistica Sinica*, 10:1281–1311.

Bauer, C., Wakefield, J., Rue, H., Self, S., Feng, Z., and Wang, Y. (2016). Bayesian penalized spline models for the analysis of spatio-temporal count data. *Statistics in Medicine*, 35(11):1848–1865.

Bayarri, M. J., Berger, J., and Datta, G. S. (2008). Objective Bayes testing of Poisson versus inflated Poisson models. In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh*, volume 3, pages 105–121. Institute of Mathematical Statistics.

Benach, J., Martínez, J. M., Yasui, Y., Borrell, C., Pasarín, M. I., Español, E., and Benach, N. (2004). *Atles de mortalitat en árees petites a Catalunya*. Universitat Pompeu Fabra.

Benach, J. and Yasui, Y. (1999). Geographical patterns of excess mortality in spain explained by two indices of deprivation. *Journal of Epidemiology & Community Health*, 53(7):423–431.

Benach, J., Yasui, Y., Borrell, C., Rosa, E., Pasarín, M. I., Benach, N., Español, E., Martínez, J. M., and Daponte, A. (2001). *Atlas de mortalidad en áreas pequeñas en España, 1987-1995*. Universitat Pompeu Fabra.

Benach de Rovira, J. and Martínez Martínez, J. M., editors (2013). *Atlas de mortalidad en municipios y unidades censales de España 1984-2004*. Fundación BBVA.

Berger, J. (2006). The case for objective Bayesian analysis. *Bayesian Analysis*, 1:385–402.

Bernardinelli, L., Clayton, D., and Montomoli, C. (1995). Bayesian estimates of disease maps: How important are priors? *Statistics in Medicine*, 14:2411–2431.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 36:192–236.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathemathics*, 43:1–21.

Best, N., Richardson, S., and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59.

Best, N. G., Arnold, Arnold, R., Thomas, A., Waller, L. A., and Conlon, E. M. (1999). Bayesian models for spatially correlated disease and exposure data. In *Bayesian Statistics 6*. Oxford University Press.

Borrell, C., Cano Serral, G., Martínez-Beneito, M. A., Dell'Olmo, M., Marc, Rodríguez Sanz, M., and Grupo MEDEA (2009). *Atlas de mortalidad en ciudades de España (1996-2003)*.

Borrell, C., Marí Dell'Olmo, M., Serrall, G., Martínez-Beneito, M. A., and Gotséns, M. (2010). Inequalities in mortality in small areas of

eleven Spanish cities (the multicenter MEDEA Project). *Health & Place*, 16:703–711.

Botella-Rocamora, P., López-Quílez, A., and Martinez-Beneito, M. A. (2012). Spatial moving average risk smoothing. *Statistics in Medicine*, 32:2595–2612.

Botella-Rocamora, P., Martinez-Beneito, M. A., and Banerjee, S. (2015). A unifying modeling framework for highly multivariate disease mapping. *Statistics in Medicine*, 34(9):1548–1559.

Botella Rocamora, P., Pérez-Panadés, J., and Martinez-Beneito, M. A. (2017). Estimating the geographical distribution of diseases: a statistical problem. *Boletín de Estadística e Investigación Operativa*, 33(1):4–21.

Brezger, A., Fahrmeir, L., and Hennerfeind, A. (2007). Adaptive Gaussian Markov random fields with applications in human brain mapping. *Applied Statistics*, 56(3):327–345.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational & Graphical Statistics*, 7:434–455.

Carter, C. K. and Kohn, R. (1996). Markov chain monte carlo in conditionally gaussian state space models. *Biometrika*, 83(3):589–601.

Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2020). Shiny: Web Application Framework for R. R package version 1.4.0.2. URL https://cran.r-project.org/web/packages/shiny/index.html.

Cheng, J., Karambelkar, B., and Xie, Y. (2019). Leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.3. URL https://CRAN.R-project.org/package=leaflet.

Congdon, P. (2008). A spatially adaptive conditional autoregressive prior for area health data. *Statistical Methodology*, 5:552–563.

Dalrymple, M. L., Hudson, I. L., and Ford, R. P. K. (2003). Finite mixture, zero-inflated Poisson and hurdle models with application to SIDS. *Computational Statistics & Data Analysis*, 41(3-4):491–504.

Denison, D. G. and Holmes, C. C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, 57:143–149.

Duncan, E. W., White, N. M., and Mengersen, K. (2017). Spatial smoothing in bayesian models: a comparison of weights matrix specifications and their impact on inference. *International Journal of Health Geographics*, 16(1):47.

Earnest, A., Morgan, G., Mengersen, K., Louise, R., Richard, S., and Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models. *International Journal of Health Geographics*, pages 6–54.

Elliott, P., Wakefield, J. C., Best, N., and Briggs, D. J., editors (2000). *Spatial Epidemiology*. Oxford University Press.

Esnaola, S., Montoya, I., Calvo, M., Aldasoro, E., Audicana, C., Ruiz, R., and Ibañez, B. (2010). *Atlas de mortalidad en áreas pequeñas de la Comunidad Autónoma del País Vasco (1996-2003)*. Departamento de Sanidad y Consumo, Vitorie-Gasteiz.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, 3 edition.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gentle, J. E. (2007). *Matrix Algebra. Theory, Computations, and Applications in Statistics*. Springer-Verlag.

Goicoa, T., Etxeberria, J., and Ugarte, M. D. (2016). Splines in disease mapping. In *Handbook of Spatial Epidemiology*, chapter 12, pages 225–238. CRC Press.

Gracia, E., López-Quílez, A., Marco, M., and Lila, M. (2017). Mapping child maltreatment risk: a 12-year spatio-temporal analysis of neighborhood influences. *International Journal of Health Geographics*, 16(38).

Graham, H. (1996). Smoking prevalence among women in the european community 1950–1990. *Social Science & Medicine*, 43(2):243 – 254.

Green, P. and Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1–16.

Gschlößl, S. and Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers*, 49(3):531–552.

Heilbron, D. C. (1994). Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36:531–547.

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, 59:317–322.

Knorr-Held, L. and Raßer, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(13-21):2045–2060.

Kuhnert, P. (2003). New methodology and comparisons for the analysis of binary data using bayesian and tree based methods.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to detects on manufacturing. *Technometrics*, 34:1–14.

Lang, S., Fronk, E.-M., and Fahrmeir, L. (2002). Function estimation with locally adaptive dynamic models. *Computational statistics*, 17(4):479–499.

Last, J., editor (2001). *A dictionary of epidemiology*. Oxford University Press, 4 edition.

Lawson, A. B. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology (3rd edition)*. CRC Press.

Lawson, A. B., Biggeri, A., Boehning, D., Lesaffre, E., Viel, J.-F., Clark, A., Schlattmann, P., and Divino, F. (2000). Disease mapping models: an empirical evaluation. *Statistics in Medicine*, 19(17/18):2217–2242.

Lawson, A. B. and Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21:359–370.

Lee, D. (2011). A comparison of conditional autoregressive models used in bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*, 2:79–89.

Leroux, B. G., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 135–78.

López-Abente, G., Aragonés, N., Pérez-Gómez, B., Pollán, M., García-Pérez, J., Ramis, R., and Fernández-Navarro, P. (2014). Time trends in municipal distribution patterns of cancer mortality in Spain. *BMC Cancer*, 14(1):535.

López-Abente, G., Pollán, M., Escolar, A., Errezola, M., and Abraira, V. (2001). *Atlas de mortalidad por cáncer y otras causas en España, 1978-1992*. Instituto de Salud Carlos III.

López-Abente, G., Ramis, R., Pollán, M., Aragonés, N., Pérez-Gómez, B., Gómez-Barroso, D., Carrasco, J. M., Lope-Carvajal, V.,

García-Pérez, J., Boldo, E., and García-Mendizábal, M. J. (2006). *Atlas municipal de mortalidad por cáncer en España, 1989-1998*. Instituto de Salud Carlos III.

Lu, H. and Carlin, B. P. (2005). Bayesian areal Wombling for geographical boundary analysis. *Geographical Analysis*, 35:265–285.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal Wombling via adjacency modelling. *Environmental and Ecological Statistics*, 14:433–452.

Lunn, D., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337.

Ma, H. and Carlin, B. P. (2007). Bayesian multivariate areal Wombling for multiple disease boundary analysis. *Bayesian Analysis*, 2(2):281–302.

Ma, H., Carlin, B. P., and Banerjee, S. (2010). Hierarchical and joint site-edge methods for medicare hospice service region boundary analysis. *Biometrics*, 66:355–364.

MacNab, Y., Kemetic, A., Gustafson, P., and Sheps, S. (2006a). An innovative application of Bayesian disease mapping methods to patient safety research: A Canadian adverse medical event study. *Statistical Methods in Medical Research*, 25:3960–3980.

MacNab, Y. C. (2007). Spline smoothing in Bayesian disease mapping. *Environmetrics*, 18:727–744.

MacNab, Y. C. (2016a). Linear models of coregionalization for multivariate lattice data: a general framework for coregionalized multivariate CAR models. *Statistics in Medicine*, pages 3827–3850.

MacNab, Y. C. (2016b). Linear models of coregionalization for multivariate lattice data: Order-dependent and order-free cMCARs. *Statistical Methods in Medical Research*, 25(4):1118–1144.

MacNab, Y. C. (2018). Some recent work on multivariate Gaussian Markov random fields. *TEST*, 27(3):497–541.

MacNab, Y. C., Kmetic, A., Gustafson, P., and Sheps, S. (2006b). An innovative application of bayesian disease mapping methods to patient safety research: A canadian adverse medical event study. *Statistics in Medicine*, 25:3960–3980.

Marcelino-Rodríguez, I., Elosua, R., del Cristo Rodríguez Pérez, M., Fernández-Bergés, D., Guembe, M. J., Alonso, T. V., Félix, F. J., González, D. A., Ortiz-Marrón, H., Rigo, F., Lapetra, J., Gavrila, D., Segura, A., Fitó, M., Peñafiel, J., Marrugat, J., and de León, A. C. (2016). On the problem of type 2 diabetes-related mortality in the canary islands, spain. the darios study. *Diabetes Research and Clinical Practice*, 111:74 – 82.

Marco, M., López-Quílez, A., Conesa, D., Gracia, E., and Lila, M. (2017). Spatio-temporal analysis of suicide-related emergency calls. *International Journal of Environmental Research and Public Health*, 14(7):735.

Martinez-Beneito, M. A. (2013). A general modelling framework for multivariate disease mapping. *Biometrika*, 100(3):539–553.

Martinez-Beneito, M. A. and Botella Rocamora, P. (2019). *Disease Mappping: From foundations to Multidimensional Modeling*. CRC Press.

Martinez-Beneito, M. A., Botella-Rocamora, P., and Banerjee, S. (2017). Towards a multidimensional approach to Bayesian disease mapping. *Bayesian Analysis*, 12:239–259.

Martínez-Beneito, M. A., López Quílez, A., Amador, A., Melchor, I., Botella Rocamora, P., Abellán, C., Abellán, J., Verdejo, F., Zurriaga, O., Vanaclocha, H., and Escolano, M. (2005). *Atlas de mortalidad de la Comunidad Valenciana, 1991-2000*. Generalitat Valenciana.

Martinez-Beneito, M. A., López-Quílez, A., and Botella-Rocamora, P. (2008). An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine*, 27:2874–2889.

Moraga, P. (2017). Spatialepiapp: A shiny web application for the analysis of spatial and spatio-temporal disease data. *Spatial and Spatio-temporal Epidemiology*, 23:47–57.

Moraga, P. (2019). *Geospatial Health Data*. CRC Press.

Morris, M. C., Marco, M., Bailey, B., Ruiz, E., Im, W., and Goodin, B. (2019). Opioid prescription rates and risk for substantiated child abuse and neglect: A bayesian spatiotemporal analysis. *Drug and Alcohol Dependence*, 205.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, 33:341–365.

Musenge, E., Freeman Chirwa, T., Kahn, K., and Vounatsou, P. (2013). Bayesian analysis of zero inflated spatiotemporal HIV/TB child mortality data through the INLA and SPDE approaches: Applied to data observed between 1992 and 2010 in rural North East South Africa. *International Journal of Applied Earth Observation and Geoinformation*, 22:86–98.

Natarajan, R. and McCulloch, C. E. (1995). A note on the existence of the posterior distribution for a class of mixed models for binomial responses. *Biometrika*, 82(3):639–643.

Neelon, B., Chang, H. H., Ling, Q., and Hastings, N. S. (2014). Spatiotemporal hurdle models for zero-inflated count data: Exploring

trends in emergency department visits. *Statistical Methods in Medical Research*, 25(6):2558–2576.

Neelon, B., Ghosh, P., and Loebs, P. F. (2013). A spatial Poisson hurdle model for exploring geographic variation in emergency department visits. *Journal of the Royal Statistical Society, Series A*, 176(2):389–413.

Neelon, B. H., O'Malley, A. J., and Normand, S.-L. T. (2010). A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use. *Stat Modelling*, 10(4):421–439.

Nieto-Barajas, L. and Bandyopadhyay, D. (2013). A zero-inflated spatial gamma process model with application to disease mapping. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(2):137–158.

Ocaña, R., Sánchez-Cantalejo, C., Fernández Ajuria, A., Ruiz Ramos, M., Mayoral Cortés, J., Méndez Martínez, C., Sáez Zafra, M., Barceló, M. A., Saurina, C., and Lertxundi, A. (2007). *Atlas de mortalidad de las capitales de provincia de Andalucía, 1992-2002*. Escuela Andaluza de Salud Pública.

Ocaña, R., Sánchez-Cantalejo, C., Toro, S., and Mayoral Cortés, J. (2010). *Atlas interactivo de mortalidad en Andalucía. 1981-2006*. Escuela Andaluza de Salud Pública.

Ocaña Riola, R. (2007). The misuse of count data aggregated over time for disease mapping. *Statistics in Medicine*, 26:4489–4504.

Puigpinos-Riera, R., Mari-Dell'Olmo, M., Gotsens, M., Borrell, C., Serral, G., Ascaso, C., Calvo, M., Daponte, A., Dominguez-Berjon, F. M., Esnaola, S., Gandarillas, A., Lopez-Abente, G., Martos, C. M., Martinez-Beneito, M. A., Montes-Garcia, A., Montoya, I., Nolasco, A., Pasarin, I. M., Rodriguez-Sanz, M., Saez, M., and Sanchez-Villegas, P. (2011). Cancer mortality inequalities in urban areas: a bayesian

small area analysis in spanish cities. *International Journal of Health Geographics*, 10:6.

Raftery, A. E. and Banfield, J. D. (1991). Stopping the Gibbs sampler, the use of morphology, and other issues in spatial statistics (discussion of Besag et al.). *Annals of the Institute of Statistical Mathemathics*, 43:32–43.

Richardson, S., Thomson, A., Best, N., and Elliot, P. (2004). Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112(9):1016–1025.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory & Applications*. Chapman & Hall/CRC.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392.

Schrödle, B. and Held, L. (2011). A primer on disease mapping and ecological regression using INLA. *Computational Statistics*, 26:241–258.

Scott, J. G. and Berger, J. O. (2010). Bayesian and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Annals of Statistics*, 38(5):2587–2619.

Sievert, C., Parmer, C., Hocking, T., Chamberlain, S., Ram, K., Corvellec, M., and Despouy, P. (2020). Plotly: Create Interactive Web Graphics via 'plotly.js'. R package version 4.9.2.1. URL https://cran.r-project.org/web/packages/plotly/index.html.

Song, H.-R., Lawson, A., D'Agostino Jr, R. B., and Liese, A. D. (2011). Modeling type 1 and type 2 diabetes mellitus incidence in youth: an application of Bayesian hierarchical regression for sparse small area data. *Spatial and Spatiotemporal Epidemiology*, 2(1):23–33.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64:583–641.

Sturtz, S., Ligges, U., and Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3):1–16.

Ugarte, M. D., Ibañez, B., and Militino, A. F. (2004). Testing for Poisson zero inflation in disease mapping. *Biometrical Journal*, 46:526–539.

Ugarte, M. D., Ibañez, B., and Militino, A. F. (2006). Modelling risks in disease mapping. *Statistical Methods*, 15:21–35.

Upfill-Brown, A. M., Lyons, H. M., Pate, M. A., Shuaib, F., Baig, S., Hu, H., Eckhoff, P. A., and Chabot-Couture, G. (2014). Predictive spatial risk model of poliovirus to aid prioritization and hasten eradication in Nigeria. *BMC Medicine*, 12(92).

Van Der Broek, J. (1995). A score test for zero inflation in a Poisson distribution. *Biometrics*, 51(2):738–743.

Xie, Y., Cheng, J., and Tan, X. (2020). DT: A Wrapper of the JavaScript Library 'DataTables'. R package version 0.13. URL https://cran.r-project.org/web/packages/DT/index.html.

Zurriaga, O., Martínez-Beneito, M. A., Botella-Rocamora, P., López-Quílez, A., Melchor, I., Amador, A., Vanaclocha, H., and Nolasco, A. (2010). Spatio-temporal mortality atlas of Comunitat Valenciana. (http://www.geeitema.org/AtlasET/index.jsp?idioma=I. Accessed: May, 2nd 2016).

Zurriaga, O., Vanaclocha, H., Martínez-Beneito, M. A., and Botella Rocamora, P. (2008). Spatio-temporal evolution of female lung cancer mortality in a region of Spain: is it worth taking migration into account? *BMC Cancer*, 8(35):1.