FACULTAT DE MATEMÀTIQUES
DEPARTAMENT D'ESTADISTICA I INVESTIGACIÓ OPERATIVA

UNIVERSITAT DE VALÈNCIA

Métodos estadísticos en la detección de focos de riesgo en brotes epidémicos

Tesis doctoral presentada por:
D. Miguel Ángel Martínez Beneito
y dirigida por:
D. Antonio López Quílez

D. Antonio López Quílez, profesor titular de Estadística e Investigación Operativa del Departament d'Estadística i Investigació Operativa de la Universitat de València

CERTIFICA que la presente memoria de investigación:

"Métodos estadísticos en la detección de focos de riesgo en brotes epidémicos"

ha sido realizada bajo su dirección en el Departament d'Estadística i Investigació Operativa por Miguel Ángel Martínez Beneito.

Y para que así conste, firma el siguiente certificado.

Burjassot, 20 de Julio de 2005

Fdo: Antonio López Quílez

A Paloma,
por todo el tiempo que le he robado con este trabajo.

Agradecimientos

Después de varios años de trabajo, una de las principales satisfacciones que produce terminar una tesis es tener la oportunidad de llenar unas líneas dedicadas a tus seres queridos. Seguramente esta sección sea la menos relevante desde el punto de vista científico, pero sin lugar a dudas es la más importante desde la perspectiva personal, y para muchos la parte que leerán con mayor atención.

Quisiera comenzar mis agradecimientos recordando a Juan Ferrándiz, Gudo para los que le conocíamos. Gudo ha formado parte de este trabajo ya que comenzó siendo su codirector junto a Antonio López. Lamentablemente nos dejó físicamente en Octubre de 2003, aunque para suerte de los que le conocimos nos acompaña y acompañará en nuestra memoria. A los que compartimos su tiempo nos enseñó que dedicarse plenamente a la estadística no significa renunciar a otras muchas inquietudes. En él hemos podido encontrar un ejemplo de humanista, en el sentido renacentista, adornado con una actitud desprendida, una sonrisa para compartir, una frase de aliento para regalar y un comentario interesante que ofrecer.

En segundo lugar me gustaría agradecer a Antonio López todo el esfuerzo y la atención que me ha dedicado durante estos años de trabajo. Antonio ha demostrado, a parte ser de un excelente director científico, saber infun-

dir una palabra de ánimo en el momento oportuno y convertir el trabajo cotidiano en un aprendizaje continuo. Gracias a él, el tiempo dedicado a realizar este estudio se ha convertido en una experiencia muy enriquecedora, el comienzo de una larga amistad y, espero que en un futuro, una fructífera colaboración.

Sin salir del ámbito científico, también quisiera nombrar a toda la gente del GEeiTema (Grup d'Estadística espacial i temporal en Epidemiología y medi ambient) ya que han sabido formar un grupo de trabajo en el que se funden el marco académico-estadístico con un ambiente jovial. Esta mezcla se suele aderezar (y en ocasiones incluso regar) de lo que gustamos en llamar semi(ce)narios, donde se manifiesta plenamente el espíritu de este grupo.

También me considero en deuda con mis compañeros de trabajo del Área de Epidemiología, ya que me han proporcionado un ambiente de trabajo cálido y un escenario en el que dar sentido a los cálculos que siempre me ha gustado hacer. Quisiera destacar además a Herme y Oscar por su sensibilidad con la estadística y los estadísticos, y por el trato tan humano con el que comparten sus horas de trabajo. Fruto de este trato hemos podido entablar una relación de amistad que va mucho más allá de lo laboral y de la que me siento especialmente orgulloso. Tampoco quisiera olvidarme de Juanjo Abellán con quien compartí varios años de trabajo en el Área de Epidemiología y quien guió mis primeros pasos en esto de la estadística espacial.

No puedo dejar pasar esta ocasión sin agradecer a mi familia (incluida la de Hondón) todo el apoyo y ánimo que me han ofrecido durante todos estos años. Después de preguntarme tantas veces "¿Y aún te queda mucho?" ahora sí que ya les puedo contestar que por fin he terminado. A mi padre, mi madre y mi hermana porque sus virtudes han sido un ejemplo y una guía

para dar sentido a este largo camino que llamamos vida. También a Carla y Andrea por ayudarme a buscar tantas veces el 5 que no me salía y que ponía fin a este trabajo. Y como no, a Paloma que ha sido la que más ha padecido esta tesis y a quien más momentos le debo por ella, espero que me los cobre todos uno a uno.

También quisiera agradecer a mis amigos de Villena, los alpisteros, su colaboración inconsciente en este trabajo ya que, sin ellos saberlo, han sabido llenar los ratos de ocio necesarios para olvidar los sinsabores y momentos de desánimo que he sufrido en el transcurso del trabajo.

Por último, quisiera recordar a todos los "voluntarios" de la ciudad de Alcoi (especialmente los casos) su colaboración indispensable para la realización del estudio. En estadística muchas veces olvidamos que los insignificantes números con los que nos "divertimos" pueden ocultar una situación personal, en ocasiones dramática. El dar soluciones a aquello que hay detrás de cada número justifica la labor de la estadística y en concreto el trabajo que aquí comienza.

Índice general

Capítulo 1

Detección de clusters en estudios epidemiológicos

El presente trabajo, ya desde su título, tiene una vocación claramente aplicada. En concreto, el campo de aplicación en el que se encuadra es el de las ciencias de la salud, en particular la epidemiología. Esta disciplina trata sobre "El estudio de la distribución y determinantes de los estados o eventos relacionados con la salud en poblaciones específicas y la aplicación de este estudio al control de los problemas de salud" Last (2001) [60]. Es en este campo de la medicina donde se ha producido una de las simbiosis más fructíferas con la estadística. La epidemiología por su parte ha supuesto para la estadística un marco donde dotar de sentido y aplicación a un gran número de sus técnicas. Mientras tanto la epidemiología se ha visto enriquecida en gran medida por la estadística, ya que le ha brindado las herramientas para el manejo, explotación y análisis de datos, la materia prima básica en la que se sustenta esta ciencia.

Como mayor prueba de la convivencia y aprovechamiento mutuo entre ambas disciplinas cabe señalar que incluso se han desarrollado líneas de investigación en estadística con aplicación casi exclusiva en el ámbito de la

epidemiología. Un ejemplo de este hecho sería el análisis de supervivencia, que ya sea desde un punto de vista clínico o poblacional, se ha desarrollado conforme los estudios epidemiológicos exigían nuevos avances metodológicos dadas las necesidades y problemas con que se encontraban. Un segundo ejemplo serían las aplicaciones de la estadística espacial a las ciencias de la salud. Concretamente, el análisis geográfico de riesgos ha supuesto un campo de aplicación y desarrollo de la estadística en el ámbito de la epidemiología que ha ocupado un gran número de páginas de la bibliografía científica de ambas áreas de conocimiento.

Este trabajo tiene como objetivo la realización de una aportación metodológica dentro del marco de la epidemiología y la estadística, aunque sin perder de vista su utilidad práctica. En concreto, su desarrollo se encuadra dentro del campo estadístico de detección de clusters o agrupaciones de observaciones, que si bien no ha sido un campo de aplicación exclusivo de la epidemiología ha alcanzado dentro de ella gran popularidad y utilidad práctica. Como aplicación específica de los desarrollos metodológicos realizados y como motivación de éstos, se presenta el estudio de una serie de brotes de neumonía por *Legionella* que se sucedieron en la ciudad de Alcoi entre septiembre de 1999 y noviembre de 2003. En los estudios previos que se realizaron de estos brotes quedó patente la necesidad de aplicar técnicas estadísticas de detección de clusters como apoyo al estudio del brote epidémico. En particular se hacía necesaria la propuesta de desarrollos metodológicos que permitieran extraer conclusiones ricas que respondieran a ciertas cuestiones específicas sobre el origen de los brotes, ese es el objetivo principal del trabajo que se va a acometer.

Respecto a la estructura de exposición del resto de la tesis, en este capítulo vamos a introducir el estudio de clusters en análisis epidemiológicos, así como una descripción de los brotes de neumonía por *Legionella* en la

ciudad de Alcoi que justifican los desarrollos metodológicos objetivo de esta tesis. En el segundo capítulo se realiza una breve introducción a la teoría de procesos puntuales y su aplicación a la detección de agrupaciones de casos en un contexto epidemiológico. En el tercer capítulo se discute la inferencia basada en modelos de mixturas de distribuciones con sus problemas y limitaciones, así como su aplicación a la detección de clusters geográficos. En el cuarto capítulo se desarrollará una propuesta, basada en modelos de mixturas, para la detección geográfica de clusters de enfermedades. En los capítulos quinto y sexto se describirán los resultados de la propuesta del capítulo cuarto a un conjunto de datos simulados, para valorar la calidad de ajuste del modelo, y al estudio de los brotes de Alcoi respectivamente. Por último, en el capítulo séptimo se presentarán las conclusiones, así como posibles líneas de desarrollo que podrían complementar y mejorar los desarrollos metodológicos propuestos.

1.1. Clustering

El análisis de *clusters* o agrupaciones de observaciones tiene una gran tradición en la literatura estadística y epidemiológica. Desde un punto de vista estadístico, el estudio de clusters se puede entender como la determinación de distintos patrones o agrupaciones sobre un conjunto de datos observados, de manera que cada uno de éstos se pueda asociar a alguna de las agrupaciones según las características de cada individuo. El procedimiento general en este tipo de aplicaciones se fundamenta en la definición de una medida de similitud entre individuos y en base a esta medida se definen las agrupaciones de casos y la asociación de cada observación a cada uno de estos grupos. A diferencia de las técnicas estadísticas de clasificación o análisis discriminante, en el análisis de clusters no existe ningún

grupo o patrón conocido a priori por lo que la definición de las agrupaciones depende exclusivamente de los datos disponibles. Es precisamente en este detalle donde reside una de las mayores dificultades de este conjunto de técnicas. Además, en un gran número de problemas ni siquiera se sabe el número de agrupaciones existentes en los datos y en muchos de estos casos dicho número es una de las características de interés objetivo del estudio estadístico.

En ocasiones el análisis de clusters se aplica sobre un conjunto de datos correspondientes a distintas localizaciones geográficas y las agrupaciones se forman en base a la ubicación de dichos individuos. El conjunto de técnicas que se aplican en estos casos se denominan herramientas de *análisis de clusters geográfico o espacial*. El objetivo de estos estudios es la determinación, si existen, de agrupaciones geográficas de casos de enfermedades en base a su distribución sobre la región de estudio. La localización de agrupaciones geográficas en la incidencia de una enfermedad denotará la existencia de alguna fuente o factor de riesgo detonante de dicha agregación. Por tanto la existencia y localización de clusters de casos proporciona información sobre el origen, factores de riesgo o cualquier otro aspecto relacionado con la enfermedad de estudio. Es este tipo de análisis en el que nos vamos a centrar en el desarrollo de la tesis.

El análisis geográfico de clusters, aún siendo una aplicación particular del análisis estadístico de clusters, tiene una gran tradición y una amplia literatura de la que existen diversas monografías específicas, valga como ejemplo Lawson y Denison (2002) [64] o Alexander y Boyle (1996) [3]. Incluso, este campo se ha convertido en una de las líneas de investigación más activas y de mayor desarrollo del mundo de la epidemiología y de la estadística espacial. En lo sucesivo no haremos distinción entre los estudios de clusters en general y los estudios de clusters de tipo geográfico, entendiendo que se

hace alusión a estos últimos mientras no se diga lo contrario.

Existe un amplio abanico de definiciones de cluster dentro del contexto epidemiológico. Dichas definiciones pueden variar desde unos términos muy específicos, "... 5 casos que representan al menos un incremento en el riesgo de al menos 5 unidades vistos por un único médico (o un reducido número de colegas) sobre un corto periodo...", hasta una definición bastante más general, "... dos o más casos aparecidos cerca ...", tal y como se expone en Wartenberg (2001) [99]. De todas maneras, una definición útil de clusters debería ser a la vez general y flexible. En Lawson (2001) [61] se da una definición de clusters que consideramos bastante apropiada por su generalidad y claridad: "cualquier área dentro de la región de estudio con riesgo significativamente elevado", donde se entiende como riesgo la probabilidad de albergar un caso. Cabe señalar que en el caso de estudiar datos de los que se conoce su localización exacta, existen planteamientos y definiciones más formales desde el punto de vista matemático de cluster geográfico, desarrollados bajo la teoría de procesos puntuales. En Cressie (1993) [31] o Diggle (2003) [39] se pueden encontrar detalles más precisos de estas formulaciones que se estudiarán con más detenimiento en el capítulo 2 de esta tesis.

Besag y Newell (1991) [14], Alexander y Boyle (1996) [3], Alexander y Boyle (2000) [4], Lawson (2001) [61] y Lawson y Denison (2002) [65] clasifican los estudios de clusters en varios tipos. No existe un criterio único claramente definido en cuanto a esta clasificación, ya que ésta depende de la metodología y las tendencias dominantes en cada época y del autor que las establezca. Sin embargo, salvando las diferencias relacionadas con la fecha de publicación de los manuscritos, parece existir un consenso más o menos amplio en ciertos aspectos que seguidamente resumimos.

En primer lugar, se suele distinguir entre estudios de clustering de tipo

individual o *general*. En el primer caso se estudia la existencia de un número excesivo de casos alrededor de ciertas localizaciones de la región de estudio concretas aunque desconocidas. En el segundo, se estudia la existencia de una distribución de los casos más heterogénea de lo que podría resultar razonable como consecuencia de la variación de la densidad de población, de forma que resulta necesaria la acción de un cambio geográfico en el riesgo para describir de forma adecuada el patrón de casos observado. En este caso, el cambio geográfico es producto de la distribución de los factores de riesgo en la población y dicha distribución suele variar de forma suave y constante.

Si consideramos una división de la región de estudio, cuando los casos sigan un patrón de agregación general observaremos que la varianza del número de casos observados sobre cada celda es superior a su valor esperado. Este hecho viola la hipótesis de que los casos siguen una distribución de Poisson con riesgo constante para todas las regiones. Este exceso de variabilidad, consecuencia de la agregación de casos al variar geográficamente la distribución de los factores de riesgo, se dice *extra-varianza* de Poisson y su detección denota la presencia de un factor que produce agregaciones en la distribución de los casos.

A diferencia de los estudios de clustering individual, los estudios de clustering general contrastan aspectos globales de la tendencia a agruparse de la variable de interés, por lo que resultan poco sensibles a agrupaciones particulares en ciertas regiones concretas. Por tanto, si se dispone de una única agrupación de la incidencia alrededor de una localización concreta, los estudios de clustering individual serán mucho más eficientes que los estudios de clustering general. Así, las metodologías de ambos tipos de estudios, obviamente, serán diferentes y se adecuarán a la situación para las que han sido ideadas. De esta manera, antes de acometer cualquier estudio de detección de clusters resulta necesario tener claro cual es el tipo de agregación que

se quiere detectar y en base a éste emplear las técnicas adecuadas para ese objetivo.

Dependiendo del tipo de datos disponibles se puede hacer una nueva clasificación de los estudios de detección de clusters. El primero de estos casos corresponde a aquellos estudios en los que se conoce la localización exacta de un conjunto de casos de cierta enfermedad. Ejemplos de este tipo de trabajos son Benes et al. (2003) [10] o Lawson y Clark (1999) [62], donde se estudian dos problemas de clustering, general e individual respectivamente. La metodología utilizada para este tipo de estudios se conoce como *análisis de procesos puntuales*. El segundo tipo de datos con los que se puede plantear un estudio de clusters corresponde al análisis de casos agregados según una división territorial. En esta situación se cuenta con el número de casos observado de cierta enfermedad en cada división y el número de observaciones esperadas según la población de dicha región. En esta ocasión el problema que se plantea es la determinación, si la hay, de una región o regiones geográficamente contiguas donde el número de casos observado sea significativamente superior al número de casos esperado. Knorr-held y Rasser (2000) [59], Ferreira et al. (2002)[43] serían ejemplos de estudios cuyo objetivo consiste en determinar las regiones en las que el riesgo es significativamente superior utilizando datos agregados según divisiones administrativas. No se ha de confundir los estudios de detección de agregaciones sobre datos agrupados con los problemas de *suavización geográfica de riesgos* o *disease mapping* en los que el objetivo fundamental consiste exclusivamente en la representación geográfica de las variaciones del riesgo. El último tipo de datos en los que se puede plantear un estudio geográfico de clusters sería en el caso que se disponga un conjunto de observaciones de un proceso sobre una serie de localizaciones fijas, a diferencia de los estudios de procesos puntuales. Este tipo de datos suele ser menos habitual en estudios epidemiológicos aunque en Diggle et al. (1998)

[41] se aplican estas técnicas a la determinación de regiones de riesgo ante infecciones por cierta bacteria.

Otra división propuesta en el estudio de clusters es entre estudios *localizados* y *no localizados*. En los estudios localizados se presupone de antemano la localización exacta de uno o más clusters asociados a distintas fuentes de riesgo. El objetivo principal en este caso consiste en valorar si la cercanía a dicha fuente de riesgo provoca un aumento en la incidencia de casos de cierta enfermedad. Por el contrario, el análisis de clusters no localizado no presupone la localización de ninguna región concreta que pudiera presentar un exceso de riesgo. Nuevamente la metodología a emplear en ambos tipos de estudio es muy diferente y las monografías dedicadas al estudio de clusters no suelen abordar la primera de estas clasificaciones. Por tanto en ningún caso se ha de confundir entre estos dos tipos de estudio ya que sus objetivos y metodología son muy distintas. En Lawson y Clark (1999) [62] se ilustran ambos tipos de estudio.

La última división utilizada se corresponde con el carácter *paramétrico* o *no paramétrico* de la modelización. Así, los modelos paramétricos realizan suposiciones sobre la forma de los clusters o sobre ciertos parámetros que definen la forma de los mismos. Por el contrario los métodos no paramétricos no suponen ninguna forma preconcebida de los clusters por lo que en principio son más flexibles que los métodos paramétricos. A su vez los métodos paramétricos son más potentes a la hora de determinar la existencia de clusters, obviamente suponiendo que la formulación de dicho modelo se adecúe al mecanismo de agregación que ha generado las observaciones. Además, los métodos paramétricos permiten una inferencia más rica, ya que permiten una extrapolación de los resultados con más posibilidades y el aprendizaje sobre los parámetros de estos modelos puede proporcionar conclusiones de gran interés epidemiológico. Por tanto ambos tipos de modelizaciones tienen

sus ventajas e inconvenientes, por lo que la aplicación de uno u otro grupo de técnicas dependerá del problema en concreto.

Uno de los problemas asociados al análisis de clusters en estudios epidemiológicos es que dichos análisis se proponen a la vista de una agrupación de casos que podría parecer anormalmente alta. Por tanto, en muchas ocasiones los análisis de clusters se realizan a posteriori, a la vista de los datos. En ese caso, se corre el peligro de caer en lo que en epidemiología se conoce como sesgo de selección. Una vez se han observado los datos existe una alta probabilidad de que cualquier hipótesis que se proponga contrastar sea aceptada, corriéndose un gran riesgo de que la agrupación que ha motivado dicho contraste haya sido generada por azar. Para evitar este hecho se han de tomar medidas de protección de error a la hora de valorar la existencia de agregaciones. Éste es un aspecto en el que no existe un consenso en absoluto y las conclusiones de los estudios dependen en gran medida del valor del estadístico que haya fijado el investigador como umbral de rechazo, que por otra parte puede variar considerablemente de investigador a investigador. Así, se ha de evitar caer en el procedimiento descrito como el del tirador tejano "Texan sharp shooter" (Bailey, 2001 [8]), en el que se sitúa la diana allá donde haya caído el disparo, es decir, contrastar la existencia de clusters allá donde parece que lo haya a la vista de los datos. En el caso de formularse una hipótesis tras la identificación de una región o regiones de riesgo a la vista de los datos, debería contrastarse dicha hipótesis mediante nuevos datos ya sean de periodos o localizaciones distintas. Este problema es de gran importancia en este tipo de análisis, hasta el punto que ha provocado una gran controversia sobre la utilidad de los estudios de clusters geográficos en enfermedades no infecciosas como se puede comprobar en Wartenberg (2001) [99]. Esta controversia se debe a que muchos de los clusters determinados corresponden a falsos positivos, mientras que la protección de error en una aplicación masiva de estos tests conllevaría la necesidad de encon-

trar evidencias muy grandes de la existencia de agregaciones de casos para concluir su existencia. Este hecho limita la potencia de estas aplicaciones.

Respecto a la interpretación de los tipos de clusters que se han descrito cabría reseñar que los clusters de tipo individual corresponden a agregaciones causadas por factores genéticos, infecciosos o por contaminantes (vertidos) que agruparían los casos en torno a ciertas localizaciones. Esta agregación se puede dar en torno al foco de contagio o a la localización del hogar del progenitor, alrededor del cual tenderían a concentrarse el resto de familiares. Sin embargo las diferencias de riesgo asociadas a un mecanismo cluster general responden a mecanismos distintos. En concreto, este tipo de agrupaciones se debe al hecho de que la distribución geográfica de una enfermedad depende de distintos factores que, en general, varían también geográficamente de forma más o menos suave. Valga como ejemplo las diferencias geográficas en hábitos relacionados con el tabaco que producirán diferencias en la incidencia de enfermedades cardiovasculares sobre cierta región de estudio. En muchas ocasiones será imposible controlar todos los factores que influyen sobre la distribución de la enfermedad, por ello se suele recurrir al uso de modelos de efectos aleatorios para el tratamiento de agrupaciones de tipo general. En general, sobre la distribución de los casos de cualquier enfermedad existirá agregación ocasionada por las diferencias en la distribución geográfica de la población o de los factores de riesgo. La cuestión es si dicha agregación es epidemiológicamente significativa o si los datos permiten o no detectar dicha variación, tal y como se expone en Wakefield et al. (2000) [98]. Por tanto no sólo el tratamiento que se le ha de dar a los clusters de tipo general e individual es distinto sino que la interpretación y motivación de ambos tipos de estudios responde a situaciones completamente diferentes, lo cual ha de tenerse en cuenta a la hora de plantear un estudio.

Además del estudio del patrón geográfico en la aparición de casos, también resulta posible el estudio de agregaciones temporales en la incidencia de cierta enfermedad (Birch et al., 2000 [18]). Incluso es posible acometer el estudio de agregaciones espacio-temporales en la incidencia de una enfermedad (Chardot et al., 1999 [28]). En este caso, no se estudia si existen agregaciones geográficas de los casos ni si existen agregaciones temporales en la incidencia de éstos, sino si en intervalos pequeños de tiempo y de espacio se observan un número de casos anormalmente alto. En esa ocasión, se podría pensar en un factor de riesgo que actúa localmente en una localización y un instante concreto.

1.2. Brotes de Neumonía por *Legionella* en Alcoi

En esta sección, se describen los brotes de legionelosis en la ciudad de Alcoi comentados anteriormente y que van a servir de ilustración para el resto del trabajo. Concretamente, en primer lugar se introducen ciertos aspectos de la *Legionella Pneumophila*, la bacteria causante de los brotes descritos en la segunda parte de esta sección. En esta segunda parte, además se detallan los estudios de los brotes realizados previamente al planteamiento del presente trabajo. Dichos estudios constituyen el estado de conocimiento previo del problema y son las bases de las líneas de desarrollo metodológico que vamos a proponer.

1.2.1. *Legionella Pneumophila*

En 1976 tuvo lugar en Philadelphia (EEUU) una convención de la legión americana, una asociación de ex-militares norteamericanos, para conmemorar el segundo centenario de la firma de la declaración de independencia de Gran Bretaña. Más de 180 delegados, alojados en el mismo hotel, contrajeron una enfermedad aguda de la que 29 de ellos murieron. Finalmente el brote concluyó con un total de 34 defunciones, algunas de ellas de simples viandantes que paseaban por la calle. Inicialmente la causa de su enfermedad resultó desconocida, aunque se sospechaba que podía deberse a una intoxicación alimentaria. Actualmente se sabe que contrajeron legionellosis, una neumonía o infección pulmonar cuyo causante es la bacteria *Legionella Pneumophila*.

La colonización y crecimiento de la bacteria *Legionella* puede darse en cualquier medio que contenga agua, siempre y cuando ésta tenga los nutrientes apropiados y la temperatura adecuada; entre 20 y 45 grados centígrados, aunque la temperatura óptima es entre 35 y 37 grados. Por tanto, son caldos de cultivo adecuados para la bacteria las duchas, baños, balnearios, fuentes ornamentales, aspersores de riego, lavaderos de coches, accesorios de dentistas, humidificadores de ambiente, torres de refrigeración, condensadores evaporativos o cualquier equipo de transferencia de masa de agua en corriente de aire con producción de aerosoles. Estos equipos resultan especialmente peligrosos en verano y otoño cuando las condiciones climáticas para el desarrollo de *Legionella* resultan más propicias. Esta bacteria habita en fuentes de agua naturales como arroyos, ríos, lagos,... por tanto no resulta sorprendente que colonice las instalaciones de riesgo que acabamos de mencionar. Si bien la utilización de biocidas específicos reduce la reproducción de estas bacterias todavía no se conoce ningún método efectivo para la erradicación de las mismas.

La trasmisión de la bacteria se produce por inhalación de la misma en aerosoles (gotas de agua microscópicas) contenidos en el aire. La ingestión de la misma resulta inofensiva. Además, puesto que la transmisión mediante contagio entre seres vivos no resulta posible, el único mecanismo posible de adquisición de la bacteria es mediante inhalación directa del medioambiente. Por tanto, cualquier sistema que trabaje en contacto con agua y que genere aerosoles se convierte en un emisor potencial de *Legionella*, por lo que supone un peligro para la salud pública si no se trata y mantiene adecuadamente. Es por ello que la existencia de las instalaciones de riesgo que se han mencionado previamente supone una amenaza para la salud de las personas de dicha población, si no siguen el mantenimiento oportuno.

La peligrosidad de estas instalaciones de riesgo puede variar según ciertos factores. En concreto, depende de la cantidad de aerosol emitido a la atmósfera, de la distancia a la que es emitido el vapor respecto a la población, de la accesibilidad al repositorio de agua para su desinfección o la temperatura a la que se mantiene el agua entre otros. Algunos de estos factores influirán sobre el radio de riesgo en el que la instalación resulta peligrosa para las personas que residen o transitan a su alrededor. Existen diferentes estudios sobre este hecho, como el de Brown et al. (1999) [23] que realiza un análisis de regresión logística apareado en el que se comparan personas que han desarrollado la enfermedad y personas que no (en determinados diseños de estudio epidemiológico se les denomina *controles*). En su estudio Brown et al. determinan que el riesgo de presentar la enfermedad depende de la distancia más próxima a la que se ha estado en contacto con la fuente de contagio, en concreto dicho riesgo disminuye en un 20 % por cada 0.1 millas de aumento en dicha distancia (aproximadamente 147 metros), estableciéndo que para distancias superiores a 0.25 millas dicho riesgo era despreciable. Por otra parte, en Bhopal et al. (1991) [17] se determina una asociación inversa entre la distancia a torres de refrigeración y el ries-

go de aparición de casos a partir de distintos estudios de casos aislados de legionelosis no asociados a brotes. Concretamente, la población residente a menos de 0.5 millas de una torre presentaba una probabilidad de infección 3 veces superior a la de la población residente a más de 1 milla de éstas. En cualquier caso, la supervivencia de la bacteria en el aire depende de distintos factores como, por ejemplo, la humedad del aire. Además, la distancia a la que se desplaza depende de otros factores como la velocidad del aire, por lo que el radio de influencia de las torres de refrigeración variará para cada problema y escenario concreto.

La infección por *Legionella* presenta dos formas clínicas perfectamente diferenciadas, la *Fiebre de Pontiac*, síndrome febril agudo de pronóstico leve y la *Enfermedad del legionario* o *Neumonía por Legionella* que cursa con fiebre alta, neumonía y cefalea. En este último caso la letalidad, proporción de muertes entre el total de afectados, de la enfermedad está alrededor del 5 % aunque si no se administra el tratamiento adecuado puede ascender hasta el 15 o el 20 %. El periodo de incubación en los enfermos puede variar entre 2 y 10 días y el riesgo de desarrollar la enfermedad dependerá de la exposición a la bacteria y las circunstancias personales de cada persona, como edad, sexo o patologías previas.

Existe un gran número de textos donde se puede ampliar información sobre la ecología de la bacteria, el cuadro clínico que produce o el tratamiento de desinfección y limpieza de las instalaciones de riesgo para que estas no sean peligrosas para la población. Esta última cuestión está incluso recogida en la legislación vigente de la Comunidad Valenciana (decretos 173/2000 de 5 de diciembre y 201/2002 de 10 de diciembre) y a nivel nacional (real decreto 865/2003 de 4 de julio). Un buen texto de referencia para todos estos aspectos es la guía editada por el Ministerio de Sanidad para el control y prevención de la legionelosis [67].

1.2.2. Descripción del brote y análisis preliminar

Alcoi es una ciudad industrial situada en el norte de la provincia de Alicante con 60.532 habitantes en el año 2004. La ciudad situada a 545 metros sobre el nivel del mar, ocupa el centro de una depresión, conocida como "la foia d'Alcoi". La población se extiende siguiendo un eje longitudinal noreste-sudoeste, ocupando una superficie de 130,6 km^2. Este territorio se encuentra atravesado por diversos cursos fluviales encastrados en profundos barrancos.

La principal industria ubicada en esta ciudad es la textil de la que vive gran parte de su población. Esta industria tiene una gran tradición en Alcoi y como consecuencia existen un gran número de empresas situadas dentro del núcleo urbano con todos los inconvenientes y peligros para la salud que ello supone.

La situación de la ciudad, en el interior de una hoya montañosa caliza e irregular, favorece un microclima con formación de brumas y tendencia a la acumulación de las emisiones urbanas en ausencia de viento suficiente que las disperse. La ubicación de la mayoría de industrias en los cauces, condiciona que las emisiones se produzcan a la altura de viviendas, calles y paseos. Por tanto las condicones orográficas-meteorológicas de Alcoi unidas a la presencia de una gran industria en el núcleo urbano, con sus instalaciones de riesgo, hacen de esta ciudad un caldo de cultivo propicio para la aparición de brotes de Legionelosis.

Entre el 16 de Septiembre y el 8 de Octubre de 2000, se produjeron en la ciudad de Alcoi 57 casos de neumonía por *Legionella*. El brote fue muy virulento ya que 40 de los 57 casos del brote iniciaron síntomas en 8 días. Se sospechaba, prácticamente desde el principio, de un origen me-

dioambiental del brote, es decir no asociado al suministro de agua de la ciudad sino asociado a alguna instalación de riesgo que estaba emitiendo *Legionella* al medioambiente. Concretamente se sospechaba de las instalaciones de refrigeración generadoras de aerosoles de las empresas y talleres emplazados en el casco urbano. Se habían tomado medidas de desinfección, limpieza e incluso en algún caso, el cierre de dichas instalaciones. Mas como se volvía a repetir una nueva onda epidémica, se sospechaba que podrían haber instalaciones de cuya existencia no tuvieran conocimiento las autoridades municipales y sanitarias, por lo cual, no habrían sido sometidas a los controles mencionados para garantizar su seguridad para la salud pública.

Ante esta situación, se hacía imprescindible la búsqueda de nuevas instalaciones potencialmente de riesgo y cuya presencia no hubiese sido comunicada a las autoridades tras el reclamo realizado a tal efecto por el ayuntamiento. Para ello, se pretendía delimitar las zonas de mayor riesgo con objeto de concentrar en ellas los esfuerzos de búsqueda de instalaciones no declaradas y que podrían haber generado el brote epidémico que se estaba analizando. Con este propósito se pusieron en marcha distintos estudios paralelos que resumimos a continuación.

Estudio meteorológico

En primer lugar se llevó a cabo un estudio del contexto meteorológico de los brotes en el que se relacionó la aparición de casos con distintas variables meteorológicas, en concreto se utilizaron mediciones diarias de la humedad, temperatura, dirección y velocidad del viento. Los datos fueron proporcionados por el Instituto Nacional de Meteorología provenientes de su estación meteorológica situada entre los términos municipales de Alcoi y Cocentaina. Para cada una de estas variables se estudió la relación entre su

serie temporal diaria desestacionalizada y la aparición de nuevos casos de la enfermedad. El periodo de estudio se circunscribe al espacio de tiempo que abarca los sucesivos brotes que se habían producido en la ciudad de forma casi sucesiva.

A partir de este estudio se concluye la existencia de una relación inversa estadísticamente significativa entre la aparición de casos cada día del periodo de estudio y el flujo de viento que sopló sobre la ciudad durante los 5 días previos. De esta forma se demuestra que el viento tiene un efecto protector sobre la aparición de casos descontando un decalaje de 5 días que coincide con el periodo de incubación promedio de la enfermedad. Este hallazgo supone una nueva evidencia de que el origen del brote era medioambiental, ya que de otra forma no habría relación entre las condiciones meteorológicas y la aparición de casos de legionelosis.

Tras el estudio meteorológico de series temporales se llevaron a cabo dos estudios más en los que se analizaba la distribución de la localización de los casos en el interior del núcleo urbano. Para la realización de ambos estudios se consideró un conjunto de 65 controles[1]. La elección de los controles se realizó de forma apareada a alguno de los casos y como criterio de apareamiento se utilizó que tanto caso y control fueran de similar edad, igual sexo y que este último hubiera tenido una consulta hospitalaria, que no tuviera como motivo legionelosis, una semana antes o después de que el caso ingresara en el mismo.

[1]Persona de similares características que los casos con la única diferencia de que los primeros no presentan la enfermedad de estudio. La elección de controles en un estudio nos permite la comparación de las características de las personas que presentan la enfermedad con las que no la presentan y de esta forma se puede indagar qué factores influyen sobre la aparición de la enfermedad en los casos. Los estudios de casos y controles es uno de los tipos de estudio con mayor uso y difusión en epidemiología.

Estudio de la movilidad urbana

El primero de los estudios geográficos planteados considera los casos agregados según las secciones censales de la ciudad. Para cada una de estas secciones censales se conoce el número de residentes que han contraido la enfermedad, así como el número esperado de casos que deberían haberse dado teniendo en cuenta la pirámide poblacional de la gente que reside en cada una de estas secciones. Para el estudio de la relación entre el número de casos observado y esperado en cada una de las secciones se consideró necesario hacer uso de alguna técnica de suavización geográfica de riesgos en áreas pequeñas. El tamaño de las divisiones administrativas elegidas, alrededor de 1000 personas por sección, aconsejaba el uso de este tipo de técnicas para obtener estimaciones estables del riesgo en cada división administrativa. Dicha estimación del riesgo en cada unidad geográfica se realizó utilizando la propuesta de Besag et al. (1991) [15] que contempla la variación del riesgo de sección a sección como combinación de 2 efectos aleatorios. El primero de los efectos aleatorios varía entre regiones de forma independiente y permite la existencia de unidades geográficas con un comportamiento diferente del resto. En el segundo de los efectos aleatorios se incorpora la dependencia de las observaciones entre distintas secciones censales, esta dependencia suele relacionarse con el emplazamiento geográfico de cada una de las regiones de estudio. Concretamente se contempla la existencia de correlación positiva entre las observaciones de secciones vecinas debido a que dichas localizaciones compartirán ciertos factores de riesgo, simplemente por hallarse emplazadas próximas geográficamente.

Sin embargo en el estudio que se llevó a cabo se consideró que la dependencia entre secciones debería definirse según otro criterio distinto a la localización geográfica, la distancia entre regiones se definió en base al tránsito de gente entre ellas. Para determinar dicha distancia se utilizó la

información recogida en la encuesta epidemiológica que se había realizado a los controles, en concreto se hizo un análisis de correspondencias de la matriz de movimientos de los encuestados entre las distintas secciones censales de la ciudad. De esta forma se pudo definir una distancia entre las distintas secciones censales de la ciudad según el trasiego de gente que había entre ellas. La definición de esta estructura de proximidad se debe al hecho que las personas no tienen un comportamiento estático sino que se mueven regularmente por el núcleo urbano de la ciudad y resulta posible que se contraiga la enfermedad en uno de estos desplazamientos. Por tanto si los residentes en cierta sección censal suelen transitar por otra sección es previsible que los habitantes de ambas tengan un riesgo similar al compartir las zonas por las que transitan. Esta dependencia en el riesgo es la que se ha tratado de incorporar en el modelo mediante la estructura de vecindad propuesta. El cálculo de las vecindades entre regiones según la movilidad urbana a partir de los controles del estudio se debe a que se quería que dichas vecindades reflejaran el tránsito de la población en general por la ciudad y no el tránsito de las personas enfermas. Como el patrón de desplazamientos de casos y controles podría diferir se ha utilizado los datos de estos últimos ya que se considera que el patrón de desplazamiento de éstos reflejará de forma más fiel el comportamiento de la población en general.

En la figura 1.1 se representan los resultados obtenidos utilizando un modelo de suavización en el que la estructura de vecindad entre regiones viene definida por contigüidad geográfica entre regiones, es decir la propuesta habitual en estudios de suavización de riesgos. Las zonas márcadas en rojo corresponden a regiones con un riesgo alto de contraer la enfermedad, ya que en éstas se ha dado un número de casos observado mayor que el que se esperaba. En esta representación se puede observar una amplia región en el centro y noreste de la ciudad donde el riesgo es más alto que en el resto de la ciudad. No obstante la zona delimitada mediante este análisis abarca

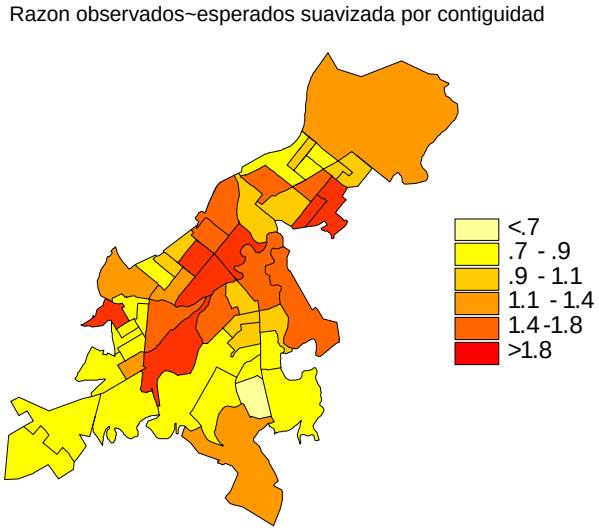

Figura 1.1: Suavización según contigüidad

una gran proporción del núcleo urbano. Por tanto estos resultados no son de excesiva utilidad a la hora de delimitar una región en la que incidir a la hora de buscar nuevas instalaciones de riesgo. Por otra parte, en la figura 1.2 se representan los resultados obtenidos al hacer uso de un modelo en el que la estructura de vecindad entre regiones viene definida por los desplazamientos de la gente que reside en ellas. En este caso se puede observar que la zona que se ha delimitado es bastante más reducida que la proporcionada por el modelo de contigüidad, por tanto la utilidad práctica de este modelo resulta mayor que la del modelo previo. Además, también se puede observar que a tenor de los resultados obtenidos parecen existir 3 zonas concretas con un riesgo mayor que el resto de secciones censales. Este hecho sugiere un posible origen multifocal del brote epidémico.

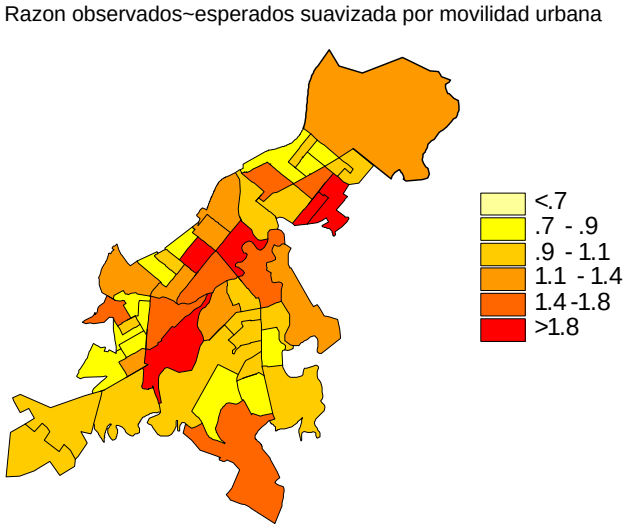

Figura 1.2: Suavización según movilidad urbana

Estudio del patrón puntual

El segundo de los estudios sobre el patrón geográfico de distribución de los casos, a diferencia del primero, utiliza la información exacta de la localización del domicilio de casos y controles en lugar de agregarse por secciones censales. En este caso se han utilizado herramientas de estudio de patrones puntuales para la comparación del patrón de distribución geográfica de casos y controles. Sin embargo, en este nuevo estudio no se tiene en cuenta la movilidad urbana de la población a la hora de establecer las regiones de riesgo.

Dada una configuración de puntos sobre una región geográfica se dice que dicho patrón sigue un proceso de aleatoriedad espacial completa si la configuración de puntos observada sigue una distribución uniforme sobre

la región de estudio. Cualquier desviación respecto a esta situación puede expresarse como una mayor regularidad entre los puntos, en cuyo caso diremos que el patrón presenta inhibición entre sus eventos, o presentando un mayor agrupamiento entre puntos, en este caso diremos que el patrón es de tipo agregado.

Mediante el estudio de patrones puntuales realizado se ha comparado el patrón observado en los casos con el de los controles del estudio. En concreto se ha llevado un a cabo un test de Monte Carlo mediante etiquetado aleatorio Diggle y Rowlingson (1994) [38] para comparar ambos patrones. Como resultado de dicho test se descarta que casos y controles se comporten de la misma forma en cuanto a su carácter de agregación o inhibición. Además, se demuestra que los casos siguen un patrón de mayor agregación que los controles. En concreto se concluye que dado un caso se tiene que en un radio de 150 metros alrededor de éste la probabilidad de observar un caso es mayor que la probabilidad de que se observe un control. Este hecho da una idea del carácter agregado del patrón de casos.

La metodología relativa a este estudio se introducirá y discutirá con mayor detalle en el capítulo 2 del presente trabajo. Se puede encontrar información más detallada sobre este trabajo en Abellán et al. (2002) [1], Martínez-Beneito et al. (2005)[66].

1.3. Objetivos del trabajo

Tras los estudios preliminares realizados de los distintos brotes epidémicos que habían tenido lugar en la ciudad de Alcoi quedan ciertas cuestiones epidemiológicas a las que se debe dar respuesta y que hasta el momento no resultan del todo claras. Los estudios realizados proporcionan ciertas

conclusiones de indudable valor y que se resumirían en las siguientes:

- A tenor del estudio meteorológico parece confirmarse el probable origen ambiental del brote, por tanto se descarta el agua de consumo como origen del brote epidémico. Este resultado permite orientar los esfuerzos para atajar la sucesión de brotes hacia la búsqueda de instalaciones de riesgo emisoras de aerosoles.

- A partir del estudio de patrones puntuales queda patente que la distribución de casos y controles es cualitativamente distinta en el sentido que la distribución de los casos muestra una agregación mayor que la de los controles. Por tanto el mecanismo de transmisión de la enfermedad actúa de forma local produciendo agrupaciones en la distribución de la incidencia de la enfermedad. Este resultado motiva el planteamiento de un estudio de localización de clusters en la distribución de la incidencia de la enfermedad dentro del núcleo urbano para determinar el origen de estas agregaciones.

- El estudio de la movilidad urbana sugiere que la distribución del riesgo no es homogénea, es más se aprecian distintas agregaciones en la distribución de la incidencia. A partir de este estudio parece sugerirse al menos 3 agrupaciones de casos en torno a localizaciones concretas. Este hecho apunta hacia un origen multifocal del brote epidémico.

La limitación fundamental de los estudios anteriores es que no responden a ciertas preguntas de gran utilidad para la gestión y prevención de futuros brotes, éstas serían ¿cuantos focos epidémicos han actuado en cada uno de los brotes?, ¿cual es el emplazamiento más probable de estos focos? El objetivo del presente trabajo será la formalización de una propuesta que permita dar una respuesta más concreta a estas y otras cuestiones concernientes al

estudio de estos brotes epidémicos y que se puede aplicar de forma general a otros estudios de similares características.

Dados los tipos de estudio de clusters que se han detallado hasta el momento y las características del brote que se han descrito, la metodología que se ha de proponer habría de cumplir ciertos requerimientos que detallamos seguidamente. En primer lugar la propuesta a realizar habría de consistir en un análisis de clusters de tipo individual, ya que el objetivo principal del estudio es la determinación de la localización y número de focos de riesgo intervinientes en el brote. Sin embargo, tal y como se ha señalado en la descripción de los estudios de agrupaciones, la hipótesis de que no existe ningún otro factor en el problema aparte de las instalaciones de riesgo que produzca agregación en la incidencia resulta poco sostenible. En concreto, la distribución de la población según edad y factores de riesgo en el interior del núcleo urbano de la ciudad condiciona la distribución de los casos. Es por ello que también se habría de considerar en la modelización alguna forma general de clustering que reflejara todos estos factores que acabamos de señalar. En caso contrario se podrían confundir las agrupaciones asociadas al mecanismo de agregación general con las agrupaciones debidas a la presencia de instalaciones de riesgo, pudiendo derivarse conclusiones erróneas del estudio emprendido.

Por otra parte la metodología a desarrollar ha de ser capaz de explotar la información con un nivel de desagregación lo más fino posible. En caso contrario se perdería información en el proceso de agregación de los casos en las distintas regiones administrativas, resultando en dicho caso un análisis de menor utilidad al ignorarse parcialmente la información disponible. De esta forma, al disponer de la dirección exacta de residencia de cada uno de los casos observados debemos plantear un estudio de clusters con datos puntuales.

Respecto al carácter localizado o no localizado del estudio, en principio ambos planteamientos tendrían sentido. Sin embargo la aplicación de un estudio localizado sobre el brote que nos ocupa presenta dos problemas fundamentales, el primero es que dicho planteamiento requiere la disposición de un censo exhaustivo de instalaciones de riesgo. En el momento de plantear el presente estudio dicho censo se encontraba en fase de elaboración, más concretamente la incorporación de instalaciones a dicho censo era un goteo constante. Por tanto la exhaustividad del censo que se estaba recogiendo era dudosa y la omisión de alguna instalación en el inventario que se estaba elaborando podría tener consecuencias muy serias sobre los resultados del estudio. En segundo lugar el censo final de instalaciones de la ciudad de Alcoi superaba las 200 unidades y en el momento en que se planteaba el presente estudio ya se intuía las dimensiones que iba a tomar dicha relación. En el caso de tener evidencias de que el brote de estudio era de tipo monofocal cabría plantearse un estudio de tipo localizado, valorando una a una la relación entre la distancia a cada instalación y el aumento en la incidencia de casos. Sin embargo el estudio de movilidad urbana parecía ofrecer evidencias en contra de la monofocalidad del brote, es más se desconocía el número de instalaciones que podrían estar interviniendo en éste. Es por ello que el número de combinaciones de instalaciones de riesgo que habría que valorar en un estudio localizado hacía intratable dicha aproximación, más si se tiene en cuenta el problema que conllevaría el número de comparaciones múltiples que se habrían de llevar a cabo. Estos dos problemas asociados a los estudios localizados desaconsejaban emprender un análisis de dicho tipo.

También se ha comentado que tanto los modelos paramétricos como los no paramétricos presentan una serie de ventajas e inconvenientes, es por ello que en principio ninguna de estas dos opciones resulta más apropiada que la otra en la propuesta que se pretende realizar. Sin embargo resultaría interesante un planteamiento que aunara las ventajas de ambas modeliza-

ciones y que de esta forma sume la potencia de los modelos paramétricos y la flexibilidad de los modelos no paramétricos en una única propuesta.

Otro detalle que resulta importante a la hora de establecer una modelización del problema es que la transmisión de la enfermedad entre personas no es factible. Por tanto, estadísticamente, los casos se habrían de considerar como condicionalmente independientes dadas la información de su lugar de residencia y cualquier otra información que se considere oportuna. Así un modelo que considere interacción entre casos directamente (la probabilidad de observar un caso aumenta la probabilidad de observar algún otro caso cerca como consecuencia de el primero) resultaría desde el punto de vista epidemiológico poco sostenible.

Así, de esta forma, se han determinado los requerimientos que habría de cumplir toda propuesta que pretendiera dar una modelización realista y realmente útil de los brotes de legionelosis de Alcoi. Además cualquier propuesta del tipo que hemos determinado también sería apropiada para cualquier brote asociado a una o varias fuentes de riesgo de emplazamiento y número desconocido en el que no exista contagio directo entre casos.

1.4. Notación

A continuación vamos a describir ciertos criterios de notación que serán utilizados repetidamente en el resto del texto.

- Notaremos por $\mathcal{N}_K(\mu, \Sigma)$ a la distribución normal K-dimensional de media μ y matriz de varianza-covarianza Σ. En el caso que K valga 1, el caso univariante, no incluiremos el subíndice en la expresión anterior y entenderemos que su segunda componente se refiere a la varianza de

la distribución normal.

- $\mathrm{Mn}(p_1, p_2, ..., p_n)$ denotará la distribución multinomial, de probabilidades $\{p_1, p_2..., p_n\}$ para cada uno de los posibles valores $\{1, 2, ..., n\}$.

- Denotaremos por $\mathcal{B}(\alpha, \beta)$ a la distribución beta de media $\alpha/(\alpha + \beta)$ y varianza:
$$\frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}.$$

- Denotaremos la generalización multivariante de la distribución beta, la distribución Dirichlet, mediante $\mathrm{Dir}(\delta_1, ..., \delta_n)$ donde la media de la i-ésima componente vale
$$\frac{\delta_i}{\sum_{j=1}^{n} \delta_j}$$
y su varianza
$$\frac{\delta_i \sum_{j \neq i} \delta_j}{(\sum_{j=1}^{n} \delta_j)^2(\sum_{j=1}^{n} \delta_j + 1)}.$$

- La distribución Gamma se representará como $\Gamma(\alpha, \beta)$ en la que la media de la distribución se corresponde con α/β y la varianza con α/β^2.

- La generalización multivariante de la distribución Gamma, la distribución Wishart, se denotará como $\mathcal{W}(\nu, \beta)$ donde ν corresponde a un escalar que controla la precisión de la distribución (los grados de libertad) y β es una matriz simétrica, definida positiva. El valor esperado de la distribución Wishart es $\nu \cdot \beta$.

- Denotaremos por $1(\mathcal{A})$ la función característica que valdrá 1 si se cumple la condición $\mathcal{A}$ y 0 en otro caso. En algunas ocasiones también emplearemos la notación $1_{[a,b]}(c)$ para la función característica que vale 1 si c pertenece al intervalo $[a, b]$ y 0 en otro caso.

- Para cualquier subconjunto A de la recta o el plano real, $|A|$ denotará la medida de dicho conjunto. Concretamente, en el resto de este trabajo utilizaremos la medida euclídea, aunque los resultados obtenidos serían igualmente válidos en el caso de utilizar cualquier otra medida.

Capítulo 2

Procesos puntuales aplicados al estudio de brotes epidémicos

La teoría de procesos puntuales proporciona una herramienta para la descripción de la variación geográfica del riesgo en estudios epidemiológicos. Estas técnicas son apropiadas para el análisis de datos individuales frente a la corriente metodológica más extendida que estudia los datos agrupados según divisiones administrativas. Sin embargo la implantación de Sistemas de Información Geográfica (GIS) en el ámbito sanitario y la mejora de la calidad de la información en las bases de datos posibilita la realización de estudios geográficos a nivel individual. Es más, dicha mejora redundará en una demanda de metodología de análisis geográfico de datos individuales a la que el estudio de procesos puntuales parece ser capaz de dar respuesta.

2.1. Procesos puntuales

Existen distintas formulaciones de la teoría de procesos puntuales más o menos rigurosas. Durante el presente trabajo no se va a abordar dicha teoría

desde una formulación excesivamente matemática sino que se van a introducir únicamente los conceptos necesarios para un correcto entendimiento de los posteriores capítulos. Para una formulación más rigurosa existen distintos textos que pueden considerarse referencia en este campo donde se tratan los detalles teóricos con mucha más profundidad, como por ejemplo Stoyan et al. (1996) [95], Möller y Waagepetersen (2004) [70] y Cressie (1993) [31].

Vamos a utilizar la siguiente definición de proceso puntual, ya que resulta muy intuitiva y no recurre a formalismos innecesarios para el desarrollo del resto del trabajo. "Un *proceso puntual* es un mecanismo estocástico que genera un conjunto contable de eventos sobre el plano" (Diggle, 2003 [39]). Por tanto, podemos entender un proceso puntual como todo mecanismo capaz de generar un conjunto aleatorio y finito de casos sobre cierta región de estudio. Notar que según la definición propuesta, en ningún momento se supone conocido el número de casos generado por el proceso, así, en un proceso puntual el número de casos generado por el proceso será también una componente estocástica de los datos. A cualquier realización de un proceso puntual se le dice *patrón puntual*. Para evitar confusión entre puntos o localizaciones de la región de estudio y los puntos que pertenecen al proceso puntual llamaremos a estos últimos *eventos* cada vez que nos refiramos a ellos.

Al igual que las variables aleatorias se resumen y describen mediante sus momentos, los procesos puntuales pueden ser resumidos mediante lo que llamamos funciones de intensidad. En concreto, si ds denota una región infinitesimal alrededor del punto s, se define la *intensidad de primer orden* de un proceso en la localización s como:

$$\lambda(s) = \lim_{|ds|\to 0} \left[\frac{E(N(ds))}{|ds|} \right] ,$$

donde $N(ds)$ es el número de eventos observados en la región ds. Por tanto,

la intensidad de primer orden de un proceso representa el número esperado de puntos en cada localización de estudio por unidad de área. En concreto, la integral de la función de intensidad sobre la región de estudio coincidirá con el número esperado de eventos en cualquier patrón puntual generado por el proceso. A $\lambda(s)$ se le conoce también como *función de intensidad* del proceso.

Para que un proceso puntual esté bien definido la integral de su función de intensidad sobre la región de estudio necesariamente ha de ser finita. En caso contrario los patrones generados no serán necesariamente contables. Por tanto suele ser habitual definir los procesos puntuales sobre una región de estudio finita. En dicho caso si la función de intensidad es finita se asegura que la integral de la función de intensidad sobre la región de estudio también lo será. Sin embargo, la elección de un dominio finito para el proceso puntual comporta otras complicaciones, concretamente habremos de tener en cuenta que pueden existir eventos, de los que no tenemos noticia, en el exterior de la región de estudio. Por tanto, habremos de tener en cuenta dicha hipótesis e incorporarla al análisis si no se quieren obtener resultados sesgados debidos a la elección de la región de estudio. La incorporación de este efecto a la estimación de un proceso puntual se le conoce como *técnicas de corrección de arista*.

La *función de intensidad de segundo orden* de un proceso puntual se define como:

$$\lambda_2(s,u) = \lim_{|ds|,|du|\to 0} \left[\frac{E(N(ds)N(du))}{|ds||du|} \right] .$$

La expresión $\lambda_2(s,u) - \lambda(s)\lambda(u)$ se puede entender como la covarianza del proceso entre las localizaciones s y u. De hecho, valores de $\lambda_2(s,u)$ superiores a $\lambda(s)\lambda(u)$ indicarán una propensión a darse eventos en s cuando se den en u y viceversa.

Obviamente se pueden definir intensidades de orden superior para cualquier proceso puntual, sin embargo dichos estadísticos suelen ser raramente utilizados y su interpretación resulta menos clara cuanto mayor es el orden de la función de intensidad. También suele ser habitual el establecimiento de hipótesis sobre las funciones de intensidad para facilitar el estudio de procesos puntuales. Así, resulta común asumir que la función de intensidad de cualquier orden del proceso es invariante ante traslaciones, en dicho caso se dice que el proceso presenta *estacionariedad*. En consecuencia, si S es la región en la que se define un proceso puntual, diremos que éste es estacionario si

$$\lambda(s) = \lambda(u),\ \lambda_2(s,u) = \lambda_2(s-u)\ \ \forall\ s,\ u \in S\ ,$$

es decir, el número de casos esperado en cada localización del proceso es constante y la dependencia de segundo orden entre dos puntos depende únicamente de su posición relativa. Por otro lado, diremos que un proceso será *isotrópico* si sus funciones de intensidad son invariantes ante rotaciones. En particular para un proceso estacionario e isotrópico la función de intensidad de segundo orden entre dos puntos dependerá unicamente de la distancia que les separa.

Los procesos puntales más simples son los *homogéneos de Poisson*, de los que se puede afirmar que suponen la referencia a partir de la que se construye la teoría de procesos puntuales. Se definen como aquellos mecanismos aleatorios que cumplen las siguientes dos propiedades:

- Existe un valor $\lambda > 0$ tal que para cualquier región A del plano el número de eventos que contiene sigue una distribución de Poisson de parámetro $\lambda|A|$,
- Para toda región A del plano, con n eventos, éstos siguen una distribución uniforme sobre la región A.

El proceso puntual generado a partir de los postulados anteriores se dice proceso de Poisson de intensidad λ, donde el valor de dicho parámetro coincide con el número esperado de eventos por unidad de superficie. En un patrón generado a partir de un proceso homogéneo de Poisson los eventos serán independientes entre sí y la probabilidad de que uno de ellos aparezca en una localización concreta será la misma para cualquier localización del plano. Los patrones generados de esta forma se dice que se rigen según *aleatoriedad espacial completa*, de ahora en adelante AEC. En problemas epidemiológicos la hipótesis de AEC no es demasiado relevante, ya que la población a riesgo varía según la localización geográfica y obviamente este factor influye sobre la distribución de casos sobre la región de estudio. Es por ello que el patrón de AEC resulta muy improbable en este tipo de estudios y es más útil comparar la distribución de los casos observados frente a la población general.

El objetivo de todo estudio de procesos puntuales es, a partir de uno o varios patrones, intentar adquirir información sobre el proceso que los ha generado. Una de las mayores complicaciones a la hora de estudiar un proceso radica en que normalmente se dispone de un único patrón para llevar a cabo el estudio, en contra de lo que suele ser habitual en otros campos de la estadística. Sin embargo, por suerte, la información que se puede extraer de un único patrón puntual puede ser muy rica, siempre y cuando la metodología que se utilice sea lo suficientemente potente. Éste es el objetivo de la inferencia en procesos puntuales, la explotación eficiente de la información disponible a la vista de un patrón puntual y la generación de conclusiones sobre el proceso que lo rige.

La primera hipótesis que se suele valorar a la hora de estudiar un patrón puntual es si éste sigue AEC, el patrón puntual estándar, antes de emprender análisis más complejos o hacer uso de modelización para describir el

mecanismo que ha generado el patrón. Si se desecha la hipótesis de aleatoriedad espacial completa diremos que el patrón es de tipo *cluster o agregado* cuando la distancia entre eventos sea menor en general que para la AEC. En caso contrario, diremos que los datos siguen un patrón de *regularidad o inhibición*. En la siguiente sección se detallan una serie de tests para contrastar la hipótesis de aleatoriedad o, de forma más general, la comparación de patrones.

2.2. Valoración de procesos puntuales

La presencia de agregación en un proceso puntual puede darse bien por la existencia de extra-variabilidad en la distribución de los casos (clustering general) o bien por la presencia de ciertas localizaciones específicas alrededor de las cuales se agruparía la incidencia de casos (clustering individual). Los planteamientos de los contrastes sobre la hipótesis de AEC se basan en 2 grandes líneas argumentales. La primera de ellas consiste en la valoración de la hipótesis de que el número de eventos en cualquier región del plano sigue una distribución de Poisson de media constante. Cualquier evidencia en contra de este hecho se interpretará como prueba de la no aleatoriedad espacial completa del patrón. El segundo enfoque empleado se basa en el estudio de la posición relativa de la localización de cada evento respecto al resto. Más concretamente, se estudiará la distribución de la distancia entre eventos, o entre puntos y eventos, y se valorará si dicha distribución es compatible o no con la hipótesis de AEC.

En el caso que no se quiera comparar el patrón disponible con la hipótesis de aleatoriedad completa, sino que se desee comparar dicho patrón con otra hipótesis de interés, como por ejemplo la distribución de la población a riesgo en la región de estudio, también se podrá hacer uso de las 2 ideas anteriores

para contrastar la igualdad de los patrones que se pretende comparar.

2.2.1. Contrastes de la hipótesis de Poisson

Tal y como se determina en el primero de los postulados de la definición de AEC, bajo esta hipótesis, la distribución del número de eventos del patrón en cualquier región del plano habrá de seguir una distribución de Poisson proporcional al tamaño de la región. Esta suposición conlleva ciertas premisas que se habrán de cumplir para que se admita como válida la hipótesis de aleatoriedad. Una de estas premisas viene dada por el hecho de que la distribución de Poisson implica que la media y la varianza de sus observaciones han de coincidir necesariamente. Esta situación se puede valorar estableciendo una partición de la región de estudio y analizando el número de eventos que recaen en cada una de las componentes de esta partición. En caso de obtener evidencias de que la media y la varianza de esta cantidad no coinciden habremos de rechazar la hipótesis de AEC.

Como ilustración de este hecho, en la figura 2.1 se puede observar la representación de un patrón que sigue AEC, otro que exhibe inhibición y un patrón agregado, respectivamente, junto a una partición de la región de estudio. Todos los patrones constan de 200 puntos y 64 celdas en la partición. En la parte inferior de la figura se ha representado un histograma con el número de eventos para cada región de la partición, junto a una curva que representa la distribución de Poisson de parámetro 200/64, el número medio de puntos por cada una de las celdas. En la figura se puede apreciar cómo el patrón de aleatoriedad espacial completa se adapta perfectamente a la hipótesis de que el número de eventos por celdas sigue una distribución de Poisson. Sin embargo, los otros dos patrones, aunque presentan un número de eventos por celda igual al patrón AEC, no parecen adaptarse tan bien

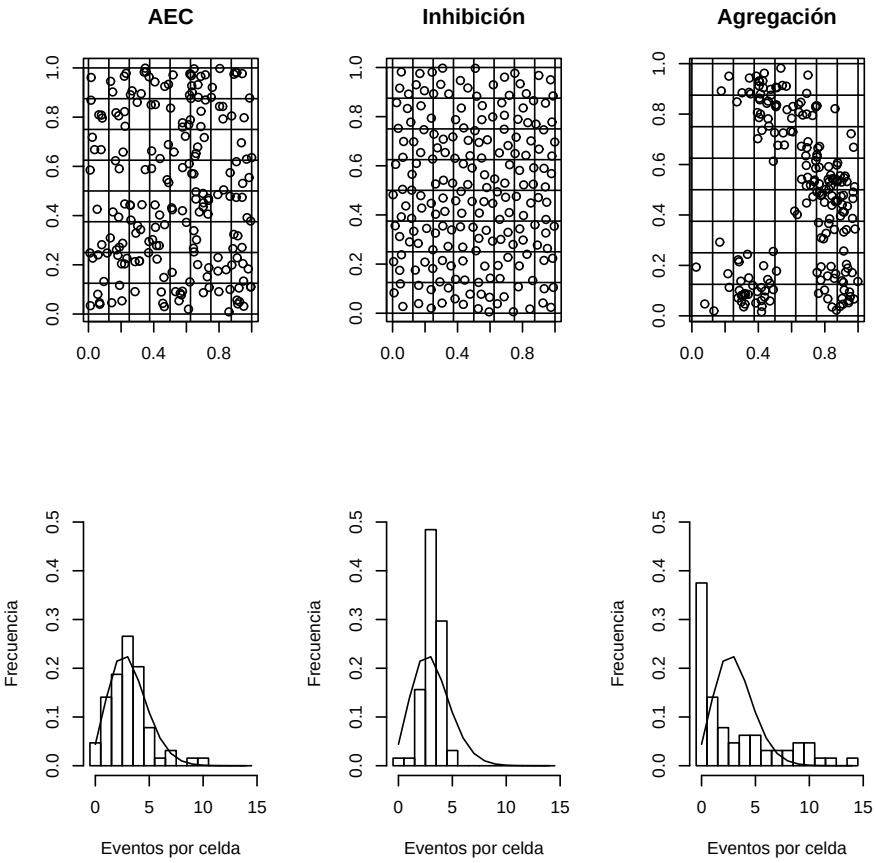

Figura 2.1: Patrones puntuales y distribución de su número de eventos por celda.

a la distribución de Poisson. Concretamente, se observa que en el patrón de inhibición la varianza del número de eventos por celda es menor que lo que correspondería a la hipótesis de Poisson. Este hecho se debe a que la inhibición impide que hayan celdas con un número excesivamente alto de puntos, al tiempo que para que quepan todos los puntos dentro de la misma región de estudio estos reproducen un patrón más ordenado que en el caso de AEC. Por tanto también resulta difícil encontrar celdas con un número pequeño de eventos o incluso vacías. Por el contrario, en el patrón agregado se puede apreciar que la varianza del número de observaciones por celda es mayor que para el caso de aleatoriedad completa. Este hecho se debe a que el proceso de agregación acumula casos en ciertas localizaciones lo que implica la existencia de celdas con un número alto de eventos mientras que para compensar este hecho otras celdas quedan más despobladas. El exceso de varianza respecto a la distribución de Poisson se dice *extravarianza de Poisson* y denotará la existencia de un factor que produce la aparición de clusters o agregaciones en el patrón puntual.

Una vez se dispone del número de eventos por celda, dados un patrón y una partición de la región de estudio, se habría de valorar si dichos datos son compatibles o no con la hipótesis de Poisson mediante un test estadístico. Un simple test χ^2 de bondad de ajuste para la distribución de Poisson sería suficiente para realizar dicha valoración. Para nuestro ejemplo de la figura 2.1 obtenemos los siguientes P-valores para el contraste de aleatoriedad espacial:

Patrón	AEC	Inhibición	Agregación
P-valor	0.14	6.3e-5	2.2e-16

En el caso que rechacemos la hipótesis de aleatoriedad la varianza rela-

tiva de Fisher (Fisher et al., 1922 [44]):

$$I = \frac{S^2}{\bar{X}} ,$$

nos proporcionará un índice de la agregación del patrón analizado. Valores de I superiores a 1 responderán a la presencia de sobredispersión, por tanto se corresponde con un patrón agregado, mientras que valores inferiores corresponderán a patrones de inhibición.

Existen distintas variantes del contraste que se acaba de proponer. Por ejemplo, en lugar de elegir una partición de la región de estudio se podría tomar una muestra de circunferencias extraídas de dicha región, o en lugar de circunferencias se podrían emplear cualquier otro tipo de regiones. Sin embargo en el caso de tomar una partición del espacio, sea cual sea la forma de los elementos de ésta, resulta posible emplear técnicas de datos en retículos (lattice) en los que el aprovechamiento de la estructura de las celdas confiere una mayor potencia al contraste de aleatoriedad. Es más, existen un gran número de tests basados en la valoración de la hipótesis de Poisson específicos para datos agregados espacialmente.

La principal ventaja de este grupo de contrastes reside en la sencillez de su implementación y su bajo coste computacional. Sin embargo, la aplicación de éstos presenta serios problemas, en primer lugar los resultados del test dependerán de la partición escogida y no existe un criterio objetivo para la elección de ésta. Concretamente, el tamaño de la partición habrá de elegirse de acuerdo con el tamaño de los clusters, lo cual puede no resultar obvio teniendo en cuenta que en ocasiones éstos ni siquiera existirán. Además, este tipo de contrastes, al estar basados en datos agregados, desaprovechan gran parte de la información que brindan los datos. Éste es un lujo que en ocasiones no se podrá permitir, sobre todo si el tamaño muestral del patrón puntual disponible es pequeño. Así, por ejemplo, este tipo de contrastes

son bastante habituales en el ámbito de estudios forestales o de botánica en general, donde el tamaño muestral del patrón puntual no suele ser un problema.

En el ámbito de la epidemiología la aplicación de estas técnicas presenta un problema añadido y es que en este caso, tal y como hemos comentado, la hipótesis de AEC no tiene excesivo sentido, por lo que querremos comparar con otras hipótesis. En este caso sería más razonable comparar con un proceso de Poisson en el que el número de casos esperados en una región sea proporcional a la población de dicha región y no a su área. Más concretamente nos interesaría determinar una partición del espacio con el mismo número de personas a riesgo (o al menos parecido) en cada región y aplicar las mismas ideas que hemos aplicado en el caso de contrastar AEC. Sin embargo, en general, no podremos elegir la partición de la región de estudio de la forma que queramos, ya que no dispondremos de información poblacional para cualquier partición. Es más, nos tendremos que limitar a las particiones de la región de estudio de las que se disponga información en fuentes de información como el censo, padrón o cualquier otra posibilidad. El problema es que la práctica en estudios epidemiológicos dista bastante de la teoría y las divisiones administrativas de las que se dispone habitualmente no tienen el mismo número de habitantes para cada región ni mucho menos. Por tanto, la viabilidad de la aplicación de estos estudios en el contexto epidemiológico es limitada. Así, por ejemplo, en Black et al. (1991) [19] se trata de solventar este problema proponiendo un método de agrupación de las regiones de estudio de forma que se obtengan unas nuevas con poblaciones similares. Sin embargo, este método propondrá divisiones menos finas que las que se disponían originalmente acrecentando el problema de la agregación de la información.

Los contrastes que hemos comentado hasta el momento valorarían la e-

xistencia de clustering general, aunque en ningún momento valoran la existencia de una agregación específica en algún lugar concreto, clustering individual. Sin embargo, Openshaw y Craft (1991) [75] proponen un método de clustering individual basado en la hipótesis de Poisson, la máquina de análisis geográfico (GAM). La idea es muy sencilla: dada una malla de puntos sobre la región de estudio y una serie de valores positivos $\{r_1, ..., r_n\}$ ascendentes, para cada punto de la malla se considera el conjunto de circunferencias centradas en éste y de radios r_1 hasta r_n. En la versión más sencilla del GAM, para cada una de las circunferencias consideradas se realiza un test χ^2 en el que se valora si el número de eventos dentro del círculo correspondería a un valor razonable suponiendo una distribución uniforme del riesgo tanto dentro como fuera de la circunferencia. Existen versiones más sofisticadas de este método pero están basadas en la misma idea. Respecto a la aplicación epidemiológica, la principal ventaja de este método es que, si bien resulta difícil saber la población a riesgo dentro de una circunferencia, resulta particularmente indicado en el caso de que se dispongan de controles o una muestra de la población. En dicho caso el test χ^2 se reduciría a una comparación del número de casos del patrón puntual y los controles tanto dentro como fuera de la circunferencia. Por tanto, la aplicación del GAM será viable, siempre que sea factible la obtención de la citada muestral poblacional o de controles. Sin embargo, la aplicación de esta técnica presenta otros inconvenientes. El principal se debe a la realización de comparaciones múltiples, ante la cual habrá de adoptarse alguna medida de protección de error. Es por ello que en distintas ocasiones se cita este método como herramienta descriptiva, no inferencial. No obstante, en la monografía de Alexander y Boyle (1996) [3] se realiza una comparativa de distintos métodos de detección de clusters y el GAM no resulta mal parado en esta comparación.

Como conclusión se puede afirmar que los métodos de valoración de

la hipótesis de Poisson presentan dificultades a la hora de su aplicación al ámbito epidemiológico. Su mayor virtud recae en su facilidad de implementación, pero dicha sencillez también produce conclusiones de bajo valor epidemiológico. En la siguiente sección se introducen los métodos de contraste basados en distancias que no agregan la información individual en grupos, a diferencia de los métodos presentados hasta el momento.

2.2.2. Contrastes basados en distancias

Estos contrastes están basados en la suposición que la posición relativa de los puntos de un patrón proporcionan información sobre el proceso puntual que los ha generado, e incluso en ocasiones son capaces de discriminar entre patrones correspondientes a distintos procesos puntuales. Existen distintas herramientas clásicas en la comparación de procesos puntuales, que valoran distintas características de éstos y en las que se basan sus comparaciones. Una de estas herramientas es la función K. Si el proceso estudiado es estacionario, ésta se define como:

$$K(t) = \lambda^{-1} E(N_0(t)) \ , \ \forall t \geq 0 \ ,$$

donde $N_0(t)$ será el número de eventos a una distancia menor que t de otro evento seleccionado al azar. Por tanto, dado un evento, la función K describe a que distancias nos encontramos nuevos eventos conforme nos alejamos del primero. En Ripley (1976) [81] se propone estimar la función K mediante la siguiente expresión:

$$\hat{K}(t) = \left(\frac{|A|}{(n-1)} \right)^{-1} \left(n^{-1} \sum_{i=1}^{n} \sum_{j \neq i} w_{ij}^{-1} 1(d_{ij} \leq t) \right) \ ,$$

donde A es la región de estudio, n el número de eventos, d_{ij} la distancia entre el evento i y el j y w_{ij} es el área de la intersección de A y el círculo

centrado en el evento i y de radio d_{ij}. Este último término se utiliza para llevar a cabo la corrección de arista sobre la estimación de la función.

Para un proceso homogéneo de Poisson se demuestra fácilmente que el valor esperado de dicha función valdrá πt^2 para cualquier valor de t. Si $K_0(t)$ denota la función K del proceso puntual con el que queramos comparar el patrón bajo estudio, la función

$$D(t) = K(t) - K_0(t), \;\; t \geq 0 \, ,$$

valorará la discrepancia entre ambos procesos, ya que aquellos valores de la función muy distintos de 0 indicarán discrepancias entre el patrón observado y el proceso puntual con el que se le compara. La expresión asintótica de la varianza para $D(t)$ se conoce en el caso homogéneo de Poisson, pero en general no resulta posible hallar dicha expresión. Por tanto, habremos de ayudarnos de métodos basados en simulación para valorar si la divergencia observada puede ser atribuida al azar o no. Para la realización de dicho contraste generaremos distintos patrones del proceso de referencia y para cada patrón que generemos calcularemos la función $D(t)$. En base a la muestra de funciones de discrepancia obtenidas se podrá calcular tanto el valor esperado de $D(t)$ para todos los valores de t que se considere como su intervalo de confianza.

En la figura 2.2 se pueden apreciar las funciones $D(t)$ resultantes de la comparación de los patrones de la figura 2.1 con un proceso homogéneo de Poisson. En ella se observa que en el primero caso las bandas de confianza (lineas discontínuas) contienen, para gran parte de las distancias valoradas, el valor 0, lo cual indica que dicho patrón es compatible con la hipótesis de AEC. Por el contrario, el intervalo de confianza en el patrón de inhibición no contiene al 0 para valores de distancias pequeños. En dicha figura se puede observar un hecho habitual en todos los patrones de inhibición, para

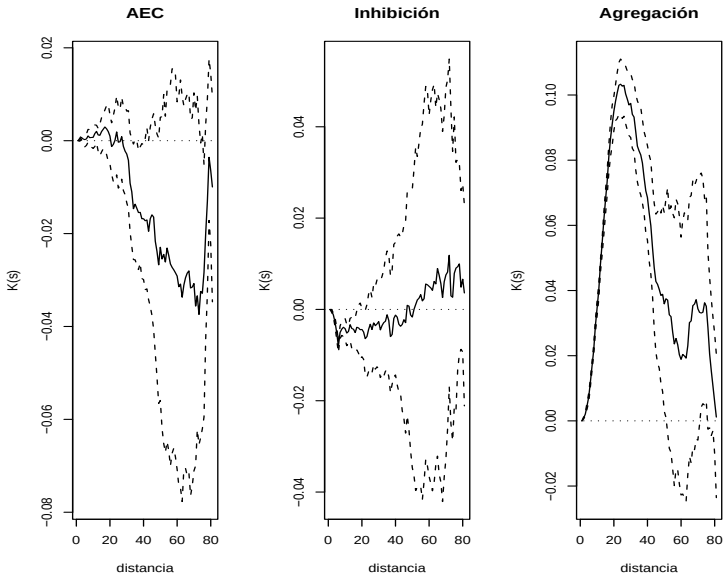

Figura 2.2: Función D y intervalo de confianza al 95 % para los patrones de la figura 2.1.

valores de distancias pequeñas la función $D(t)$ toma valores negativos. Esta situación se debe a que para patrones de inhibición nos encontramos menos eventos juntos de lo esperado bajo la hipótesis de AEC. Esto conlleva que $D(t)$ para dichos rangos de valores para las distancias sea negativo. Al contrario, en un patrón de agrupación esperamos observar valores positivos en la función $D(t)$ para distancias pequeñas, tal y como se ha dado en el tercero de los patrones donde se aprecia este efecto de forma muy acusada. La función $D(t)$ nos da información sobre si a cierta distancia t de cualquier evento nos encontramos más o menos eventos de los que esperaríamos, pero para ello se ha de realizar para cada valor de t un contraste. Por tanto, no nos bastará observar si el intervalo de confianza contiene en todo momento al 0 ya que es probable que este hecho no se de aunque el patrón de estudio cumpla AEC, simplemente por la probabilidad de que se den falsos negativos al realizarse un gran número de contrastes. Como solución a este problema se propone el cálculo del estadístico

$$D = \int_0^{d_0} \frac{D(t)}{\sqrt{Var(D(t))}} dt \ ,$$

para el patrón estudiado y utilizar este estadístico para contrastar globalmente si el patrón es compatible con el proceso que se quiere comparar, y no para un único t. En la expresión anterior d_0 es un parámetro fijado de antemano que se supone sería la distancia máxima a la que se espera encontrar un exceso o carencia de eventos. El contraste anterior se hará nuevamente comparando el valor del estadístico D para el patrón de estudio y para distintos patrones generados a partir del proceso puntual que se quiera contrastar. En concreto, para el ejemplo de la figura 2.1 y un valor de d_0 de 0.8 unidades (el rango de valores observable en la figura) obtenemos los siguientes resultados:

Patrón	AEC	Inhibición	Agregación
P-valor	0.264	0.044	0

Por tanto podemos concluir que para el segundo y tercer patrón, observando hasta una distancia de 0.8 unidades, encontramos evidencias de que no se sigue un proceso homogéneo de Poisson. Además, aunque para una distancia concreta de 0.35 unidades el primer patrón presenta una agregación menor de lo que se esperaría bajo la hipótesis de AEC globalmente el comportamiento de este patrón es completamente compatible con esta hipótesis.

A parte de la función K existen otras funciones que valoran otras facetas del patrón puntual y que complementarán la información proporcionada por la función K. Así, la función G se define como la función de distribución de la distancia desde un evento concreto hasta su evento más próximo. La expresión empírica de dicha función se expresa como:

$$\hat{G}(t) = n^{-1} \sum_{i=1}^{n} 1(d_i \leq t),\ \forall\ t \geq 0\ ,$$

donde d_i es la distancia del i-ésimo evento a su vecino más próximo. Por otra parte, se define F como la función de distribución de la distancia de cualquier punto de la región de estudio a su evento más próximo. Así, dado un conjunto de m componentes de la región de estudio, estimamos dicha función mediante:

$$\hat{F}(t) = m^{-1} \sum_{i=1}^{m} 1(d_i \leq t),\ \forall\ t \geq 0\ ,$$

donde d_i es la distancia desde el i-ésimo punto hasta su evento más próximo. Para un proceso homogéneo de Poisson se tiene $F(t) = G(t),\ \ \forall t$. Por tanto la función J definida como:

$$J(t) = \frac{1 - G(t)}{1 - F(t)},\ \ t \geq 0\ ,$$

valdrá 1 para dicho proceso. Concretamente, valores de $J(t)$ inferiores a 1 para valores pequeños de t indican agregación, mientras que valores superiores evidencian inhibición entre eventos.

De la misma forma que se ha descrito para la función K se podrán hacer contrastes de hipótesis basados en las funciones F, G y J. Es decir, mediante la generación de patrones de un proceso puntual se podrá comparar dicho proceso con un patrón sometido a estudio. Al igual que se hecho con la función K, también resulta posible la adaptación de la función D para comparar las nuevas funciones que se acaban de definir. De esta forma se dispondrán de nuevos criterios en los que basarse a la hora de comparar patrones puntuales con distintos procesos.

La aplicación de estas ideas al ámbito epidemiológico requiere ciertas adaptaciones ya que, tal y como se ha señalado, en este caso la comparación con la hipótesis homogénea de Poisson no tiene demasiado sentido. Además, no suele ser posible disponer de múltiples muestras poblacionales con las que comparar el patrón bajo estudio. En este caso la solución propuesta se basa en la utilización de una única muestra de controles[1]. Una vez se ha tomado la muestra de controles el objetivo será su comparación con la distribución de los casos. Conviene destacar que si ambas distribuciones coinciden, necesariamente habrán de coincidir sus estadísticos K, F, G y J. Por tanto podremos contrastar la igualdad de ambas distribuciones basándonos en las funciones anteriores. Ilustraremos el procedimiento mediante la función K aunque se puede proceder de forma análoga para el resto de funciones. El estadístico que vamos a utilizar para la comparación de ambos patrones es:

$$D(t) = K_{casos}(t) - K_{controles}(t), \quad t \geq 0 \; .$$

Los valores de D distintos de 0 indicarán discrepancias entre la distribución de casos y controles. Por tanto, se habrá de valorar si la magnitud de los valores de D se deben al azar o responden a diferencias entre la distribución de

[1]Recordamos que un control será un individuo de características similares a los casos y que les diferencia de éstos el no presentar la enfermedad bajo estudio

casos y controles. Ante la imposibilidad de recurrir a un número indefinido de muestras de controles para comparar con el patrón de los casos, se recurrirá a un procedimiento de *etiquetado aleatorio*. En éste se unen los n_{casos} y $n_{controles}$ en un único grupo, de ellos se eligen n_{casos} al azar, quedando a su vez $n_{controles}$ observaciones formando otro grupo. Para los dos nuevos grupos que se han muestreado se calcula una nueva función D. Repitiendo este proceso las veces que se considere oportuno se obtendrá la distribución del estadístico D bajo la hipótesis de etiquetado aleatorio de ambos conjuntos. De esta forma podremos comparar la función D original con la nueva distribución que acabamos de calcular y, por tanto, valorar la hipótesis de etiquetado independiente entre ambos conjuntos. Notar que la verificación de dicha hipótesis no implica que ambas distribuciones coincidan pero sí que será condición necesaria para ello.

La comparación de la función D haciendo uso de etiquetado aleatorio se describe en Diggle y Chetwynd (1991) [40]. Sin embargo la utilización de controles para tener en cuenta la distribución poblacional ya fue utilizada anteriormente en Cuzick y Edwards (1990) [32]. En dicho trabajo se introduce la idea de etiquetado aleatorio de casos y controles, aunque el estadístico utilizado en este trabajo es el número de casos entre los k eventos más próximos de cada caso. Este estadístico es menos potente que el utilizado en Diggle y Chetwynd (1991) [40]. Además las conclusiones del test de Cuzick y Edwards resultan poco claras ya que son del tipo "existe agregación atendiendo a los 6 vecinos más próximos". Entonces se define la magnitud de la agregación en términos de distancias entre vecinos, en lugar de distancia física. Por tanto el test original ha sido reemplazado por el propuesto por Diggle y Chetwynd, que es más potente y proporciona conclusiones más intuitivas.

Los procedimientos descritos en el presente capítulo contrastan la exis-

tencia de un patrón de agregación general, es decir, si los eventos tienden a agruparse entre sí aunque no necesariamente alrededor de una localización concreta. Entonces, en caso de aceptar la existencia de agregación en el patrón bajo estudio, los contrastes utilizados no proporcionarán información sobre la localización de las posibles agregaciones. Es por ello que una vez se haya determinado la existencia de agregación en el patrón puntual se habrá de estimar su función de intensidad para determinar la localización de aquellos lugares con mayor riesgo de presentar un caso.

2.3. Estimación no paramétrica de la función de intensidad

La estimación de la función de intensidad de forma no paramétrica se lleva a cabo haciendo uso de métodos de estimación mediante funciones kernel (Diggle, 2003 [39], Cressie, 1993 [31]). Ésta se basa en el establecimiento de una elevación del riesgo alrededor de cada evento del proceso puntual y en base a la agregación de dichas elevaciones se definirá la función de intensidad. De esta forma allá donde haya una acumulación de casos la función de intensidad tomará valores superiores como consecuencia de esta agregación. Para poner en marcha la estimación de la función de intensidad se ha de determinar previamente la función kernel que define la elevación del riesgo alrededor de cada evento. Suele ser habitual la función kernel cuártica

$$k(u) = \begin{cases} 3\pi^{-1}(1-u^2)^2 & 0 \leq u \leq 1 \\ 0 & u > 1 \end{cases},$$

aunque también existen otras posibilidades como la elección de un kernel constante alrededor de un disco centrado en cada evento, funciones que decaen de forma exponencial o simplemente funciones normales bivariantes. Una vez se ha determinado la función kernel a utilizar, se define la función

de intensidad como

$$\hat{\lambda}_h(u) = \frac{1}{p_h(u)} \left\{ \sum_{i=1}^{n} k\left(\frac{u-u_i}{h}\right) \right\} ,$$

donde $p_h(u)$ es un factor que lleva a cabo lleva a cabo la corrección de arista en la estimación de la intensidad, u_i la localización de los eventos del patrón puntual. El parámetro h se dice la ventana de la función kernel y controla el grado de suavización de la función de intensidad ajustada. Este parámetro controla la distancia hasta la que se produce una elevación del riesgo como consecuencia de la presencia de un evento. Resulta interesante señalar que el efecto de la ventana del kernel sobre la estimación de la función de intensidad resulta mucho más crítico que el efecto de la propia función kernel elegida. Es más, la elección del parámetro de la función de suavización tendrá un gran efecto sobre la apariencia de la función de intensidad estimada. Por ello se ha de llevar especial cuidado en la elección de este parámetro. En Diggle (1985) [37] se propone un método para la elección del parámetro de suavización. Dicho método se basa en la minimización del error cuadrático medio de la estimación kernel y la función de intensidad para distintos valores del parámetro de ventana. No obstante la función a minimizar suele ser bastante plana alrededor de su mínimo, señalando un conjunto de valores posibles más que un único valor, por lo que en muchos casos resulta un tanto ambiguo el estimador puntual proporcionado por este criterio.

Sin embargo, tal y como se ha señalado ya, cuando se disponga de un conjunto de eventos asociados a un brote epidémico el objetivo fundamental puede que no sea la estimación de la función de intensidad. Esto se debe a que las localizaciones de mayor intensidad no han de ser necesariamente consecuencia de un mayor riesgo de enfermar en dicho lugar, sino que puede responder a la presencia de una densidad de población mayor. En este caso resulta de mayor interés la estimación de la *función de riesgo* que describe las diferencias entre las funciones de intensidad de dos procesos, en este

caso el proceso puntual de los casos y el de la población o un conjunto de controles. Más concretamente, en el caso de disponer de un conjunto de casos y otro de controles, la función de riesgo $r(u)$, en cualquier punto de la región de estudio S, vendrá definida por

$$r(u) = \frac{\lambda_{casos}(u)}{\lambda_{controles}(u)} \quad u \in S \ .$$

En consecuencia, si dos patrones se distribuyen geográficamente de la misma forma, sus funciones de intensidad serán proporcionales y la función de riesgo que compara ambos procesos será constante.

La estimación de la función de riesgo también se realizará mediante métodos kernel. Concretamente, se empleará una estimación kernel de la intensidad tanto para el patrón de los casos como para el de los controles y en base a un cociente de estas estimaciones se determinará la función de riesgo. En Kelsall y Diggle (1998) [58] se describe un procedimiento para la estimación del parámetro de suavización de la función kernel siempre que se disponga de una muestra de casos y controles. En dicha ocasión el procedimiento se basa en la validación cruzada de la característica caso-control de cada uno de los eventos disponibles para distintos valores del parámetro de ventana. Respecto a la significatividad de la superficie de intensidad en cada uno los puntos de la región de estudio en Martínez-Beneito et al. (2005) [66] se describe un método para la valoración de dicha significatividad. Además, en este mismo trabajo junto a Abellán et al. (2002) [1] se describe la aplicación de las técnicas de patrones puntuales descritas, en los brotes de legionelosis de Alcoi introducidos en el capítulo anterior.

La modelización kernel no paramétrica de la función de intensidad supone una herramienta muy útil para la descripción de un proceso puntual. Sin embargo habrá ciertas cuestiones a las que dicha estimación no proporcionará respuesta. Por ejemplo, ¿cuál es la estimación del número medio de

agrupaciones en el proceso de los casos? y, si es posible, ¿cuál es la distribución de este valor? y ¿hasta qué distancia se extiende cada una de las agrupaciones?. Además, la metodología propuesta hasta el momento es muy adecuada para el estudio de agrupaciones generales, pero tiene una baja potencia estadística a la hora de valorar la existencia de ciertos clusters específicos. La respuesta a este tipo de preguntas tiene que venir dada por algún método de estimación que contemple la existencia de clusters individuales. Por tanto se habrá de recurrir a la modelización de la superficie de intensidad para responder a este tipo de cuestiones.

2.4. Modelización de procesos puntuales

Tal y como ya se ha comentado, los procesos homogéneos de Poisson constituyen los procesos puntuales más sencillos. Sin embargo, éstos no son ni mucho menos los únicos procesos posibles. La primera alternativa obvia a los procesos homogéneos consiste en la relajación de la estacionariedad de la media de este proceso. El resultado de esta relajación da lugar a los *procesos no homogéneos de Poisson* que se definen como aquellos procesos que cumplen las siguientes condiciones.

- La variable aleatoria $N(A)$ que describe el número de casos observados en cualquier región A, sigue una distribución de Poisson de media
$$\lambda(A) = \int_A \lambda(s)ds \ .$$
- Condicionado al número de eventos generados por el proceso en la región A, las localizaciones de éstos son una muestra aleatoria de realizaciones independientes con función de densidad
$$\frac{\lambda(s)}{\int_A \lambda(s)} \ . \tag{2.1}$$

La definición anterior está sujeta a que la función de intensidad tenga integral finita para cualquier subconjunto de la región en la que se ha definido el proceso. Por tanto, a diferencia del proceso homogéneo, el caso no homogéneo permite que la intensidad del proceso varíe en función de la localización geográfica. Los procesos de Poisson no homogéneos son una de las alternativas para describir un patrón en el que la hipótesis de aleatoriedad espacial completa no resulte sostenible. Además, la propuesta no homogénea permite la incorporación de covariables asociadas a localizaciones geográficas, la introducción de gradientes geográficos o dotar de una forma concreta la superficie de intensidad.

La definición de la superficie de intensidad para el caso no homogéneo requerirá la elección de una forma paramétrica para ésta. Dicha forma paramétrica habrá de tomar necesariamente valores positivos sobre la región de estudio. Así suele ser habitual definir la intensidad como exponencial de una función que puede depender o bien de covariables o de las coordenadas geográficas. En ese caso la superficie de intensidad dependerá de una serie de parámetros que se habrán de estimar recurriendo a los métodos de inferencia estadística oportunos. Sin embargo, estas opciones suelen ser poco flexibles para describir la estructura de un proceso puntual, por lo que suele ser común recurrir a formas más elaboradas para definir la función de intensidad.

Para la estimación de los parámetros de la intensidad del proceso, se habrá de tener en cuenta que el número de eventos generados es una variable más del proceso. Es decir, un proceso puntual generará conjuntos de casos en los que una de las cuestiones que se desconocen es su número de observaciones. En la definición de proceso de Poisson no homogéneo, ecuación (2.1), se muestra la función de verosimilitud de los datos condicionada al número de observaciones del patrón, que tal y como se ha comentado es

desconocido antes de disponer los datos. La función de verosimilitud que habremos de utilizar para la estimación de un proceso de Poisson en el caso de que el patrón dispuesto no haya sido generado condicionado a un cierto número de observaciones será:

$$f((s_1, ..., s_n), n) = \frac{\prod_{i=1}^{n} \lambda(s_i)}{n!} \exp\left(-\int_A \lambda(s)ds\right) , \tag{2.2}$$

donde A es la región de estudio en la que se ha definido el proceso. Notar que en la expresión anterior se podrá evaluar la probabilidad de cualquier patrón independientemente del número de eventos que disponga, es más, la suma de todas estas probabilidades valdrá 1. Por tanto, un proceso puntual establece la probabilidad de cada patrón puntual independientemente del número de eventos que disponga, es más, el dominio de la función de probabilidad se corresponde con todos los conjuntos de puntos finitos sobre la región de estudio. La deducción de la expresión (2.2) se desarrolla con todo detalle en las páginas 621-622 de Cressie (1993) [31].

Los procesos no homogéneos de Poisson suponen una generalización de la versión homogénea, que puede servir como alternativa a ésta cuando su ajuste evidencie sobredispersión. En dicho caso un modelo no homogéneo posibilitará que la función de intensidad se adapte a las zonas con mayor y menor presencia de puntos. De esta forma el valor de $\int_A \lambda(s)ds$ se adecuará mejor a los valores de $N(A)$ para cualquier conjunto A, por lo que disminuirá la varianza de dicha variable y en consecuencia la sobredispersión del proceso. Sin embargo, en muchas ocasiones los procesos de Poisson no tienen la flexibilidad suficiente para describir el comportamiento de un patrón puntual de forma apropiada. Este hecho suele ser particularmente usual en epidemiología, donde la presencia de factores de riesgo ambientales, que actúan de forma distinta en cada localización de la región de estudio, condicionan el riesgo asociado a cada emplazamiento en el que se estudia el proceso. La presencia de estos factores ambientales producirán agregaciones

de tipo general ocasionadas por el reparto, de forma más o menos suave, de los factores de riesgo en la población. El efecto de esta heterogeneidad medioambiental será muchas veces imprevisible, en consecuencia, resulta difícil elegir una forma paramétrica concreta para un proceso de Poisson a la que el patrón bajo estudio se adapte plenamente. Es decir, la propuesta de modelos de Poisson no homogéneos requiere la definición de una forma paramétrica que no siempre resultará evidente, ni siquiera intuitiva. De esta manera, la utilización de modelos de Poisson no suele suponer una alternativa sencilla en el caso que un modelo homogéneo no sea apropiado para describir un patrón puntual.

Una segunda alternativa a los procesos homogéneos de Poisson consistiría en considerar la heterogeneidad medioambiental, a la que nos referíamos anteriormente, como de origen aleatorio y tratar ésta como tal. En ese caso, el proceso estocástico que explicaría la generación de los casos es el conocido como *proceso de Cox*, introducido originalmente en Cox (1955) [30] para problemas unidimensionales. La definición formal de este tipo de procesos viene dada por las siguientes dos condiciones:

- $\{\Lambda(s) : s \in A\}$ es un proceso estocástico de valores positivos,
- Condicionado a $\{\Lambda(s) = \lambda(s) : s \in A\}$, los eventos forman un proceso de Poisson homogéneo de intensidad $\lambda(s)$.

Por tanto, una vez conocida la función de intensidad, un proceso de Cox se reduce a un proceso de Poisson. La particularidad de los procesos de Cox es que la función de intensidad en éstos es estocástica, a diferencia de los procesos de Poisson en los que la intensidad pertenece a una familia paramétrica de la que habremos de estimar sus parámetros. Por ello, a los procesos de Cox también se les conoce como *procesos doblemente estocásti-*

cos. Existe un paralelismo directo entre la modelización estadística mediante efectos fijos y los procesos de Poisson mientras por otro lado los procesos de Cox pueden ser considerados como modelos de efectos aleatorios.

Además de los procesos de Poisson y Cox existen otros procesos puntuales que no están basados únicamente en una superficie de intensidad. Estos procesos no suponen los eventos como realizaciones independientes de un mecanismo cuya función de probabilidad viene dada por una superficie de intensidad sino que se incorpora dependencia entre los eventos del proceso. Suele ser habitual modelizar dicha dependencia en función de la distancia entre pares de puntos, aunque también es posible hacerlo en función de la posición relativa de conjuntos de 3 o más puntos. En un *proceso puntual de interacción entre pares*, la función de verosimilitud de cualquier patrón puntual es de la forma:

$$f(\{x\}) \propto \prod_{\xi \in \{x\}} \phi(\xi) \prod_{\{\xi,\eta\} \subseteq \{x\}} \phi(\{\xi,\eta\}) \, .$$

En la expresión anterior $\{x\}$ representa el conjunto de observaciones del patrón puntual. Los procesos de Poisson son un caso particular de los procesos de interacción en los que $\phi(\{\cdot,\cdot\}) = 1$. Dentro de este tipo de procesos resultan particularmente utilizados los *procesos de Strauss*, en los que la función de interacción entre puntos se define como:

$$\phi(\{\xi,\eta\}) = \gamma^{1(\|\xi-\eta\| \leq R)} \, , \tag{2.3}$$

donde $0 \leq \gamma \leq 1$ y $R > 0$. Notar que el proceso de Strauss así definido establece inhibición entre los eventos del proceso. Esta inhibición dependerá de R, la distancia máxima a la que existe interacción entre eventos, y γ, que controla la potencia del efecto inhibidor.

De aquí en adelante no dedicaremos más atención a los modelos de interacción entre eventos ya que éstos no resultan adecuados para el estudio

de los problemas epidemiológicos que nos ocupan. Este hecho se debe a que el mecanismo de trasmisión de la legionela no responde a contagio directo entre personas. Por lo que, la presencia de un caso en cualquier localización de la región de estudio no supondrá un aumento de riesgo en sus alrededores asociado a la presencia de dicho caso. De esta manera, la inclusión de interacción en la modelización de enfermedades no contagiosas no tiene sentido desde un punto de vista epidemiológico. Por todo ello, a partir de este momento obviaremos su estudio.

Una vez se ha descrito el marco teórico en el que se fundamenta gran parte de la teoría de procesos puntuales vamos a introducir algunos modelos de particular interés. Los modelos en que nos vamos a detener, según se califica en Möller (2003) [68], suponen las propuestas de mayor utilización y proyección en la aplicación de la teoría de procesos puntuales en estudios epidemiológicos. Nuestra opinión sobre la importancia de estos modelos coincide en este sentido, ya que por su flexibilidad estas propuestas suponen unas herramientas de gran utilidad para describir variaciones geográficas de riesgos.

2.4.1. Procesos basados en teselaciones

Los procesos puntuales basados en teselaciones hacen uso de estas construcciones matemáticas para la definición de la superficie de intensidad en un proceso de Poisson no homogéneo. Una *teselación* consiste en una partición de la región de estudio en distintos polígonos según cierto criterio. Dado un conjunto de puntos $\{g_k\}$ sobre cierta región finita, se define la *teselación de Voronoi* asociada como, aquella partición tal que cada localización pertenece al polígono que contiene al punto de $\{g_k\}$ más cercano a dicha localización. Existen otras construcciones que darán lugar a otros tipos de teselaciones

o triangulaciones del espacio aunque la teselación de Voronoi es la más utilizada en el ámbito de procesos puntuales.

En Arjas y Gasbarra (1994) [5], Arjas y Heikkinen (1997) [6], Heikkinen y Arjas (1998)[55] se utiliza el conjunto de las teselaciones de Voronoi de la región de estudio para definir la función de intensidad de un proceso puntual, tanto en la recta real como en una región del plano. La inferencia para esta propuesta se realiza desde un enfoque bayesiano. En concreto, si $\{A_k\}$ es la teselación asociada al conjunto de puntos $\{g_k\}$, Heikkinen y Arjas [55] definen el soporte de la función de intensidad como el conjunto de funciones escalonadas:

$$\{\Lambda(s) = \sum_k \Lambda_k 1(s \in A_k) : \Lambda_k > 0, \{g_k\} \subseteq A\} \ .$$

Para completar la definición del proceso, $\{g_k\}$ corresponde corresponde a la realización de un proceso de Poisson homogéneo sobre la región de estudio A. Por otra parte a $\{log(\Lambda_k)\}$, para todos los valores de k, se les da una distribución inicial simultánea autoregresiva (SAR) de estructura espacial Cressie (1993) [31], en la que se considera dependencia entre aquellos valores correspondientes a polígonos adyacentes en la teselación.

La estimación de la función de intensidad se lleva a cabo mediante métodos de Monte Carlo mediante Cadenas de Markov (MCMC) (Gilks et al., 1995 [50]). Como consecuencia de la simulación, en la que en cada iteración se dispondrán de nuevos conjuntos para $\{g_k\}$ y $\{\Lambda_k\}$, obtendremos una colección de funciones de intensidad escalonadas procedentes del proceso de estimación. Así, aunque el resultado de la simulación sea un conjunto de funciones escalonadas, la estimación final de la función de intensidad será el promedio de todas estas funciones. En consecuencia, dicha estimación no será una función escalonada, sino que será una superficie suave al ser promedio de distintas funciones continuas salvo en un conjunto de me-

dida nula que cambiará de iteración a iteración. El planteamiento de este modelo desde un enfoque frecuentista no tiene excesivo sentido ya que en dicho caso se obtendría el estimador puntual más verosímil de la teselación. Por tanto la función de intensidad resultante será una función escalonada con discontinuidades, de limitada interpretación y utilidad a fines prácticos.

La estructura espacial del valor de la función de intensidad en cada polígono proporcionará una configuración de la superficie de intensidad con variaciones suaves. De esta forma se penalizan los cambios de la función de intensidad excesivamente abruptos que podrían no ser demasiado razonables. También se evita de esta forma la existencia de soluciones degeneradas en las que cada punto del proceso se modelice mediante un polígono de la teselación de área pequeña e intensidad grande, mientras que el resto de polígonos tome valores muy bajos del parámetro de intensidad. Además, en la modelización propuesta, al penalizar cambios abruptos de la función de intensidad, se compensa dicha penalización mediante la inclusión de un mayor número de puntos de la teselación en las zonas con mayor cambio de la intensidad. De esta manera, las regiones del proceso en la que la intensidad es más cambiante, reciben mayor atención en la modelización al utilizarse un número mayor de polígonos de la teselación en estas zonas. Este hecho es un efecto secundario muy deseable de la modelización de los valores de la función de intensidad.

El uso de teselaciones como método no paramétrico de descripción de funciones de distribución o de intensidad no es patrimonio exclusivo de la teoría de procesos puntuales, sino que está experimentando un auge también en otro tipo de aplicaciones. Así, dentro del contexto de la estadística espacial, en Knorr-held y Rasser (2000) [59] y Ferreira et al. (2002) [43], entre otros, se aplican el uso de estas técnicas en el contexto de la distribución geográfica de riesgos para datos agregados. Mientras, en Stephenson

et al. (2003) [94] se aplica este tipo de ideas para el tratamiento de datos geoestadísticos en el que no se cumple la estacionariedad de segundo orden.

La aplicación de estas técnicas para la estimación de funciones de intensidad presentan, desde nuestro punto de vista, ciertos problemas. Así, en los artículos donde se desarrollan estas técnicas se fijan los valores de los hiperparámetros del modelo de forma un tanto arbitraria y no existe una forma clara de definir dichos valores. Como solución se podría incluir alguna capa más en el modelo para expresar el desconocimiento que tenemos de estos parámetros a priori. Sin embargo, hasta donde nosotros conocemos, se ha realizado muy poco trabajo en este sentido a pesar que desde nuestro punto de vista, habría que dedicarle más atención. Por otro lado, si bien la estimación de la superficie de intensidad es muy flexible y expresará de forma adecuada la información resultante de los datos, las conclusiones de la inferencia no serán del todo adecuadas para el problema que abordamos en el presente trabajo. Así, tras la estimación de la superficie de intensidad desconoceremos la distribución del número de agrupaciones existente a la vista de la distribución de los eventos o cual es la forma y extensión de estas agrupaciones.

2.4.2. Procesos cluster de Poisson

Los *procesos cluster de Poisson* o *procesos de Neyman-Scott* (Neyman-Scott, 1958 [73]) suponen una propuesta muy adecuada para la modelización de agrupaciones individuales de eventos. El resto de modelos introducidos tienen una vocación más orientada hacia la detección de agrupaciones generales, es decir la modelización de la heterogeneidad medioambiental. Los procesos cluster de Poisson se definen como aquellos procesos que cumplen las siguientes propiedades:

- Existe un proceso de "padres" que sigue un proceso de Poisson de intensidad ρ.
- Cada padre produce un número aleatorio de descendientes M, dichos valores se distribuyen independientes e idénticamente distribuidos *(i.i.d.)* según una función de distribución discreta p_m.
- Las posiciones de los descendientes respecto a su padre se distribuyen i.i.d. de acuerdo con una distribución de probabilidad $h(\cdot)$.

Cuando h sigue una distribución normal bivariante de matriz de varianza-covarianza proporcional a la matriz identidad, el proceso cluster se dice *proceso de Thomas*. Si h sigue una distribución uniforme sobre un disco centrado en la localización de los padres diremos que observamos un *proceso cluster de Matérn*.

A diferencia de los modelos vistos hasta ahora, el proceso cluster de Poisson contempla una estructura de padres-hijos en los datos para modelizar la agregación de sus eventos. Una de las dificultades añadidas en la estimación de este tipo de procesos radica en que únicamente se conoce la localización de los eventos descendientes, por tanto no se conoce ni el número de padres ni mucho menos sus localizaciones. La distribución de estos últimos será uno de los principales objetivos de la inferencia estadística.

En Cressie (1993) [31] se demuestra que el proceso de Neyman-Scott en el que el número de casos en cada cluster sigue una distribución de Poisson, es matemáticamente equivalente a un proceso de Cox donde la superficie de intensidad es de la forma:

$$\lambda(s) = \alpha \sum_{c \in C} h(s - c) \, , \tag{2.4}$$

donde c son valores generados de un proceso de Poisson de intensidad ρ y α es el número de casos esperado de la distribución de Poisson p_m.

La aplicación del proceso de Neyman-Scott para la modelización del riesgo en problemas epidemiológicos tiene la ventaja de contemplar específicamente los clusters de tipo individual de los eventos. Este hecho supondrá una mejora de la potencia en la detección de ese tipo de agregaciones respecto a otras modelizaciones de la superficie de riesgo, como por ejemplo la de teselaciones. Por el contrario, los procesos cluster de Poisson, en general no incorporan la heterogeneidad medioambiental en la definición de la superficie de riesgo, por lo que ignoran esta fuente de variación inherente a la mayoría de problemas epidemiológicos y que en ciertos casos podrá influir en gran medida sobre los resultados de la inferencia del proceso.

La inferencia en este tipo de modelos se puede llevar a cabo tanto desde un enfoque clásico como bayesiano. En particular, en Castelloe (1998) [26] se realiza un amplio estudio de la aplicación de ambas aproximaciones. El número de parámetros a estimar en un proceso cluster de Poisson dependerá del número de padres del proceso, que en general es desconocido. Desde la perspectiva bayesiana este parámetro se puede tratar como una variable y aprender sobre ella, mientras que, desde el punto de vista clásico, este parámetro se habrá de fijar a un valor concreto y se estimarán el resto de parámetros del modelo condicionados a dicho valor. El aprendizaje sobre el número de padres de un proceso de Poisson suele ser una de las características más interesantes sobre la que realizar la inferencia en este tipo de modelos. Por tanto, la aproximación bayesiana goza de particular interés al proporcionar información sobre la distribución de este valor.

La aplicación de este tipo de modelización en el ámbito de la epidemiología, aunque no está demasiado extendida, ya goza de algún antecedente.

Concretamente, Lawson y Clark (1999) [62] valoran la existencia de una o más agregaciones específicas de casos en torno a uno o varios emplazamientos concretos, desde una perspectiva bayesiana. Posteriormente ahondaremos más en la aplicación de estos modelos en el ámbito de estudios epidemiológicos.

2.4.3. Modelización Poisson/Gamma

La tercera propuesta que vamos a destacar dentro del análisis de procesos epidémicos no contagiosos también consiste en un proceso de Cox. La modelización empleada se realiza dentro del marco bayesiano y se introduce en Wolpert y Ickstadt (1998) [100] e Ickstadt y Wolpert (1999) [56]. Se ha incluido una versión discreta de esta propuesta en la última versión del software WinBUGS, 1.4.1, por lo que esperamos que la popularidad de esta modelización se acreciente en un futuro.

La propuesta de modelización Poisson/Gamma generaliza la idea de regresión mediante medias móviles y la regresión local mediante herramientas de kernel al campo de los procesos puntuales. Así, si $k(u,s)$ es una función kernel, es decir

$$\int_U k(u,s)du = \int_S k(u,s)ds = 1 \ ,$$

la superficie de intensidad en el modelo de Cox Poisson/Gamma viene definida como

$$\Lambda(s) = \int_B k(u,s)\Gamma(du) \ ,$$

donde Γ es un campo de variables aleatorias que tienen como distribución

$$\Gamma(du) \sim \text{Gamma}(\alpha(du), \tau(u)) \ .$$

De esta manera, la superficie de intensidad del proceso Poisson/Gamma consiste de la media local ponderada de infinitas variables con distribución

Gamma, donde los pesos de la media vienen definidos por la función kernel. Γ se define sobre una región B que contiene a la zona de estudio A de forma que se minimiza el efecto de la frontera de la región de estudio.

La implementación del presente modelo requiere la generación de un campo de valores aleatorios sobre la región de estudio. En Wolpert y Ickstadt (1998) [100] se describe un proceso mediante el cual sólo se ha de generar un número de valores finito de dicho campo. Mediante el proceso propuesto se generarán únicamente los valores del campo aleatorio superiores a cierto umbral mientras que los valores inferiores a dicho umbral son ignorados. Teniendo en cuenta la discretización propuesta, la función de intensidad se puede reexpresar como:

$$\Lambda(s) = \sum_j k(s, u_j)\gamma_j \ , \tag{2.5}$$

donde el parámetro j toma un número de valores finito, γ_j son valores de una distribución Gamma superior a cierto umbral y u_j son distintas localizaciones de la región de estudio. Por tanto, el método de inferencia propuesto es aproximado, aunque estableciendo el citado umbral en valores muy bajos (por ejemplo la tolerancia de la máquina que realice el proceso de simulación) el error cometido puede resultar mínimo. La inferencia del modelo se realiza mediante métodos de simulación MCMC.

El modelo Poisson/Gamma se puede entender como una suavización, mediante funciones kernel, de los impulsos del campo aleatorio Γ. El grado de suavización en cualquier punto s de la región de estudio dependerá de la función kernel escogida, donde puede elegirse por ejemplo un disco centrado en s o una distribución normal centrada en dicho punto. Los parámetros de la función kernel, bien el radio del disco o la precisión de la normal, también pueden tratarse como valores desconocidos dentro del proceso de inferencia bayesiana si se considera oportuno. Además, también se pueden

modelizar los parámetros de la distribución del campo Γ en función de ciertos parámetros e intentar aprender sobre éstos. De esta forma, el modelo ganará en flexibilidad y será capaz de captar distintos comportamientos en los datos o incluso dotarles de cierta estructura.

La ecuación (2.5) establece cierto paralelismo entre la modelización mediante procesos Poisson/Gamma y los procesos cluster de Poisson. De hecho los primeros se pueden entender como límite de los procesos cluster de Poisson dentro del marco de los *shot-noise G cox processes* (Brix, 1999 [20]). En este modelo la superficie de intensidad toma la forma (2.5), pero el campo γ_j no sigue una distribución Gamma sino una generalización de ésta. Por tanto podemos entender los procesos Poisson/Gamma como una generalización de los procesos cluster de Poisson en los que cada padre tiene asociado un peso distinto (con distribución Gamma) y en los que el número de padres del proceso no es finito.

La aplicación de este tipo de modelización en problemas epidémicos, como por ejemplo en Best et al. (2000) [16], goza de particular interés, ya que permite la incorporación de información con distintos niveles de desagregación de forma consistente. Dicha consistencia concierne a la agregación espacial de datos, es decir, si se discretiza la superficie de intensidad según ciertas regiones, la función de intensidad en cada una de éstas coincidirá con la integral de la superficie de intensidad sobre dicha región. De esta forma resulta natural extrapolar los resultados obtenidos a cierto nivel de resolución a desagregaciones geográficas menos finas. Este hecho que puede parecer tan obvio no se da en muchas otras propuestas de modelización.

Sin embargo, la modelización Poisson/Gamma tampoco contempla específicamente la existencia de agregaciones individuales, sino que supone variabilidad de tipo general en toda la región de estudio. En consecuencia,

esta propuesta resulta más adecuada para la modelización de heterogeneidad medioambiental ya que no tendrá excesiva potencia para detectar agregaciones específicas. Más aún, esta propuesta presenta los mismos problemas de interpretación que los modelos de teselaciones, ya que no responde a cuestiones como ¿cuántas agrupaciones existen?, ¿cuáles son sus localizaciones? o ¿cuál es la extensión de dichas agrupaciones?. Los procesos cluster de Poisson respondían de mejor forma a estas cuestiones ya que realizaban una modelización específica de estos aspectos.

2.4.4. Procesos de Cox log-gaussianos

La última de las modelizaciones que vamos a destacar también se corresponde con un proceso de Cox. Este modelo se introduce por primera vez en Möller et al. (1998) [69] y se ha aplicado, entre otros, al estudio geográfico de la incidencia de encefalitis en Benes et al. (2003) [10]. El *proceso de Cox log-gaussiano* generaliza la idea de autocorrelación del estudio de series temporales al campo de los procesos puntuales.

En un proceso de Cox log-gaussiano la función de intensidad toma la forma

$$\Lambda(s) = \exp(Z(s)) \ ,$$

donde $Z(s)$ es un campo de variables gaussianas definidas sobre la región de estudio A. En este caso la estructura espacial viene definida por la existencia de correlación entre los valores del campo aleatorio Gaussiano, por lo que la covarianza de la variable Z entre dos localizaciones de la región de estudio $c(s,t) = Cov(Z(s), Z(t))$ se hace depender de la posición relativa de ambas localizaciones. Para que el proceso puntual esté bien definido, c habrá de cumplir ciertas condiciones de suavidad que se estudian en Möller et al. (1998) [69].

Suele ser habitual suponer que el proceso Z es estacionario e isotrópico, en cuyo caso se tiene que

$$c(s,t) = c(\|s-t\|) \; \forall \; s,t \in A \; .$$

Existe una amplia colección de funciones que pueden utilizarse para definir la covarianza entre distintas localizaciones del proceso y que cumplen la condición anterior. Citamos como ejemplo la clase de funciones *power-exponential*

$$c(u|\phi,\kappa,\sigma) = \sigma^2 \exp(-\phi u^{\kappa}) \; ,$$

en la que el parámetro ϕ controla la máxima distancia a la que se puede admitir que los datos son dependientes y el parámetro κ controla la suavidad de la superficie de intensidad resultante. En Schlather (1999) [86] se describe con detalle distintas familias de funciones válidas para definir estructuras de covarianza.

La estimación de un proceso de Cox log-gaussiano requiere la estimación del campo aleatorio subyacente Z. Sin embargo, dicha estimación habrá de hacer uso de alguna simplificación ya que dicho campo consta de un número infinito de valores, lo que la convierte en computacionalmente inabordable. La simplificación que se adopta en este caso se basa en la consideración de una fina malla que divide la región de estudio en pequeños trozos, y en cada una de estas secciones se considera un valor constante del campo aleatorio gaussiano. De esta forma, la estimación del campo aleatorio resulta manejable y se puede hacer más precisa conforme más fina sea la partición escogida de la región de estudio. La inferencia en este tipo de procesos se lleva a cabo desde un enfoque bayesiano y el proceso de estimación hace uso de métodos de simulación MCMC.

Respecto a las ventajas de esta modelización cabría señalar su flexibilidad, ya que el campo aleatorio de variables gaussianas es capaz de adaptarse

en gran medida a un amplio abanico de situaciones. Además, el término exponencial permite a esta modelización describir de forma razonable incluso agrupaciones de casos muy acusadas. Sin embargo, la detección de agrupaciones aisladas puede llegar a resultar difícil para esta propuesta ya que en el caso de que exista poca heterogeneidad ambiental, el modelo log-gaussiano incorporará poca variabilidad en su estimación. Por tanto, en este caso, le resultará difícil adaptarse a una única agregación de casos concreta. Este hecho redunda en la falta de potencia de este método ante agregaciones específicas, tal y como le ocurría al modelo de teselaciones y el modelo Poisson/Gamma. Además, al igual que en dichos modelos, las conclusiones del modelo Log-gaussiano no son tan potentes como las de los procesos clusters de Poisson, pues éstas respondían a cuestiones de gran interés para el estudio epidemiológico, como por ejemplo, la distribución del número de agregaciones que se pueden observar en el patrón.

A la vista de los modelos que acabamos de detallar podemos observar que tenemos varias posibilidades para definir dependencia entre las localizaciones de un proceso puntual. Así vemos que podemos definir dicha dependencia mediante una función escalonada, es decir, imponiendo que para cada punto del proceso, salvo un conjunto de medida nula, exista un entorno de dicho punto en el que la función de intensidad tomará el mismo valor en cada iteración del proceso. Por otra parte, también podemos definir dependencia entre localizaciones como consecuencia de la proximidad a ciertos "focos" o padres. En ese caso el proceso puntual viene definido en dos fases y la dependencia entre eventos proviene de la relación entre los mismos en la primera fase del proceso. También se puede definir relación espacial entre localizaciones haciendo depender la función de intensidad de la media local de un proceso estocástico, de esta forma localizaciones próximas tendrán medias locales similares y por tanto funciones de intensidad parecidas. Por último, se puede definir dependencia espacial entre localizaciones recurrien-

do a un proceso estocástico autoregresivo en el que localizaciones próximas guardan correlación entre sí. Así pues, se disponen de distintas formas de definir funciones de intensidad o de riesgo en la que exista dependencia entre localizaciones próximas. La elección de uno de estos modelos a la hora de describir un patrón puntual dependerá de distintos aspectos, como la información externa que se disponga sobre el mecanismo que ha generado el patrón o la bondad de ajuste de los distintos propuestas al patrón puntual estudiado.

Capítulo 3

Modelos de mixturas

Según se ha expuesto en el capítulo 2, cada una de las propuestas de modelización de la función de intensidad presenta distintas ventajas e inconvenientes. Concretamente, observamos que la modelización cluster de Poisson presentaba ciertas ventajas que el resto de propuestas no tenía. En dicho modelo se contempla específicamente la existencia de agregaciones individuales de casos alrededor de ciertas localizaciones específicas, mientras que en el resto de propuestas se ajusta la extravarianza de los datos mediante la modelización de la heterogeneidad medioambiental. Por tanto, la modelización cluster de Poisson se adapta de forma satisfactoria al tipo de situaciones que nos ocupa, la determinación de los distintos focos de riesgo en brotes epidémicos. Es por ello que nos vamos a detener en su estudio, así como en el de los modelos de mixturas, que constituyen una formulación alternativa de los procesos cluster de Poisson.

En los procesos puntuales que se han introducido resulta necesaria la utilización de modelos muy flexibles para describir las superficies de intensidad. Este hecho se debe a que la estructura de algunos datos es demasiado compleja como para poder ser descrita por un modelo paramétrico sencillo,

especialmente si dichos datos presentan ciertas características poco comunes, como por ejemplo multimodalidad, que no suelen contemplarse en las funciones de distribución habituales. Este es el motivo por el que se suele recurrir a métodos no paramétricos para la modelización de las superficies de intensidad en procesos puntuales. De esta forma se intenta garantizar la adecuación de la intensidad al patrón estudiado. Sin embargo, la modelización mediante mixturas se suele presentar como una alternativa a los modelos no paramétricos suficientemente flexible que puede llegar a describir la localización de los eventos de forma satisfactoria.

Dada una familia de funciones de distribución $f(\cdot|\theta)$, diremos que un conjunto de valores sigue una distribución de *mixtura finita de la familia f*, si para cada observación su función de distribución toma la forma:

$$p(x_i|w,\theta) = \sum_{j=1}^{m} w_j f(x_i|\theta_j) \ , \tag{3.1}$$

donde el vector w se dice vector de *pesos de la mixtura* y cumple que su suma vale 1. Este tipo de propuesta resulta muy apropiada en el caso de que dispongamos de varios grupos heterogéneos en nuestra población, ya que en ese caso cada grupo se modelizaría mediante una componente de la mixtura. Los pesos de la mixtura corresponden a la proporción de población que el modelo asigna a cada grupo de ésta.

Los modelos de mixturas de distribuciones suponen una extensión de la modelización mediante las familias de distribuciones habituales, que permite contemplar multimodalidad o la existencia de agrupaciones heterogéneas en los datos. No obstante, ésta no es la única aplicación de los modelos de mixturas ya que también se suelen utilizar para la estimación flexible de funciones de distribución dada su facilidad para adaptarse a los datos. Así, si bien las mixturas no pueden considerarse un método de estimación no parámétrico, suponen una alternativa a este tipo de métodos en ciertas

aplicaciones.

Existe una gran similitud entre la formulación del proceso de mixturas, ecuación (3.1) y la formulación del proceso cluster de Poisson como proceso de Cox, ecuación (2.4). De hecho, tal y como se demuestra en las páginas 36-38 de Castellos (1998) [26], dado el número de padres de un proceso cluster de Poisson, éste se puede considerar un caso particular de modelización mediante mixturas de distribuciones, en el que todas las componentes reciben el mismo peso y todas las componentes de la mixtura son iguales salvo su localización. Esta doble visión de los procesos cluster, tanto como procesos puntuales como procesos de mixturas, nos va a permitir aprovechar los resultados y ventajas de ambas aproximaciones. En concreto, la modelización de un proceso cluster de Poisson mediante un modelo de mixturas tiene ciertas ventajas, ya que permitirá de forma sencilla considerar agrupaciones con pesos y formas distintas.

La función de distribución f empleada en la definición de la mixtura puede ser tanto univariante como multivariante, concretamente el caso bidimensional resultará especialmente interesante para nosotros ya que será útil para definir la superficie de riesgo en problemas con componente geográfica. En la figura 3.1 se puede observar la función de intensidad resultante de un proceso cluster de Poisson con distribución homogénea de los padres, de intensidad ρ=6. En dicho proceso la distribución de los hijos alrededor de los padres sigue una distribución normal bivariante de matriz de varianzas-covarianzas

$$\begin{pmatrix} 0{,}01 & 0{,}001 \\ 0{,}001 & 0{,}01 \end{pmatrix}$$

y el número de hijos esperados para cada padre es de 10 eventos.

Por otra parte, en la figura 3.2 podemos observar un conjunto de datos

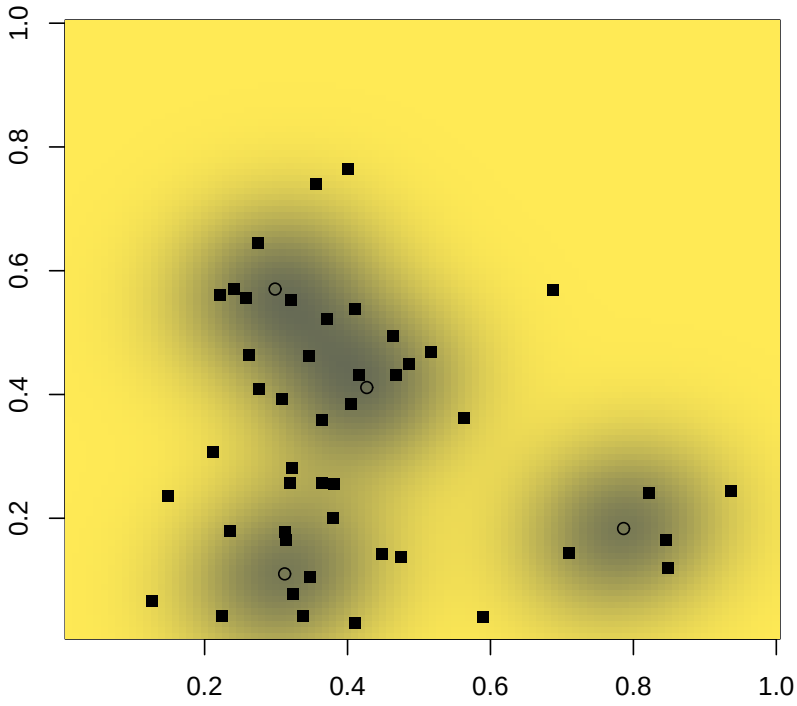

Figura 3.1: Proceso cluster de Poisson ρ=6, distribución normal bivariante de los hijos alrededor de los padres y 10 hijos esperados por cada padre.

generado a partir de un modelo de mixturas de distribuciones normales bivariantes de 6 componentes. El número de eventos generado es 60, los pesos de la mixtura siguen una distribución Dirichlet de parámetros (1,1,1,1,1,1), las medias siguen una distribución Uniforme sobre la región de estudio, $[0,1] \times [0,1]$. Por último, las matrices de varianza-covarianza de las distintas componentes siguen una distribución

$$\mathcal{W}\left(5, \frac{1}{5}\begin{pmatrix} 0{,}01 & 0{,}001 \\ 0{,}001 & 0{,}01 \end{pmatrix}\right) .$$

Así, aunque ambas formulaciones puedan generan procesos equivalentes, podemos observar algunas diferencias entre los patrones generados por ambas propuestas. En el proceso cluster el número de padres del proceso es una variable aleatoria, en concreto el patrón puntual que se ha utilizado como ilustración dispone únicamente de 4 padres (puntos huecos) mientras que el número esperado de éstos es 6. Sin embargo, en el modelo de mixturas dicho parámetro no es necesariamente variable. Así, en el ejemplo que se ha utilizado, el número de componentes se ha fijado a 6. Por otra parte, observamos que el modelo de mixturas tiene componentes de efecto casi inapreciable mientras que otras componentes aglutinan una gran proporción de los eventos del proceso. Por el contrario, en el proceso cluster todas las agrupaciones generan aproximadamente los mismos casos, 10 en promedio, aunque el número exacto de eventos por agrupación también es una cantidad variable. Este hecho se debe a la posibilidad que brinda el modelo de mixturas de incluir pesos diferentes en sus distintas componentes. Además, se puede observar que en la formulación mediante mixturas la orientación de las agrupaciones puede diferir para los distintos grupos, mientras que en la formulación como proceso puntual todas las componentes han de tener necesariamente la misma forma. Por último, el proceso puntual generará un número aleatorio de casos mientras que el modelo de mixturas se define

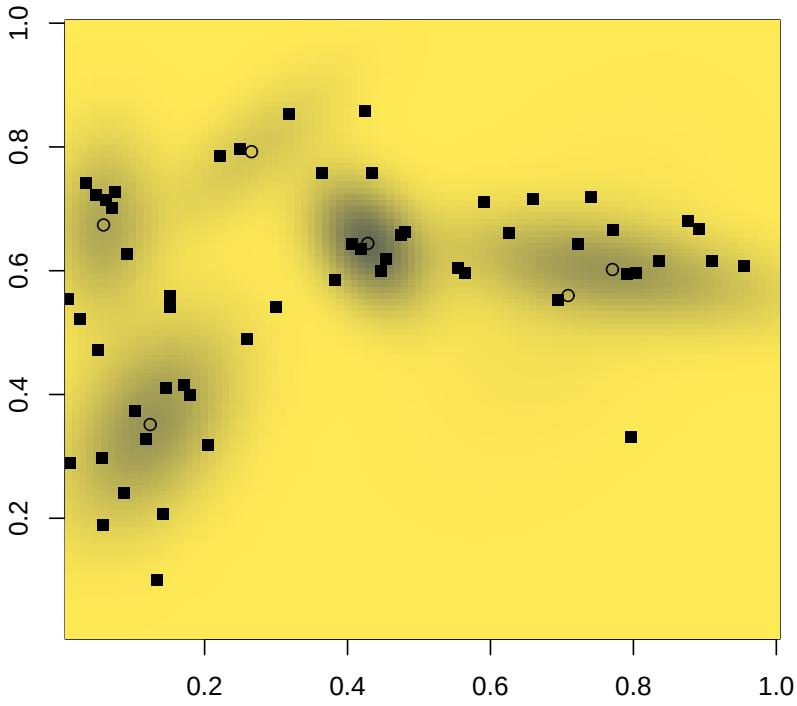

Figura 3.2: Modelo de mixturas de distribuciones normales bivariantes de 6 componentes.

condicionado al número de observaciones que se disponen. Este condicionamiento puede ser realista o no, dependiendo del problema en cuestión. Así, tanto una formulación como la otra presentan alguna ventaja concreta, por ello sería aconsejable proponer un planteamiento común de los dos modelos que unificara las ventajas de ambos.

La formulación de los modelos de mixturas resulta posible en el contexto tanto de la estadística clásica como el de la estadística bayesiana. En ambos marcos estos modelos gozan de una gran tradición, estudios teóricos e innumerables aplicaciones. Sin embargo, sea cual sea el marco estadístico en el que se decida llevar a cabo la inferencia sobre una distribución de tipo mixtura, existe una representación del modelo, haciendo uso de datos aumentados, que ha demostrado tener una gran utilidad. En dicha representación se hace uso, para cada observación x_i, de una variable auxiliar Z_i que indica la componente de la mixtura a la que pertenece la i-ésima observación. En ese caso la verosimilitud de los datos, condicionados al valor de las variables $\{Z_1, ..., Z_n\}$ toma el siguiente valor:

$$P(\{x_1, ..., x_n\}|\{Z_1, ..., Z_n\}, w, \theta) = \prod_{i=1}^{n} \prod_{j=1}^{m} w_j^{1(Z_i=j)} (f(x_i|\theta_j))^{1(Z_i=j)} \ , \quad (3.2)$$

donde la función $1(\cdot)$ valdrá 1 si la condición a la que hace referencia se cumple y 0 en otro caso. La ventaja de esta representación del modelo de mixturas es que la maximización de la función (3.2) resulta bastante más sencilla que la maximización de la función de verosimilitud original:

$$P(\{x_1, ..., x_n\}|w, \theta) = \prod_{i=1}^{n} \left(\sum_{j=1}^{m} w_j f(x_i|\theta_j) \right) \ ,$$

en la que el alto número de sumandos resultante (m^n) conlleva una expresión final de difícil manejo. La formulación del modelo de mixturas mediante el uso de datos aumentados se conoce como *formulación completa* de un modelo de mixtura.

Desde el enfoque de la estadística frecuentista, la inferencia en un modelo de mixturas se reduce a la maximización de (3.2) respecto a los parámetros w y θ. El problema que se presenta al maximizar dicha expresión es que en ella se desconocen los valores de los parámetros $\{Z_1, ..., Z_n\}$, por lo que se habrá de realizar la maximización marginalizando sus valores. Para ello suele ser habitual hacer uso del algoritmo EM (Dempster et al., 1977 [36]). En este procedimiento se procede de forma secuencial maximizando la función de verosimilitud respecto a w y θ dados unos valores de $\{Z_1, ..., Z_n\}$. Posteriormente se calculan los valores esperados de $\{Z_1, ..., Z_n\}$ dadas las estimaciones de w y θ obtenidas en la fase anterior. Este proceso se repite hasta que se alcanza la convergencia de los parámetros. Gracias al algoritmo EM resulta posible la inferencia de los modelos de mixturas desde el punto de vista clásico de una forma computacionalmente razonable.

La inferencia desde el enfoque bayesiano se realiza mediante los métodos propios de simulación en modelos jerárquicos. Concretamente, las mixturas desde este enfoque se plantean como modelos jerárquicos, ya que según se ha visto la función de verosimilitud depende de parámetros desconocidos $(\{Z_1, ..., Z_n\}, w, \theta)$ que se habrán de definir en sucesivas capas del modelo. La formulación habitual de una mixtura desde el enfoque bayesiano toma la siguiente forma:

$$P(x_i|Z_i, w, \theta) = f(x_i|\theta_{Z_i}) \; i = 1, ..., n \; ,$$

$$P(Z_i|w) = \mathrm{Mn}(w_1, ..., w_m) \; i = 1, ..., n \; ,$$

$$P(w) = ... \; , \; P(\theta) = ... \; .$$

donde, si se marginalizan los parámetros $\{Z_1, ..., Z_n\}$, resulta sencillo observar que la formulación de datos aumentados es equivalente a la formulación original del modelo de mixturas. También resulta posible plantear la formulación original del modelo de mixturas, añadiendo las distribuciones iniciales necesarias al resto de parámetros del modelo. No obstante,

dicha formulación resulta bastante más ineficiente computacionalmente y no aporta ninguna información sobre la probabilidad de que cada individuo concreto pertenezca a cierta componente específica de la mixtura, cuando esta información puede ser de interés desde un punto de visto epidemiológico.

En la formulación que se ha hecho hasta el momento de los modelos de mixturas se ha considerado el número de componentes m como un parámetro fijo del modelo. Sin embargo, en ocasiones éste será uno de los parámetros del modelo que mayor interés se tendrá en aprender. Así, en la detección de focos de riesgo en un brote epidémico este parámetro responderá a una de las cuestiones de mayor interés de análisis estadístico, ésta es: ¿cuantos focos han intervenido en la difusión del brote? Así, nos interesará disponer de un procedimiento estadístico que proponer alguna información sobre esta cuestión.

En el caso de que se considere el número de componentes de la mixtura como un valor fijo, no resultará obvio establecer un valor adecuado para este parámetro y los resultados que se derivarán de la mixtura dependerán en gran medida del valor que se establezca para m. Es por ello que la elección que se hace de este valor se ha de realizar con particular cuidado y resulta conveniente comparar el ajuste obtenido para distintos valores de este parámetro. Desde el enfoque de la estadística clásica se suele hacer uso de estadísticos de selección de modelos como el AIC (Akaike, 1973 [2]) o el BIC (Schwarz, 1978 [87]), entre otros, para la valoración de las distintas propuestas sobre el número de componentes de una mixtura. El estadístico DIC (Spiegelhalter et al., 2002 [90]) no suele ser apropiado para la valoración de modelos de mixturas, ya que no resulta aconsejable su aplicación en el caso de que la función de verosimilitud sea multimodal. No obstante se ha realizado alguna propuesta de adaptación de este estadístico para su

aplicación a modelos de mixturas, como por ejemplo Celeux et al. (2003) [27], que permite la utilización del estadístico DIC también en este contexto. En base a los valores obtenidos según estos criterios se suele determinar un estimador puntual del número de componentes de la mixtura que describe de forma más adecuada los datos analizados.

En Raftery (1993) [79] se describe un proceso para aproximar la probabilidad del número de componentes de una mixtura a la vista de los datos, en función del estadístico BIC. Esta estimación de la función de probabilidad de este valor nos permitirá promediar los resultados obtenidos para los modelos de mixturas con diferentes números de componentes. De esta forma resulta posible obtener estimaciones combinadas de todos los parámetros del modelo incorporando la incertidumbre sobre el número de componentes de la mixtura. Esta técnica se conoce como "*composite EM estimation*". Sin embargo, la estimación mediante composite EM consiste en el promedio de las estimaciones para distintos valores de m, según su probabilidad, pero no incorpora la variabilidad de las estimaciones para cada valor de m. Es por ello que bajo esta aproximación se obtendrán estimaciones de la superficie de intensidad más precisas de lo que cabría esperar.

Tal y como hemos podido comprobar, resulta posible desde un enfoque clásico estimar la superficie de intensidad en modelos de mixturas que incorporen la incertidumbre sobre el número de parámetros del modelo. Sin embargo, también resulta posible un planteamiento completamente bayesiano de estos modelos (Richardson y Green, 1997 [80]) en el que el número de componentes de la mixtura se incluye como una variable más en la formulación. En ese caso, la inferencia sobre el número de componentes de la mixtura se realiza de forma más natural, como cualquier otro parámetro del problema, sin tener que recurrir al ajuste de distintos modelos y una posterior puesta en común de todos los resultados. Para la estimación bayesiana

de los modelos de mixturas con un número de componentes desconocido se hará uso de métodos de simulación de modelos jerárquicos que sean capaces de incorporar incertidumbre sobre el número de parámetros del problema, por ejemplo la *simulación de salto reversible* descrita en Green (1995) [52]. En Castelloe (1998) [26] se hace uso de ambos tipos de estimación, composite EM y simulación de salto reversible, sobre un modelo de mixturas y se comparan los resultados obtenidos en un gran número de bancos de datos simulados. Castelloe concluye en su estudio que, desde su punto de vista, resulta preferible la formulación bayesiana de los procesos cluster de Poisson, ya que en la inferencia mediante métodos EM no resulta sencillo incorporar modificaciones al modelo de mixturas. Además, según se señala en Stephens (1999) [91] la estimación de mixturas desde un punto de vista clásico presenta ciertos problemas debido a que para un gran número de familias paramétricas la función de verosimilitud no está acotada. Así, por ejemplo, en el caso de una mixtura de distribuciones gaussianas la verosimilitud viene dada por

$$L(y|w,\mu,\sigma) = \prod_{j=1}^{n} \left(w_1 \mathcal{N}(y_j|\mu_1,\sigma_1^2) + ... + w_m \mathcal{N}(y_j|\mu_m,\sigma_m^2) \right) ,$$

pero dicha función es no acotada para $x_i = \mu_j$ para cualquier i, j conforme $\sigma_j \to 0$. Por tanto, la función de verosimilitud no sólo es no acotada, sino que tiene múltiples máximos locales. Este problema suele ser resuelto mediante la imposición de igualdad entre todas las varianzas o la imposición de una cota inferior para las varianzas de las componentes. De todas formas, dichas condiciones no siempre resultan demasiado apropiadas en el contexto del problema estudiado.

Nuestra impresión es similar a la de Castelloe, es más, creemos que el enmarcar la modelización mediante mixturas dentro del contexto de los modelos jerárquicos bayesianos aporta muchas ventajas, tales como la modificación del modelo de una forma sencilla o la utilización de los métodos de

simulación y validación de este tipo de propuestas. Además, nuestra opinión es que la inferencia sobre el número de componentes de la mixtura se ha de realizar considerando este parámetro como variable y no de otra forma que trate de suplir la consideración de este valor como constante. De todas formas, la aproximación bayesiana al modelo de mixturas no está exenta de problemas, tal y como quedará patente en el resto de esta tesis. Aun así, consideramos más apropiada la modelización bayesiana del modelo de mixturas en nuestro problema y a partir de este momento nos vamos a centrar en su desarrollo.

3.1. Modelización bayesiana de mixturas gaussianas

La modelización de procesos puntuales mediante mixturas de funciones de densidad gaussianas bidimensionales con pesos constantes es, según se ha comentado, uno de los procesos cluster de Poisson de mayor tradición. El modelo de mixturas gaussiano resulta apropiado para modelizar la difusión de una enfermedad alrededor de varios focos de emisión. Este hecho se debe a que dicha distribución decrece como función de la distancia a su media, por lo que ésta puede ser utilizada para describir el decrecimiento de la función de riesgo conforme aumenta la distancia a uno de los focos del brote. Además, la distribución gaussiana permite, mediante la modelización de la matriz de varianzas-covarianzas, inducir de forma sencilla direcciones en las que el riesgo se difunda con mayor facilidad. Obviamente, también se pueden utilizar otras funciones de distribución para modelizar un proceso cluster de Poisson. En concreto, en Stephens (1999) [91] se emplean distribuciones t multivariantes con pocos grados de libertad como alternativa robusta a la utilización de distribuciones gaussianas. Sin embargo, en el desarrollo

del presente trabajo nos centraremos en el estudio de mixturas gaussianas, aunque se podría hacer un estudio similar empleando cualquier otra función de densidad.

Comenzaremos introduciendo el modelo de mixturas gaussiano para el caso univariante. Obviamente, la formulación que vamos a proponer no es única y sobre ésta se pueden realizar modificaciones, sobre todo respecto a la distribución previa inicial de sus parámetros. El siguiente modelo hace uso de la formulación completa, ya que tal y como se ha comentado ésta tiene ciertas ventajas computacionales:

$$\begin{aligned} y_i|z_i,\mu,\sigma &\sim \mathcal{N}(y_i|\mu_{z_i},\sigma^2_{z_i}) \;\; i=1,...,n \\ z_i|w &\sim \mathrm{Mn}(w_1,...,w_m) \;\; i=1,...,n \\ \mu_j &\sim \mathcal{N}(\xi,\kappa^{-1}) \;\; j=1,...,m \\ \sigma_j^{-2} &\sim \Gamma(\alpha,\beta) \;\; j=1,...,m \\ w &\sim \mathrm{Dir}(\delta,...,\delta)\ . \end{aligned} \tag{3.3}$$

Como se puede observar, dada la componente a la que pertenece cada observación, z_i, ésta se distribuye según una distribución normal con los parámetros propios de dicha componente. Por otro lado, la componente a la que pertenece cada observación se distribuye mediante una distribución multinomial, cuyos parámetros coinciden con los pesos de la mixtura. Dichos pesos siguen una distribución Dirichlet en la que todos los parámetros toman el mismo valor, δ. De esta forma, todos los pesos tienen a priori la misma distribución marginal. Por último, para las medias de las componentes de la mixtura se emplea una distribución normal que cubra ampliamente su rango de valores razonable, mientras que para las precisiones se emplea una distribución Gamma. La especificación de los hiperpárametros en el modelo presentado en (3.3) dependerá en gran parte de los estudios y el autor que lo haya empleado.

La formulación anterior se puede considerar una mixtura estándar de distribuciones normales a la que se podrán realizar modificaciones. Sin embargo, supone un buen punto de partida para otros modelos de mixturas, ya que la inferencia para éste resulta muy sencilla. Ello se debe a que en la formulación empleada todas las distribuciones iniciales, dado el resto de variables, son conjugadas de su función de verosimilitud. Por tanto la simulación de la distribución posterior de las variables del modelo resulta muy sencilla. En concreto, se puede hacer uso de Gibbs Sampling (Geman y Geman, 1984 [49]) para la generación de todos los valores. En Stephens (1999) [91] o Richardson y Green (1997) [80] se describen las funciones necesarias para muestrear la distribución posterior de los parámetros, éstas son:

$$\begin{aligned} p(z_i = j|...) &\propto w_j \mathcal{N}(y_i|\mu_j, \sigma_j^2) \;\; j = 1, ..., m \;\; i = 1, ..., n \\ w &\sim \text{Dir}(\delta + n_1, ...\delta + n_m) \\ \mu_j &\sim \mathcal{N}(\tfrac{\sigma_j^{-2} \sum_{\{i:z_i=j\}} y_i + \kappa\xi}{\sigma_j^{-2} n_j + \kappa}, (\sigma_j^{-2} n_j + \kappa)^{-1}) \;\; j = 1, ..., m \\ \sigma_j^{-2} &\sim \Gamma(\alpha + \tfrac{1}{2} n_j, \beta + \tfrac{1}{2} \textstyle\sum_{\{i:z_j=i\}} (y_i - \mu_j)^2) \;\; j = 1, ..., m \; . \end{aligned} \tag{3.4}$$

Por tanto, aunque señalábamos que la inferencia en un modelo de mixturas podía ser complicada debido a la forma de su función de su verosimilitud, resulta posible muestrear de forma sencilla los parámetros del modelo haciendo uso únicamente de generadores de valores aleatorios de las distribuciones Normal, Gamma, Dirichlet y Multinomial.

Entre las distintas modificaciones que se pueden hacer del modelo descrito en (3.3), tiene un particular interés el hecho de considerar como desconocido el número de componentes de la mixtura. Tal y como se ha comentado, ésta es una de las principales razones por las que nos hemos decantado por el planteamiento bayesiano para el presente trabajo. Richardson y Green (1997) [80] supone la referencia seminal en la consideración de modelos de mixturas con un número de componentes desconocidas. En dicho artículo se extiende la propuesta de modelo de mixturas que se acaba de definir

considerando tanto β como m variables aleatorias con distribuciones

$$\begin{aligned} \beta &\sim \Gamma(g, h) \\ m &\sim \mathcal{U}(1, M) \ . \end{aligned} \tag{3.5}$$

En adelante nos referiremos al modelo jerárquico definido por las expresiones (3.3) y (3.5) como el modelo de mixturas con un número de componentes desconocidas de Richardson y Green (RG). En dicha propuesta, la consideración que se realiza de β como variable del modelo va a permitir considerar precisiones distintas en las componentes de las mixturas, al tiempo que se penalizan grandes diferencias entre las distintas precisiones. De esta forma, se evitan los problemas relativos a los máximos locales de la función de verosimilitud que se planteaban en la versión no jerárquica del modelo de mixturas manteniendo la variabilidad en cuanto a las distintas precisiones de la mixtura. Precisaremos con más detalle esta cuestión al introducir el valor de los hiperparámetros utilizados en el modelo RG. Respecto a la distribución previa inicial de m, se ha adoptado una distribución Uniforme discreta entre 1 y un valor umbral máximo que podría tomarse como el número de observaciones o incluso algún valor menor que se considere razonable y que no coarte los resultados del modelo. Otros autores han tomado otras distribuciones iniciales para este parámetro, como por ejemplo una distribución de Poisson, más tarde se discutirá estas otras propuestas utilizadas.

Respecto a la elección de los hiperparámetros del modelo, ξ se establece como la media de los datos, mientras que para κ^{-1} se toma la longitud de su rango al cuadrado, R^2. De esta forma, se adopta una distribución inicial para las medias de las componentes que cubre sobradamente el rango de valores razonables para este valor, sin recurrir a distribuciones iniciales extremadamente vagas que darían verosimilitudes considerables a valores disparatados. En el caso de recurrir a distribuciones iniciales demasiado

vagas, las componentes de la mixtura que no tengan asociadas ninguna observación pueden tomar valores poco razonables y alterar la convergencia del mecanismo de simulación.

En cuanto a los hiperparámetros de las precisiones de las componentes de la mixtura se ha señalado que β será considerada variable, sin embargo, para la elección de α habremos de tener en cuenta que queremos penalizar las diferencias excesivas entre las distintas precisiones del modelo. En el caso de elegir un valor de $\alpha = 1$ se tendría:

$$E[\sigma_i^{-2}] = \frac{\alpha}{\beta} = \frac{1}{\beta}, \quad var[\sigma_i^{-2}] = \frac{\alpha}{\beta^2} = \frac{1}{\beta^2} \quad i = 1, ..., n \ ,$$

es decir, la media y la desviación típica de las precisiones coinciden, lo cual puede resultar poco restrictivo y podría resultar más apropiado penalizar en mayor medida las diferencias entre los valores de las precisiones. Teniendo en cuenta que la varianza y la media de las precisiones siguen la siguiente relación:

$$var[\sigma_i^{-2}] = \frac{\alpha}{\beta^2} = \frac{1}{\alpha}\left(\frac{\alpha}{\beta}\right)^2 = \frac{1}{\alpha}(E[\sigma_i^{-2}])^2 \ ,$$

si se aumenta el valor de α, se disminuirá la variabilidad de las precisiones en relación a su media. En concreto, en el modelo RG se propone un valor de $\alpha = 2$. Sin embargo, como no se quiere ser restrictivo en cuanto al valor promedio que tomarán las precisiones se toma una distribución vaga para la variable β. Para ello, se toma un valor para g de 0.2 siguiendo un razonamiento análogo al que se acaba de emplear para la determinación del valor de α. Respecto a h se propone como valor

$$h = \frac{100g}{\alpha R^2} \ ,$$

que refleja una media a priori para β de

$$\frac{\alpha R^2}{100} \ ,$$

lo cual se traduce, dado el valor esperado de β en una media a priori para las precisiones de $\frac{100}{R^2}$, es decir un valor esperado para las desviaciones típicas de las componentes de $\frac{R}{10}$ que puede resultar razonable teniendo en cuenta la gran variabilidad que se ha inducido en la distribución de β. Notar que se podría haber optado por una distribución menos informativa para las precisiones de las componentes de la mixtura, pero en ese caso se tendrían problemas con los máximos locales del modelo al permitirse la utilización de precisiones grandes sin ninguna penalización.

Por último, en cuanto al valor del parámetro δ, que define la distribución de los pesos de la mixtura, se tiene que su distribución marginal en función de delta sigue una distribución $\mathcal{B}(\delta, (m-1)\delta)$, por lo que la media y varianza de cada uno de los pesos vienen dadas por las siguientes expresiones:

$$E[w_i] = \frac{1}{m}, \quad var[w_i] = \frac{(m-1)}{m^2(m\delta+1)}, \quad i = 1, ..., m \ .$$

Así pues, valores grandes de δ se corresponderán con varianzas pequeñas en la distribución de los pesos, por lo que los pesos de la mixtura serán todos similares y, en consecuencia, las proporciones de observaciones correspondientes a cada componente de la mixtura. Por el contrario, valores pequeños de δ favorecerán componentes de la mixtura de tamaño muy distinto. En el modelo *RG*, para cada valor de m se ha adoptado un valor de $\delta = 1$, que corresponde a una distribución uniforme sobre el rango de valores posible del vector w.

La generalización al caso bidimensional del modelo de Richardson y Green resulta bastante directa. Stephens (1999) [91] realiza dicha generalización mediante la siguiente propuesta

$$
\begin{aligned}
y_i|z_i, \mu, \sigma &\sim \mathcal{N}_2(y_i|\mu_{z_i}, \Sigma^2_{z_i}) \;\; i = 1, ..., n \\
z_i|w &\sim \text{Mn}(w_1, ..., w_m) \;\; i = 1, ..., n \\
\mu_j &\sim \mathcal{N}_2(\xi, \kappa^{-1}) \;\; j = 1, ..., m \\
(\Sigma^2_j)^{-1}|\beta &\sim \mathcal{W}(2\alpha, (2\beta)^{-1}) \;\; j = 1, ..., m \\
\beta &\sim \mathcal{W}(2g, (2h)^{-1}) \\
w &\sim \text{Dir}(\delta, ..., \delta) \\
m &\sim \mathcal{U}(1, M) \; .
\end{aligned}
\tag{3.6}
$$

Para llevar a cabo esta generalización únicamente se ha hecho uso de las versiones multivariantes de las distribuciones del caso unidimensional. Concretamente se han sustituido las distribuciones normales univariantes por distribuciones normales bivariantes, denotadas como $\mathcal{N}_2$. Mientras, las distribuciones Gamma han sido sustituidas por distribuciones Wishart, la generalización multivariante de esta distribución. El valor esperado de una distribución $\mathcal{W}(\nu, S)$ es $\nu \cdot S$, donde ν es un escalar que controla la dispersión de la distribución, y valores pequeños de ν se corresponden con una distribución muy dispersa, mientras que S es una matriz que define la forma esperada a priori de la matriz de varianzas-covarianzas.

Respecto a los hiperparámetros del modelo bidimensional de mixturas propuesto por Stephens, ξ y κ se eligen de la misma forma que en el caso unidimensional, en base al punto medio de los observaciones y su rango. δ se toma igual a 1, es decir una distribución Uniforme sobre los valores posibles de los pesos. Respecto a α y g, Stephens propone tomar unos valores ligeramente superiores al caso unidimensional, 3 y 0.3 respectivamente. De esta forma se trata de ser algo más informativo a la hora de dar la distribución inicial de las precisiones, a la vez que también se penalizan en mayor medida

los óptimos locales del modelo de mixturas. h se define de la siguiente forma

$$h = \begin{pmatrix} \frac{100g}{\alpha R_1^2} & 0 \\ 0 & \frac{100g}{\alpha R_2^2} \end{pmatrix} ,$$

donde R_1, R_2 coinciden con el rango de los datos en su primera y segunda dimensión. De esta forma se han seguido las mismas especificaciones que para el modelo RG para el modelo unidimensional.

Respecto a la inferencia del modelo, en el caso que el valor de m fuera conocido resultaría posible la simulación de todos los parámetros del modelo mediante Gibbs Sampling. Stephens describe las distribuciones de cada parámetro condicionado al resto, con las que podremos muestrear las variables que componen el modelo de mixturas bidimensional. Dichas expresiones son:

$$\begin{gathered} p(z_i = j|...) \propto w_j \mathcal{N}_2(y_i|\mu_j, \Sigma_j^2) \;\; j = 1, ..., m, \;\; i = 1, ..., n \\ w \sim \mathrm{Dir}(\delta + n_1, ...\delta + n_m) \\ \mu_j \sim \mathcal{N}_2\left((\Sigma_j^{-2} n_j + \kappa)^{-1}\left(\Sigma_j^{-2}\left(\textstyle\sum_{i:z_i=j} y_i\right) + \kappa\xi\right), (\Sigma_j^{-2} n_j + \kappa)^{-1}\right) \;\; j = 1, ..., m \\ \Sigma_j^{-2} \sim \mathcal{W}\left(2\alpha + n_j, \left(2\beta + \textstyle\sum_{i:z_j=i}(y_i - \mu_j)(y_i - \mu_j)^t\right)^{-1}\right) \;\; j = 1, ..., m \\ \beta \sim \mathcal{W}\left(2g + 2m\alpha, \left(2h + 2\textstyle\sum_j \Sigma_j^{-2}\right)^{-1}\right) . \end{gathered} \tag{3.7}$$

Sin embargo, en el modelo propuesto se desconoce el valor de m, por lo que el número de variables a simular también es desconocido. En este caso, hemos de emplear un método de inferencia que sea capaz de simular modelos con un número indeterminado de parámetros. Dichos métodos de simulación van a ser tratados en la siguiente sección.

3.2. Simulación trans-dimensional

En estadística suelen ser habituales los "problemas en los que el número de cosas que se desconocen es una de las cosas que se desconocen". Este tipo de situaciones se presentará cuando la incertidumbre ante un problema de inferencia no se ciña únicamente a la estimación de los parámetros de un modelo, sino que dicha incertidumbre afecte también a la elección del modelo apropiado para describir los datos bajo estudio. Ejemplos habituales de este tipo de situaciones serían los problemas de selección de variables en un modelo de regresión, la determinación de la estructura de una serie temporal o la elección del número de componentes en un modelo de mixturas. En cualquiera de los problemas anteriores la elección de un modelo u otro conllevará la estimación de un número diferente de parámetros. Por otra parte, en la mayoría de estas situaciones el conocimiento previo del problema impedirá que se opte por un modelo concreto con ciertas garantías. Además, la elección de un modelo concreto de entre todas las posibilidades puede coartar las conclusiones del estudio e incluso conducir a resultados que no describan apropiadamente los datos. Por tanto, en este tipo de situaciones resulta de gran valor los métodos de inferencia "trans-dimensionales" que permiten la simulación conjunta de distintos modelos estadísticos candidatos a describir los datos.

A diferencia de los criterios de selección de modelos, que determinan el modelo estadístico que mejor se adapta al conjunto de datos observado, la simulación trans-dimensional se moverá entre dichos modelos permaneciendo más tiempo en aquellos que son considerados más probables. Por tanto, los métodos de simulación trans-dimensional proporcionarán la distribución posterior de los modelos propuestos, a diferencia de los métodos de selección de modelos que proporcionan un único estimador puntual del modelo con-

siderado como más probable, ignorando la variabilidad existente en dicha estimación.

La técnica de simulación trans-dimensional más extendida hasta la fecha es la simulación mediante *Métodos de Monte Carlo basados en Cadenas de Markov de salto reversible* (RJMCMC), introducida por primera vez en Green (1995) [52]. También se puede encontrar un desarrollo más pedagógico de este técnica de simulación en Waagepetersen y Sorensen (2001) [97].

La simulación de salto reversible se basa en la determinación de un mecanismo de generación de valores que cumpla la condición de *detailed balance*:

$$\int_{(x,x')\in A\times B} \pi(dx)P(x,dx') = \int_{(x,x')\in A\times B} \pi(dx')P(x',dx) \ , \qquad (3.8)$$

donde π es la distribución que se está interesado en muestrear, mientras que $P(x_1, x_2)$ es la probabilidad de pasar de x_1 a x_2 aplicando el mecanismo de simulación propuesto. Esta expresión establece una condición suficiente para que el mecanismo de generación de valores produzca una muestra de la distribución que se desea muestrear, π. La simulación mediante el método de Metropolis-Hastings, uno de los mecanismos de simulación más habituales, también se basa en la determinación de una probabilidad de aceptación en cada movimiento de la cadena de Markov, de forma que la simulación cumpla la condición anterior.

La simulación de salto reversible proporciona una muestra de valores de entre un conjunto de modelos estadísticos, en base a la probabilidad de cada uno de éstos. Es decir, la cadena de Markov generada por un algoritmo de salto reversible se moverá entre los distintos modelos que se han considerado candidatos a describir los datos, de forma que la cadena permanecerá más tiempo en aquellos modelos que mejor los describan. Los

algoritmos RJMCMC suelen constar de distintos tipos de movimientos que cambian el estado de la cadena de Markov de un modelo a otro, llamaremos estos movimientos *trans-dimensionales*. Además, la cadena de Markov tendrá movimientos que se realizan interiormente en cada modelo visitado por el proceso de simulación, estos movimientos se dicen intra-dimensionales y se llevan a cabo mediante los métodos de simulación MCMC habituales, como por ejemplo Gibbs Sampling o el algoritmo de Metropolis-Hastings. Es en los movimientos de tipo trans-dimensional donde la simulación de salto reversible realiza su contribución. Ésta se basa en la definición de pares de movimientos opuestos M_{ab} y M_{ba} entre distintos modelos Θ_a, Θ_b, posiblemente de dimensiones distintas. Denotaremos por θ_a el vector de parámetros, de dimensión $d(\theta_a)$, del modelo Θ_a y por θ_b los parámetros de Θ_b, de dimensión $d(\theta_b)$. Para el movimiento entre los modelos Θ_a y Θ_b, se hará uso de un vector aleatorio u_a, de dimensión $d(u_a)$, u_b de dimensión $d(u_b)$ y de las aplicaciones biyectivas y diferenciables:

$$\begin{array}{rccc} M_{ab}: & \Theta_a \times U_a & \rightarrow & \Theta_b \times U_b \\ & (\theta_a, u_a) & \rightarrow & (\theta_b, u_b) \end{array} \qquad \begin{array}{rccc} M_{ba}: & \Theta_b \times U_b & \rightarrow & \Theta_a \times U_a \\ & (\theta_b, u_b) & \rightarrow & (\theta_a, u_a) \end{array} ,$$

donde M_{ab} es la aplicación inversa de M_{ba}. Para que esto sea posible se habrá de cumplir $d(\theta_a)+d(u_a) = d(\theta_b)+d(u_b)$. Por tanto, para que la cadena se mueva entre los modelos Θ_a y Θ_b se habrá de generar un vector aleatorio u_a y la aplicación M_{ab} transformará el conjunto de valores actual, θ_a, junto con el conjunto de valores aleatorios generados, para obtener un valor, θ_b, del nuevo modelo. Una vez se ha generado el nuevo valor candidato a integrar la cadena de Markov se somete a un proceso de aceptación-rechazo. Así, en el caso que se quiera realizar un movimiento del modelo Θ_a al modelo Θ_b la nueva propuesta se aceptará con probabilidad:

$$\alpha(\theta_a, \theta_b) = min\left(1, \frac{\pi(\theta_b|x) j_{M_{ba}}(\theta_b) g(u_b)}{\pi(\theta_a|x) j_{M_{ab}}(\theta_a) g(u_a)} \left|\frac{\partial(\theta_b, u_b)}{\partial(\theta_a, u_a)}\right|\right) ,$$

donde $\pi(\cdot|x)$ será el valor en la distribución posterior que se intenta mues-

trear, $g(\cdot)$ es la distribución de probabilidad del vector aleatorio necesario para el salto entre los distintos modelos y $j_{M_{ba}}(\theta_b)$ es la probabilidad de que en la cadena de Markov se proponga el movimiento M_{ba} cuando la cadena se halla en θ_b. Conviene destacar que la existencia del Jacobiano de la transformación entre (θ_a, u_a) y (θ_b, u_b) viene garantizada por la diferenciabilidad de la aplicación M_{ab}, y que la biyectividad de dicha aplicación garantiza que el jacobiano de dicha transformación será generalmente distinto de 0.

En el caso de que se proponga un movimiento del modelo Θ_b a Θ_a la probabilidad de aceptación vendrá dado por

$$\alpha(\theta_b, \theta_a) = min\left(1, \frac{\pi(\theta_a|x) j_{M_{ab}}(\theta_a) g(u_a)}{\pi(\theta_b|x) j_{M_{ba}}(\theta_b) g(u_b)} \left|\frac{\partial(\theta_a, u_a)}{\partial(\theta_b, u_b)}\right|\right) ,$$

donde en este caso la existencia del jacobiano nuevamente viene garantizada por la diferenciabilidad de la aplicación M_{ba}. Además como M_{ab} es la aplicación inversa de M_{ba} resulta:

$$\left|\frac{\partial(\theta_a, u_a)}{\partial(\theta_b, u_b)}\right| = \left|\frac{\partial(\theta_b, u_b)}{\partial(\theta_a, u_a)}\right|^{-1} ,$$

por lo que

$$\alpha(\theta_b, \theta_a) = min\left(1, \frac{1}{\frac{\pi(\theta_b|x) j_{M_a}(\theta_b) g(u_b)}{\pi(\theta_a|x) j_{M_b}(\theta_a) g(u_a)} \left|\frac{\partial(\theta_a, u_a)}{\partial(\theta_b, u_b)}\right|^{-1}}\right) = min\left(1, \alpha(\theta_a, \theta_b)^{-1}\right) .$$

De esta manera, la probabilidad de aceptación de un paso viene determinada por la probabilidad de aceptación del paso inverso del mecanismo trans-dimensional.

La reiteración sucesiva de los movimientos M_{ab} y M_{ba} junto con los movimientos intra-dimensionales que se consideren oportunos generará una

cadena de Markov que varía entre los distintos modelos propuestos tal y como se pretendía.

No obstante, la simulación de salto reversible no es el único método de simulación trans-dimensional propuesto hasta la fecha. Así, en Greenander y Miller (1994) [54] se propone una metodología de simulación que ellos mismos denominan *jump diffusion*. Dicha metodología se basa en la simulación de un proceso continuo en el tiempo en el que se producen saltos entre modelos en instantes aleatorios. Entre salto y salto trans-dimensional la simulación efectúa distintas simulaciones intra-dimensionales mediante un proceso de Langevin (Besag, 1994 [13]; Roberts y Tweedie, 1996 [84]). Posteriormente ahondaremos en este tipo de simulación. Esta formulación ha tenido un éxito bastante más reducido al de la simulación de salto reversible, seguramente debido a su complejidad. No obstante, Besag (1994) [13] demuestra que, salvo una discretización con fines prácticos que se realiza en el algoritmo de *jump diffusion*, éste se puede considerar como una formulación particular de la simulación de salto reversible.

En Stephens (2000) [92] se realiza otra propuesta de simulación trans-dimensional basada en un proceso de nacimiento-muerte sugerido anteriormente en Ripley (1977) [82]. Éste se basa también en la generación de un proceso en tiempo continuo, en el que se añadirán (nacimiento) o eliminirán (muerte) alguna de las componentes del modelo actual, en instantes aleatorios que dependerán de la verosimilitud del estado de la simulación en cada momento. Así, cuando el proceso se encuentre en un estado de gran verosimilitud la cadena tardará un gran tiempo en saltar a otro modelo, mientras que cuando el proceso visite un estado poco compatible con los datos, éste será abandonado rápidamente. La cadena de Markov de la simulación se consigue mediante una discretización del proceso continuo que se ha descrito, del que se extraerá su valor cada t unidades de tiempo. Notar que dada

la formulación de los algoritmos de nacimiento-muerte, éstos resultan especialmente indicados en el caso que los modelos disponibles sean anidados, como es el caso de los modelos de mixturas con un número de componentes desconocidas. Cappé et al. (2003) [24] desarrollan un completo análisis de la relación entre la simulación de salto reversible y la de nacimiento-muerte, aplicando sus resultados también a la simulación de modelos de mixturas.

Como alternativa a los métodos que se han descrito, que muestrean valores sobre la unión de distintos modelos, en Carlin y Chib (1995) [25] se muestrean conjuntamente valores de todos los modelos propuestos y una vez muestreados se elige uno de estos valores en base a su verosimilitud. Obviamente, esta alternativa resulta poco eficiente computacionalmente al muestrear en cada paso valores de todos los modelos posibles. Sin embargo, existen distintas variantes de este método que mejoran este aspecto reutilizando los valores muestreados y que no han sido incluidas en la cadena de Markov (Dellaportas, 2002 [34]). Además, este tipo de simulación tiene una limitación natural sobre el número de modelos considerado ya que éste habrá de ser necesariamente finito para poder ser simulado. Por último, esta propuesta hace uso de lo que en su día llamaron los autores *pseudo-prior*. Si θ es el conjunto de parámetros de todos los modelos y θ_{-k} son todos los parámetros de θ salvo los del modelo k, la pseudo-prior se define como $P(\theta_{-k}|\theta_k)$. Esta distribución se toma de forma completamente arbitraria ya que no influye sobre la distribución posterior obtenida, sin embargo dicha elección condiciona la eficiencia y el comportamiento de la cadena de Markov simulada. Por tanto la determinación de la pseudo-prior supone un problema adicional para la aplicación de esta metodología. Godsill (2000) [51] unifica esta última propuesta y la simulación de salto reversible en un único marco teórico común. De esta forma, resultará posible aunar las ventajas de ambos tipos de simulación en base a este planteamiento.

En Greean (2003) [53] y Sisson (2004) [88] se revisa el estado actual de las técnicas de simulación trans-dimensional. En dichos trabajos se pueden encontrar mas detalles sobre los métodos de simulación aquí descritos, así como las posibles perspectivas abiertas actualmente en este campo.

3.2.1. Simulación del modelo de mixturas unidimensional

Tras introducir los modelos de mixturas con un número desconocido de componentes y los métodos de simulación trans-dimensionales, vamos a describir como aplicar estos algoritmos para llevar a cabo la inferencia en dichos modelos. Concretamente, se hará uso de un algoritmo de salto reversible que es la propuesta de simulación que se utilizará posteriormente para abordar el problema de la detección de focos en brotes epidémicos. El algoritmo de salto reversible que se va a describir corresponde a la propuesta realizada en Richardson y Green (1997) [80]. No obstante, se ha considerado interesante incluir su desarrollo en el presente texto para facilitar la comprensión de la sección anterior y el resto del trabajo.

Recordamos que el modelo de mixturas *RG* unidimensional, en su formulación completa, se definía de la siguiente forma

$$\begin{aligned} y_i|z_i,\mu,\sigma &\sim \mathcal{N}(y_i|\mu_{z_i},\sigma^2_{z_i}) \;\; i=1,...,n \\ z_i|w &\sim \mathrm{Mn}(w_1,...,w_m) \;\; i=1,...,n \\ \mu_j &\sim \mathcal{N}(\xi,\kappa^{-1}) \;\; j=1,...,m \\ \sigma_j^{-2} &\sim \Gamma(\alpha,\beta) \;\; j=1,...,m \\ w &\sim \mathrm{Dir}(\delta,...,\delta) \\ \beta &\sim \Gamma(g,h) \\ m &\sim \mathcal{U}(1,M)\ , \end{aligned} \tag{3.9}$$

donde $\delta, \xi, \kappa, M, g, h$ serán los hiperparámetros del modelo.

Tal y como se ha comentado los algoritmos de salto reversible constan de movimientos intra y trans-dimensionales. En concreto, para la simulación del modelo (3.9) se hará uso de dos pares de movimientos trans-dimensionales, nacimiento-muerte y división-combinación, y 5 movimientos de tipo intra-dimensional en los que se actualizarán los distintos paramétros de los modelos.

Los movimientos de tipo intra-dimensional consisten en el muestreo de los parámetros $\{\beta, \mu_j, \sigma_j^{-2}, w, z_i : i = 1, ..., n \quad j = 1, ..., m\}$. Como en este tipo de movimiento no cambiamos de modelo, las simulaciones se pueden realizar mediante el muestreo de Gibbs descrito en la ecuación (3.4).

Para la realización del proceso de división-combinación en primer lugar se realizará una elección aleatoria entre los movimientos de división y combinación según las probabilidades d_k y $c_k = 1 - d_k$, dependientes del número de componentes en la mixtura, k. Suele ser habitual tomar $d_k = b_k = 0{,}5$ para $1 < k < M$, $d_M = b_1 = 1$ y $d_1 = b_M = 0$. El proceso de combinación procede eligiendo de forma aleatoria 2 componentes de la mixtura y combinándolas en una única componente que habrá de cumplir las siguientes condiciones

$$\begin{aligned} w_{j^*} &= w_{j1} + w_{j2} \ , \\ w_{j^*}\mu_{j^*} &= w_{j1}\mu_{j1} + w_{j2}\mu_{j2}, \\ w_{j^*}(\mu_{j^*}^2 + \sigma_{j^*}^2) &= w_{j1}(\mu_{j1}^2 + \sigma_{j1}^2) + w_{j2}(\mu_{j2}^2 + \sigma_{j2}^2) \ , \end{aligned} \tag{3.10}$$

donde los subíndices $j1$ y $j2$ corresponden a las componentes de la mixtura que se van a combinar, mientras que el subíndice j^* alude a la componente resultado de dicha combinación. El cumplimiento de las condiciones anteriores garantiza que los momentos de orden 0, 1 y 2 de la nueva componente y la suma de dichos momentos en las componentes originales coinciden. De esta forma, se intenta que las propuestas generadas tras la combinación tengan

una probabilidad de aceptación razonable y así la cadena de Markov resultante presente un buen comportamiento. Las observaciones perteneciantes previamente a las componentes $j1$ o $j2$ se asignarán tras la combinación a la componente j^* mientras que el resto de observaciones se mantendrán en las mismas componentes a las que pertenecían anteriormente. Para la realización del proceso de división se hará uso de las siguientes variables auxiliares

$$u_1 \sim \mathcal{B}(2,2)\ , \quad u_2 \sim \mathcal{B}(2,2), \quad u_3 \sim \mathcal{B}(1,1)\ ,$$

y en base a éstas se dividirá una componente de la mixtura elegida al azar, j^*, según:

$$\begin{array}{cc} w_{j1} = w_{j^*} u1\ , & w_{j2} = w_{j^*}(1-u_1)\ , \\ \mu_{j1} = \mu_{j^*} - u_2\sigma_{j^*}\sqrt{\frac{w_{j2}}{w_{j1}}}\ , & \mu_{j2} = \mu_{j^*} + u_2\sigma_{j^*}\sqrt{\frac{w_{j1}}{w_{j2}}}\ , \\ \sigma^2_{j1} = u_3(1-u_2^2)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j1}}\ , & \sigma^2_{j2} = (1-u_3)(1-u_2^2)\sigma^2_{j^*}\frac{w_{j^*}}{w_{j2}}\ . \end{array} \tag{3.11}$$

La asignación de las observaciones asociadas a la componente j^* a las nuevas componentes $j1$ y $j2$ se realizará en base a la verosimilitud de que cada una de éstas pertenezca a la primera o a la segunda. Es decir, si $L_1(x)$ es la verosimilitud de que la observación x pertenezca a la primera componente y $L_2(x)$ es la verosimilitud de que pertenezca a la segunda, la probabilidad de que se asigne dicha observación a $j1$ vendrá dado por el valor

$$\frac{L_1(x)}{L_1(x)+L_2(x)}\ .$$

Las distribuciones utilizadas de las variables auxiliares u_1, u_2, u_3 se han elegido de forma que se garantice una probabilidad de aceptación razonablemente alta, no obstante, también resultan posibles otras elecciones alternativas que podrían suponer una mejora de ciertas características de la cadena de Markov generada.

Si tras un movimiento de división se realiza una combinación de las componentes generadas, se volverá al estado de la cadena previo a la realización de ambos movimientos. Así, los movimientos de división y combinación realizan acciones opuestas, además las ecuaciones (3.11) establecen una biyección entre los parámetros $\{w_{j1}, \mu_{j1}, \sigma_{j1}, w_{j2}, \mu_{j2}, \sigma_{j2}\}$ y $\{w_{j^*}, \mu_{j^*}, \sigma_{j^*}, u_1, u_2, u_3\}$, tal y como resulta necesario para la implementación del algoritmo del paso reversible.

Una vez se han propuesto los nuevos valores a generar en un movimiento de división, la probabilidad de aceptación de dicha propuesta viene dada por $min\{1, A\}$, donde A se corresponde con:

$$\begin{aligned} A = \quad & \frac{L(y_i^*|z_i^*,\mu^*,\sigma^*)}{L(y_i|z_i,\mu,\sigma)} \frac{p(k+1)}{p(k)} (k+1) \frac{w_{j_1}^{\delta-1+l_1} w_{j_2}^{\delta-1+l_2}}{w_{j^*}^{\delta-1+l_1+l_2} B(\delta, k\delta)} \\ & \times \sqrt{\frac{\kappa}{2\pi}} \exp(-\tfrac{1}{2}\kappa\{(\mu_{j_1}-\xi)^2 + (\mu_{j_2}-\xi)^2 + (\mu_{j_1}-\xi)^2)\}) \\ & \times \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{\sigma_{j_1}^2 \sigma_{j_2}^2}{\sigma_{j^*}^2}\right)^{-\alpha-1} \exp(-\beta(\sigma_{j_1}^2 + \sigma_{j_2}^2 + \sigma_{j^*}^2)) \\ & \times \frac{d_{k+1}}{b_k P_{alloc}} \{g_{2,2}(u_1) g_{2,2}(u_2) g_{1,1}(u_3)\}^{-1} \\ & \times \frac{w_{j^*} \|\mu_{j_1} - \mu_{j_2}\| \sigma_{j_1}^2 \sigma_{j_2}^2}{u_2(1-u_2)^2 u_3(1-u3) \sigma_{j^*}^2} \ , \end{aligned} \tag{3.12}$$

donde, l_1 se corresponde con el número de observaciones correspondientes a la primera componente resultante de la división; l_2 sería el equivalente para la segunda componente; $p(\cdot)$ se corresponde con la distribución inicial del número de componentes de la mixtura (Green y Richardson utilizan una distribución Uniforme); $B(\cdot,\cdot)$ denota la función beta; $L(\cdot|\cdot)$ la función de verosimilitud de cualquier observación; $g_{a,b}(\cdot)$ la función de densidad beta de parámetros a y b; y P_{alloc} se corresponde con la probabilidad de las asignaciones de las observaciones a cada componente tras el proceso de división.

En la ecuación (3.12) las 3 primeras líneas hacen referencia al cociente de distribuciones posteriores de los parámetros del modelo antes y después del

movimiento de división, mientras que la cuarta línea contiene el cociente de las probabilidades de propuesta de un movimiento de división y combinación junto con el cociente de las probabilidades de los valores generados para la realización del movimiento. La última línea de (3.11) se corresponde con el jacobiano de las trasformaciones necesarias para la transición entre los distintos espacios que comporta el movimiento de división.

En el caso del movimiento de combinación, la probabilidad de aceptación viene dada por $min\{1, A^{-1}\}$ según se ha expuesto previamente en el desarrollo de los algoritmos de salto reversible.

El proceso de nacimiento-muerte, el segundo movimiento propuesto por Richardson y Green, procede de la siguiente forma. En primer lugar se realiza una elección aleatoria entre los movimientos de nacimiento y muerte, con probabilidades b_k y $d_k = 1 - b_k$ respectivamente, donde $b_1 = 1$, $b_k = 0{,}5$ para $1 < k < M$ y $b_M = 0$. En el caso de que se lleve a cabo un nacimiento, se añadirá una nueva componente al modelo de mixtura actual muestreando la media y la precisión de sus distribuciones iniciales, mientras que el nuevo peso se genera de una distribución $\mathcal{B}(1, k)$, la distribución marginal de los pesos de la mixtura tras la generación de la nueva componente. El resto de pesos de la mixtura se modifican según la relación $w'_j = w_j \cdot (1 - w^*_j)$ para acomodar el peso de la nueva componente que se ha creado. En el caso de tener que realizarse una muerte, se elegirá al azar una componente de la actual mixtura j^* y se eliminará de ésta. Los pesos de las componentes restantes se reescalarán según $w'_j = w_j/(1 - w^*_j)$ para que vuelvan a sumar 1. Una vez llevados a cabo la generación de la nueva componente o la eliminación de ésta, el nuevo estado generado se habrá de someter a un proceso de aceptación-rechazo. La probabilidad de aceptación del nuevo estado viene dada por $min(1, A)$ para el proceso de nacimiento y de $min(1, A^{-1})$ para la

muerte, donde A toma la siguiente forma

$$A = \frac{p(k+1)}{p(k)} \frac{1}{B(k\delta, \delta)} w_{j^*}^{\delta-1} (1 - w_{j^*})^{n+k\delta} (k+1) \frac{d_{k+1}}{(k_0+1)b_k} \frac{1}{g_{1,k}(w_{j^*})} \ .$$

Por tanto resulta evidente que los movimientos de nacimiento y muerte son de carácter trans-dimensional, ya que, o bien aumentan o disminuyen el número de componentes de la mixtura, lo que conlleva un aumento o disminución del número de parámetros muestreado. Además, ambos movimientos son opuestos ya que el nacimiento de una componente se puede ver revertido mediante la muerte de dicha componente y viceversa.

Una vez se han definido los pares de movimientos reversibles nacimiento-muerte, división-combinación y los movimientos intra-dimensionales de actualización de los parámetros del modelo, el algoritmo de simulación del modelo propuesto procede mediante la repetición sistemática de los siguientes pasos:

- Actualización intradimensional mediante Gibbs Sampling de los parámetros del modelo.
- Movimiento de división-combinación
- Movimiento de nacimiento-muerte

Seguidamente ilustramos la aplicación del algoritmo que se acaba de detallar, sobre un conjunto de datos utilizado repetidamente en el ámbito de los modelos de mixturas.

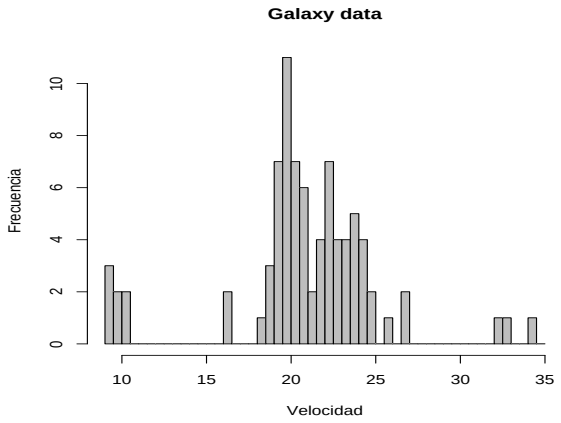

Figura 3.3: Representación del banco de datos de velocidad de galaxias.

Análisis de datos de galaxias

Si hay un banco de datos histórico sobre el que se han aplicado sistemáticamente los distintos avances en los modelos de mixturas unidimensionales, este es el *galaxy data*, utilizado entre otros en Escobar y West (1995) [42], Richardson y Green (1997) [80], Stephens (2000) [93] o Cappé et al. (2003) [24]. Este banco de datos consta de 82 mediciones de la velocidad de distintas galaxias. En la figura 3.3 se puede observar una representación de dichos datos, en la que se puede apreciar la multimodalidad presente. Dicha característica hace idónea la utilización de modelos de mixturas para su estudio.

Uno de los principales problemas a la hora de analizar estos datos es la determinación del número de componentes de la mixtura en las que resulta razonable agrupar los datos. A simple vista, en la representación anterior se puede observar que podrían ser adecuados modelos de mixturas desde 3

hasta 7 componentes o incluso más para describir el comportamiento de los datos. No obstante la representación mediante histogramas puede resultar engañosa y al cambiar la amplitud de las barras en dicha representación podríamos aventurar un número de componentes algo diferente. Por tanto, la elección de un modelo concreto con un número de componentes determinadas no tiene demasiado sentido en este caso, ya que cabe la posibilidad de errar en la elección del número de componentes del modelo utilizado. Además, la no incorporación de dicha incertidumbre en la modelización de la mixtura puede derivar en una estimación de la función de distribución con menor variabilidad de la que debería.

La aplicación del algoritmo de Richardson y Green sobre los datos de galaxias proporciona la siguiente estimación del número de componentes de la mixtura:

k	1	2	3	4	5	6	7	8
p(k)	0	0	0.061	0.128	0.182	0.199	0.160	0.109
k	9	10	11	12	13	14	15	>15
p(k)	0.071	0.040	0.023	0.013	0.006	0.003	0.002	0.003

Tal y como parecía a simple vista un número de componentes en la mixtura inferior a 3 resulta improbable, siendo 6 el número de componentes más probable para describir el conjunto de datos. Por otro lado, también se observa que resultan bastante probables los modelos de mixturas de hasta 10 componentes como máximo, para un número de componentes superior la probabilidad es inferior al 3 %. Recordamos que un modelo de más de 7 componentes parecía improbable a la vista de la representación. Se observa además que la cola derecha de la distribución del número de componentes es bastante alargada. Este hecho es bastante habitual en este tipo de modelización y resulta un inconveniente a la hora de determinar una estimación puntual del número de componentes de la mixtura. Este efecto en la esti-

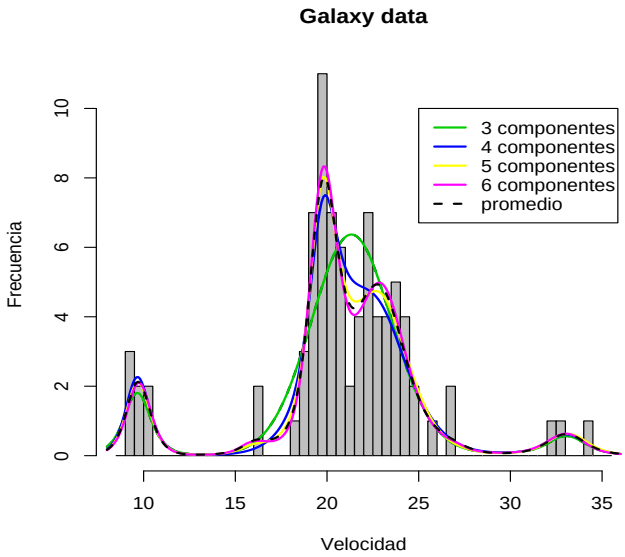

Figura 3.4: Estimación de la función de densidad según el modelo de Richardson y Green. Estimación condicionada al número de componentes (entre 3 y 6) y estimación no condicionada.

mación se debe a que si un modelo resulta creíble, el mismo modelo con una componente más pero con un pequeño peso también resultará razonablemente creíble. Sin embargo, la eliminación de una componente del estado anterior puede conducir a un estado bastante de verosimilitud bastante menor, lo cual produce la asimetría sobre la distribución posterior del número de parámetros.

En la figura 3.4 se puede observar la estimación de la función de intensidad obtenida, representada mediante líneas discontinuas. Además, se incluye también la estimación condicionada al número de componentes en cada iteración del algoritmo cuando dicho número varía entre 3 y 6. Se puede

observar que a partir de 4 componentes la estimación obtenida es bastante similar, siendo casi indistinguibles para valores del número de componentes superiores a 6.

De la misma forma que se ha establecido el algoritmo para el caso de mixturas normales unidimensionales, resulta posible la adaptación de dicho algoritmo al caso multidimensional. De hecho Berger, Castelloe, Dellaportas, Lawson, Roeder o Stephens, entre otros autores, han realizado distintas aportaciones en este sentido. No nos vamos a extender más en esta sección detallando la generalización multidimensional del algoritmo de Richardson y Green ya que ésta resulta obvia en muchos aspectos y se va a tratar con bastante más detalle en el próximo capítulo.

3.3. Otras propuestas

Para terminar el presente capítulo vamos a revisar algunos trabajos dentro del campo de mixturas en los que se plantea el problema de la determinación del número de componentes necesarias para explicar los datos, señalando las particularidades de cada uno de ellos.

En su tesis doctoral, Stephens (1999) [91] estudia distintos problemas asociados a la inferencia bayesiana en modelos de mixturas con un número indeterminado de componentes, tanto para el caso unidimensional como en el bidimensional. Además, Stephens propone la simulación mediante procesos continuos de nacimiento y muerte como alternativa a la simulación de salto reversible. Con el objetivo de acotar la variabilidad de la cola derecha de la distribución del número de componentes, Stephens propone utilizar como distribución inicial de este parámetro una Poisson de media 1. Esta propuesta penaliza sobre la distribución posterior los valores altos del

número de componentes de la mixtura reduciendo la cola derecha de su distribución. Sin embargo, desde nuestro punto de vista la utilización de información inicial contrapuesta a los datos supone una forma artificial de alterar los resultados. Además, la elección de la media utilizada en la distribución de Poisson es arbitraria y no responde a ningún conocimiento previo del problema, por lo que cualquier otro valor propuesto podría ser igualmente válido y la distribución final del número de componentes será sensible a esta elección.

En Cappé et al. (2003) [24] se comparan las 2 técnicas trans-dimensionales utilizadas para la simulación de modelos de mixturas, la simulación mediante procesos de nacimiento-muerte y la de algoritmos de salto reversible. De este trabajo se derivan varios resultados interesantes que utilizaremos a la hora de realizar la generalización bidimensional del algoritmo de Richardson y Green.

En Lawson y Clark (1999) [62] se propone un proceso cluster de Poisson modificado, similar a un proceso de mixturas, para la modelización de la mortalidad por cáncer respiratorio en Armadale, Escocia. En dicho proceso la superficie de intensidad viene dada por la siguiente función de intensidad

$$\lambda(x) = \rho \left(1 + \sum_{j=1}^{m} \mathcal{N}_2(x|\mu_j, \Sigma^2) \right) ,$$

es decir, dicha superficie combina una parte con forma de mixtura de normales bivariantes esféricas, con pesos y varianzas comunes, junto a otra superficie de riesgo uniforme. Esta última componente se utiliza para contemplar el hecho que se pueden observar muertes por cáncer no agrupadas en la población, que se distribuyen de forma aleatoria en la región de estudio. Lawson y Clark también tratan de controlar la cola derecha de la distribución del número de componentes de la mixtura, para ello utilizan

como distribución de las medias de las componentes un proceso de inhibición de Strauss, descrito en el capítulo anterior (2.3). No obstante, los valores de los parámetros de la distribución de Strauss se fijan de forma arbitraria y sospechamos que la distribución final del número de componentes será muy sensible a dichos valores. Además, se señala que la consideración de dichos parámetros como variables en el modelo conlleva muchos problemas de convergencia en el proceso de simulación. Por tanto los resultados de esta propuesta pueden estar muy condicionados a la elección de los parámetros del proceso de Strauss. En este trabajo también se tiene en cuenta la heterogeneidad medioambiental haciendo uso de efectos aleatorios con estructura espacial.

En Castelloe (1998) [26], dentro del marco de procesos puntuales, se estudian los procesos cluster de Poisson sobre un plano desde un planteamiento bayesiano. Dicho proceso consiste en un modelo de mixturas normales bidimensionales en el que todas las componentes comparten el mismo peso y la misma matriz de varianza-covarianza, no necesariamente esférica. Esta última simplificación facilita en gran medida la adaptación bidimensional del algoritmo de salto reversible. Castelloe dedica especial atención a la comparación de la inferencia sobre el número de componentes de la mixtura desde el marco bayesiano y el frecuentista. Desde el marco bayesiano utiliza como técnica de inferencia la simulación de salto reversible mientras que para el planteamiento frecuentista la probabilidad posterior del número de componentes de la mixtura se aproxima, tal y como se propone entre otros en Fraley y Raftery (2002) [46], mediante:

$$\widehat{p(k)} \simeq \frac{\exp(\frac{1}{2}BIC_k^{EM})}{\sum_q \exp(\frac{1}{2}BIC_q^{EM})} ,$$

donde BIC_k^{EM} corresponde al estadístico BIC (Schwarz, 1978 [87]) del modelo de mixturas con k componentes, estimado mediante el algoritmo EM. No obstante, también se estudia la aproximación de la probabilidad del

número de componentes mediante otros estadísticos de selección de modelos como el AIC (Akaike, 1973 [2]) o el AWE (Approximate Weight of Evidence) (Banfield y Raftery, 1993 [9]). A tenor de los resultados de Castelloe parece observarse que, en ocasiones, dichas aproximaciones dan resultados muy diferentes por tanto los resultados obtenidos desde la aproximación clásica dependen en gran medida de la aproximación utilizada. Además, la estimación frecuentista parece mostrar tanta variabilidad como la estimación mediante la aproximación bayesiana.

En Dellaportas y Papageorgiou (2004) [35] se propone un algoritmo de salto reversible para el modelo de mixturas normales multidimensional, a diferencia de todas las propuestas anteriores que son a lo sumo bidimensionales. La mayor dificultad que conlleva la generalización multidimensional del algoritmo RG radica en la definición de la biyección del movimiento de división-combinación. En dicho movimiento el número de parámetros implicados aumenta de forma cuadrática con la dimensión de los datos, por tanto el jacobiano de la transformación propuesta puede llegar a resultar muy complicado de calcular. Dellaportas hace uso de la descomposición en valores propios de la matriz de varianzas-covarianzas de la componente a dividir para definir las nuevas matrices, gracias a lo cual el jacobiano de la transformación resultante se simplifica en gran medida.

En Pérez y Berger (2001) [77] discuten la elección de distribuciones iniciales para la media y la varianza en los modelos de mixturas normales. En su trabajo tratan de definir un análisis por defecto (*default analysis*) para los modelos de mixturas. Concretamente, la aplicación con la que ilustran su propuesta corresponde a un modelo de mixturas bidimensional. En este trabajo se argumenta que la elección de una distribución inicial impropia para dichos parámetros conllevaría una distribución posterior también impropia para ellos, por lo que no se puede recurrir a semejantes distribuciones

como propuesta de distribución inicial no informativa. Por ello, proponen la utilización de *expected posterior priors* (Pérez y Berger, 2002 [78]), como propuesta no informativa a diferencia de la propuesta de Richardson y Green que utilizaban modelos jerárquicos con distribuciones vagas basadas en los datos, en una forma quizás un tanto "ad hoc". En estos trabajos se propone la utilización de dos variantes de las expected posterior priors, el método base y el método empírico. En Pérez y Berger (1999) [76] se compara la propuesta mediante expected posterior priors sobre un banco de datos analizado por Richardson y Green. Los resultados de Pérez y Berger, en cuanto al número de componentes de la mixtura difieren según la variante empleada, así el método empírico proporciona una estimación sustancialmente menor (2-4) que el método base (3-7). Por otro lado, el método base proporciona un rango de valores similar a los obtenidos por Richardson y Green, aunque se pueden apreciar ostensibles diferencias en la posterior del número de componentes según ambos métodos.

Nobile (2004) [74] estudia de forma teórica la distribución del número de componentes en un modelo de mixturas. En su trabajo se deriva una expresión de la distribución inicial del número de componentes no vacías en función de la distribución inicial del número de componentes y el parámetro de la distribución (Dirichlet) de los pesos de la mixtura. Además, en este trabajo se cuestiona la utilidad de los modelos bayesianos de mixturas para realizar inferencia sobre su número de componentes. Dicho cuestionamiento se argumenta en que desde el enfoque bayesiano el número de componentes en el modelo responde a cuantas componentes son compatibles con los datos. Bajo esta última interpretación, si un modelo es compatible con los datos, el mismo modelo con una nueva componente vacía tendrá también una alta verosimilitud. Nobile atribuye a este hecho la cola derecha en la distribución del número de componentes de la mixtura y propone el número de componentes no vacías como un indicador mejor del número de compo-

nentes necesarias para explicar los datos. Según Nobile este hecho no se da en el enfoque frecuentista ya que desde ese punto de vista se determina el menor número de componentes necesarias para describir los datos.

Otra aproximación a la determinación del número de componentes en un modelo de mixturas viene dada por los modelos de mixturas de procesos Dirichlet descritos entre otros en Escobar y West (1995) [42] y en Ishwaran y Zarepour (2000) [57]. En dicha aproximación se supone que cada individuo sigue una distribución normal de parámetros exclusivos para dicho individuo. Sin embargo la distribución inicial de dichos parámetros sigue un proceso Dirichlet, distribución no paramétrica en la que la probabilidad de que dos observaciones compartan un mismo valor es positiva. Es más, si $\Theta = \{\theta_1, ..., \theta_m\}$ es el conjunto de parámetros distintos de las observaciones $\{y_1, ..., y_n\}$ en un proceso Dirichlet de parámetro $\alpha > 0$, la probabilidad de que los parámetros de una nueva observación y_{n+1} pertenezcan a Θ será $1 - 1/(n + \alpha)$. Por tanto el proceso Dirichlet para las componentes de la mixtura tenderá a agrupar los datos en un número de grupos inferior al número de observaciones. En Ishwaran y Zarepour (2000) [57] se señala que las mixturas de procesos Dirichlet permiten aproximar el número de componentes de una mixtura, mediante el estudio del número de componentes de la mixtura no vacías. Sin embargo, según se señala en dicho artículo, generalmente este procedimiento sobreestima dicho número. Por otra parte, Escobar y West estudian también el ejemplo de las galaxias. En el siguiente cuadro se puede observar tanto la distribución inicial utilizada para el número de componentes como la distribución posterior obtenida:

k	1	2	3	4	5	6	7	8	9
p(k)	.01	.06	.14	.21	.21	.17	.11	.06	.02
p(k|X)	.03	.13	.26	.26	.18	.09	.04	.01	.00

La distribución inicial utilizada para este parámetro no se ha definido de forma explícita sino que es consecuencia de las distribuciones iniciales establecidas para otros parámetros del proceso. A diferencia de la propuesta de Richardson y Green, las mixturas de procesos Dirichlet no incorporan el número de componentes de la mixtura como un parámetro propio del modelo, por lo que estos modelos no permiten un control directo de este parámetro. Este hecho supone desde nuestro punto de vista un problema de este tipo de modelización ya que, por ejemplo, en el ejemplo estudiado no resulta claro el efecto de la distribución inicial del número de agrupaciones sobre la distribución posterior de este valor.

Distintos autores han enfocado el problema de la determinación del número de componentes de una mixtura desde el punto de vista de selección de modelos. Las aproximaciones comentadas hasta ahora determinan la probabilidad de cada número de componentes y promedian el resultado de todas las propuestas posibles en función de su probabilidad. Esta aproximación se conoce como *model averaging* o promediado de modelos. Por el contrario, los métodos de selección de modelos ignoran la incertidumbre en la distribución de los posibles modelos que podrían describir satisfactoriamente los datos. De hecho el objetivo de estos métodos consistirá en la elección del modelo que parezca más apropiado para describir los datos ignorando el resto a la hora de elaborar las conclusiones. En este tipo de aproximaciones suele ser habitual la utilización de factores Bayes para la elección del número de componentes de la mixtura. Desde el enfoque clásico Raftery tiene un gran número de trabajos (Fraley y Raftery, 2002 [46]; Fraley y Raftery, 1998 [45]; Banfield y Raftery, 1993 [9]) en los que determina el número de componentes de la mixtura aproximando el factor bayes de varios modelos como función de su BIC, de forma similar a la que hacía Castelloe en su trabajo. Dentro del marco bayesiano Roeder y Wasserman (1987) [85] también hayan la probabilidad de cada número de componentes

para el ejemplo de las galaxias, haciendo uso del estadístico BIC y dotando a las distribuciones iniciales de los parámetros de estructura Markoviana (*distribuciones iniciales parcialmente propias*). Roeder y Wasserman obtienen una probabilidad de que el número de componentes sea 3 superior a 0.999. También desde el marco bayesiano Moreno y Liseo (2003) [71] hacen uso de factores Bayes para la determinación de la probabilidad de cada número de componentes, pero en este caso no se hace uso del BIC para aproximar los factores Bayes. Para el ejemplo de las galaxias, Moreno y Liseo obtienen una probabilidad para 3 componentes de 0.84 y para 4 componentes de 0.16.

De todas formas nuestra opinión es que, al menos desde el enfoque bayesiano, resulta aconsejable incorporar al proceso de inferencia la incertidumbre respecto a la determinación del modelo más apropiado para la descripción de los datos. En caso contrario se obtendrán resultados más precisos de lo que se debería. Por ello la aproximación mediante promediado de modelos nos parece más atractiva ya que considera conjuntamente todos los modelos posibles y no calcula para cada uno de ellos su probabilidad específica.

Tal y como se puede observar existe una gran variedad de propuestas de estimación del número de componentes de una mixtura. Por tanto este campo supone una línea de investigación bastante activa en la actualidad. Sin embargo, resulta desalentador comprobar cómo los resultados de las distintas propuestas difieren entre sí por lo que pensamos que se debe invertir todavía bastante más esfuerzo en el estudio de estos modelos.

Capítulo 4

Modelización de procesos puntuales mediante mixturas

Hasta el momento se han descrito los procesos puntuales como el marco apropiado para el estudio de brotes epidémicos en los que se dispone de la localización exacta a la que se pueden atribuir los casos observados. Por otro lado se ha introducido la modelización estadística mediante modelos de mixturas para la descripción de datos agrupados, resultado por ejemplo del efecto de distintas fuentes de riesgo como el brote que nos ocupa. La modelización mediante mixturas supone una generalización de los Procesos Cluster de Poisson, tal y como se expone en Castelloe (1998) [26]. Concretamente, los modelos de mixturas contemplan el que las distintas agrupaciones tengan tamaños diferentes en cuanto al número de casos que agrupan y la distribución de los casos alrededor de su centroide sea distinta para cada agrupación en los casos. Por tanto la modelización de procesos puntuales mediante mixturas introduce nuevas posibilidades de modelización dentro de este marco, que nos va a permitir un ajuste más flexible que el de los propios procesos cluster de Poisson.

La modelización mediante mixturas, nos va aportar una herramienta útil para la descripción de clusters de tipo individual, pero no para las agregaciones de tipo general, consecuencia de la heterogeneidad medioambiental que no se contemplan expresamente en este tipo de modelización. Lawson y Clark (1999) [62], hasta donde nosotros conocemos, supone el primer trabajo en la modelización de procesos epidémicos mediante procesos puntuales que considera explícitamente ambos procesos de agrupación, el general y el individual. Concretamente, estos autores consideran un proceso de Poisson donde la superficie de riesgo empleada para describir un patrón geográfico de mortalidad tiene la siguiente forma:

$$\lambda(x_i|\theta) = \rho \cdot g(x_i) \cdot \left(1 + \sum_{k=1}^{K} \mathcal{N}_2(x_i|\mu_j, \sigma^2 I_2)\right) \cdot \exp(u_i + v_i)\ . \qquad (4.1)$$

Donde I_2 denota la matriz identidad de dimensión 2. En la superficie anterior ρ ajusta el valor promedio del proceso en cualquier punto. Este valor, junto con el área de la región de estudio, determina el número de observaciones esperadas en el proceso. Por otro lado, $g(\cdot)$ estima la distribución geográfica de la población a riesgo en la región de estudio. Dicha función se basa en una muestra de controles obtenida previamente, a partir de éstos se propone una estimación no paramétrica del riesgo mediante métodos kernel. El tercer término de la función de intensidad anterior corresponde a la modelización del clustering individual del proceso, es decir, aquellas agrupaciones en torno a ciertas localizaciones concretas (aunque desconocidas) que son consecuencia de la existencia de una fuente de riesgo en dicho emplazamiento. Este término se expresa como una mixtura de distribuciones normales bivariantes en la que todas las componentes comparten el mismo peso y la misma matriz de varianza-covarianza, más un término uniforme sobre la región de estudio. Este último término responde a que la enfermedad estudiada (cáncer respiratorio) no es de tipo infeccioso y se espera que una proporción de los casos haya surgido de forma aleatoria y semejante a

la distribución de la población en la región de estudio. Por tanto, el proceso puntual considerado por Lawson y Clark contempla 2 mecanismos de generación de casos, un mecanismo que produce agrupaciones en torno a ciertas localizaciones desconocidas y un segundo mecanismo que genera casos de forma semejante a la distribución de la población en la región de estudio.

El último término de la función de intensidad responde a un mecanismo de clustering general que se ha creído conveniente contemplar en la modelización. Este mecanismo consta de 2 efectos aleatorios que tomarán valores diferentes para cada individuo, contemplándose así las variaciones en la susceptibilidad de la enfermedad en la población a estudio. El primero de los efectos aleatorios se distribuye de forma independiente para cada individuo, mientras el segundo sigue una distribución normal multivariante en la que la correlación entre puntos depende de la distancia que les separa. Así pues, el comportamiento de estos dos efectos aleatorios imitan los modelos de suavización geográfica de riesgos, como el de Besag et al. (1991) [15], en los que también se incluyen 2 efectos aleatorios con estructuras de covarianza similares a las empleadas en el trabajo de Lawson y Clark. Para la modelización de la heterogeneidad medioambiental ambos autores hacen uso de efectos aleatorios, ya que su flexibilidad resulta idónea para la descripción de este tipo de agregación, de la que no se suele tener ninguna idea a priori sobre su forma. Por el contrario, para la modelización de las agrupaciones individuales emplean una forma funcional definida ya que este proceso de agrupación se intuye centrado alrededor de ciertas localizaciones concretas.

El trabajo de Lawson y Clark supone una propuesta de indudable interés tanto estadístico como epidemiológico, ya que contemplan en un único modelo las agrupaciones de origen individual y general. No obstante, desde nuestro punto de vista existen distintos aspectos en dicho trabajo susceptibles de mejora. En primer lugar, la parte de clustering individual se podría

generalizar mediante un modelo de mixturas con un número indeterminado de componentes como los que han sido descritos en el capítulo 3. De esta forma, el modelo incorporaría distintas matrices de varianza-covarianza y pesos para cada componente. Además el modelo de mixturas que se vaya a considerar, se puede generalizar también contemplando matrices de varianza-covarianza no esféricas para sus componentes. Por otro lado, sobre el término de clustering individual de la propuesta de Lawson y Clark intervienen dos componentes, una contempla los casos que aparecen agrupados y la otra contempla los casos que aparecen aleatoriamente en la población; ambos términos pueden ser incluidos también en un modelo de mixturas, tal y como se expone en Dasgupta y Raftery (1998) [33]. Sin embargo, desde nuestro punto de vista, la formulación utilizada por Lawson y Clark resulta muy rígida, ya que el número de casos que se espera que genere la componente uniforme es $\rho \int_A 1dx = \rho|A|$, donde A es la región de estudio, mientras que el número de casos que se espera que genere cada agrupación es $\rho \int_A \mathcal{N}_2(x|\mu_j, \sigma^2 I_2)dx \approx \rho$. En la expresión anterior la igualdad no se da de forma estricta debido al efecto frontera, consecuencia de considerar una región de estudio finita. La cuestión es que si el área de la región de estudio es grande, el tamaño de las agrupaciones alrededor de las fuentes de riesgo serán pequeñas en relación al número de casos generados por la componente uniforme. De la misma forma si el área de la región de estudio es pequeña, la importancia del término uniforme respecto a las agrupaciones es pequeña. En cualquier caso, tal y como se ha planteado el modelo, el tamaño de la región de estudio determina la importancia de la componente uniforme dentro del proceso de agrupación individual. Dasgupta y Raftery (1998) [33] solventan este problema considerando pesos tanto para la componente uniforme como para las gaussianas bivariantes. Este problema también podría solucionarse mediante la utilización de modelos de mixturas.

Sobre el término de clustering general también existen ciertos aspectos

que consideramos mejorables o sobre los que incluso discrepamos. Tanto el efecto aleatorio espacial como el heterogéneo se han referido en la definición del proceso a los individuos en lugar de a sus localizaciones, es decir, se definen tantos valores de los efectos aleatorios como individuos han sido observados. Sin embargo la ecuación (4.1) corresponde a la superficie de intensidad de un proceso de Poisson, por tanto debería estar referida a las localizaciones de la región de estudio y no en función de las observaciones del proceso. Además, la estructura de covarianza del efecto aleatorio espacial depende de la localización de las observaciones del proceso, por tanto de los datos analizados. Así, nuestra opinión sobre el planteamiento propuesto por Lawson y Clark en su trabajo es que se está utilizando dos veces la información de los datos, una como datos en sí y otra a la hora de definir la estructura del efecto aleatorio espacial. En comunicación personal con uno de los autores, éste justificaba la utilización de la localización de las observaciones para definir el efecto aleatorio espacial como una aproximación empírico-bayesiana. En nuestro caso preferiríamos utilizar un término espacial que no duplicara la información de los datos ya que no tenemos idea de qué efecto puede tener dicho artefacto sobre el resultado final.

A tenor de lo expuesto parece que la modelización conjunta del proceso de agregación general e individual en un mismo modelo no es un tema completamente resuelto. Hasta donde nosotros conocemos la única propuesta realizada, al menos en aplicaciones epidemiológicas, es la de Lawson y Clark y, tal y como acabamos de exponer, encontramos ciertas aspectos mejorables en ella. Por ello se hace necesaria la formulación de alguna alternativa que incorpore de forma satisfactoria ambos mecanismos de agregación. Ese será el propósito de la modelización que nos disponemos a introducir.

4.1. Procesos cluster de Poisson y modelos de mixturas

En la sección 2.4.2 se describían los procesos cluster de Poisson como propuesta para la modelización de clustering individual en el estudio de procesos puntuales. En dichos procesos los datos se agrupaban según un proceso de padres e hijos donde tanto el número de padres, como el número de hijos para cada padre es desconocido. En dicha sección se señala también que dicho proceso se puede considerar como un proceso de Poisson con superficie de intensidad:

$$\lambda(s) = \alpha \sum_{c \in C} h(s - c) \ , \tag{4.2}$$

donde c son los valores generados a partir de un proceso de Poisson homogéneo de intensidad ρ, α es el número de hijos esperado para cada padre y h una función de distribución determinada. La ecuación (4.2) recuerda en cierta medida a un modelo de mixturas; en concreto, si para cada componente c, consideramos un número esperado de hijos distinto α_c y la distribución $h(\cdot)$ como una distribución normal bivariante centrada en c y con matriz de varianzas-covarianzas propia, entonces dicha ecuación resulta:

$$\lambda(s) = \sum_{c \in C} \alpha_c \mathcal{N}_2(s|c, \Sigma_c^2) \ ,$$

donde la única diferencia con los modelos de mixturas bidimensionales considerados en el capítulo anterior es que en este último la suma de los distintos α_c vale necesariamente 1. Pero la ecuación anterior puede ser reexpresada como:

$$\lambda(s) = \left(\sum_{c \in C} \alpha_c \right) \cdot \left(\sum_{c \in C} \left(\frac{\alpha_c}{\sum_{c \in C} \alpha_c} \right) \mathcal{N}_2(s|c, \Sigma_c^2) \right) \ , \tag{4.3}$$

donde el segundo término se corresponde con un modelo de mixturas y el primero con el número total de casos esperados por el proceso. Por tanto

observamos que mediante una generalización sencilla, los modelos cluster de Poisson se pueden expresar como modelos de Poisson donde la superficie de intensidad es proporcional a una superficie con forma de mixtura.

El proceso de Poisson definido por (4.3) no sólo generaliza a los procesos cluster de Poisson, sino que también supone una generalización de los procesos de mixturas al contemplar éstos como procesos de Poisson. Dicha generalización permite considerar el número de observaciones de la mixtura como un valor desconocido, lo cual no resultaba posible fuera del marco de los procesos de Poisson, por ejemplo, en el modelo *RG*. Por tanto, el planteamiento del modelo de mixturas como proceso de Poisson resulta más apropiado para el análisis de datos en los que el número de observaciones no venga determinado por el diseño de la recogida de datos, sino que el tamaño muestral en el problema sea aleatorio. Así, cualquier proceso de Poisson donde la función de intensidad sea de la forma (4.3) se puede considerar como una generalización tanto de los procesos cluster de Poisson en los que los hijos se distribuyen según una distribución normal bivariante alrededor de los padres, como de los modelos de mixturas normales bivariantes.

Los procesos de Poisson que tienen la forma de la ecuación (4.3) no han sido posibles de abordar hasta el desarrollo de los métodos de simulación trans-dimensional. Este hecho se debe a que en los procesos cluster de Poisson el número de padres no es un valor conocido, por lo que la intersección de ese tipo de modelos con los modelos de mixturas se da únicamente si el número de componentes en éstas también es indeterminado. Así pues, hasta que la simulación trans-dimensional no ha hecho posible contemplar modelos como el de Richardson y Green no ha sido posible considerar conjuntamente los procesos cluster de Poisson y los modelos de mixturas.

Además, el marco conjunto que acabamos de proponer permite enrique-

cer la modelización mediante mixturas con las ventajas que conllevan los procesos de Poisson. En concreto, recordamos que los modelos de Poisson se generalizaban como procesos de Cox y que disponíamos en la literatura de varios procesos de este tipo bastante indicados para la modelización de la heterogeneidad medioambiental. En la siguiente sección haremos uso de un proceso log-gaussiano para modelizar dicha heterogeneidad. De esta forma, el modelo (4.3) podrá incorporar la modelización del clustering general sin tener que recurrir a la introducción de efectos aleatorios asociados a las observaciones y que pueden duplicar la información utilizada en el modelo.

4.2. Propuesta de modelización univariante

Una vez se han descrito los esfuerzos de otros autores por modelizar conjuntamente los procesos de agrupación de tipo individual y general, pasamos a describir nuestra propuesta. Comenzamos formulando la versión univariante, es decir, la localización de los casos vienen dados por una única coordenada. Este caso carece de interés epidemiológico, al menos en estudios de tipo espacial, pero su planteamiento nos permitirá comparar los resultados de nuestro modelo con otros, como el de Richardson y Green. Dicha comparación permitirá afrontar el problema bidimensional con un conocimiento más profundo de las posibilidades y carencias de nuestra modelización.

Nuestra formulación consistirá en describir los datos como un proceso de Cox que integrará un proceso log-gaussiano junto con una superficie de riesgo con forma de mixtura de distribuciones normales. En la figura 4.1 se representa la combinación de ambos procesos. Por un lado, tenemos el proceso log-gaussiano que representará las fluctuaciones aleatorias de estructura espacial, que atienden a la variabilidad medioambiental de la población. Por

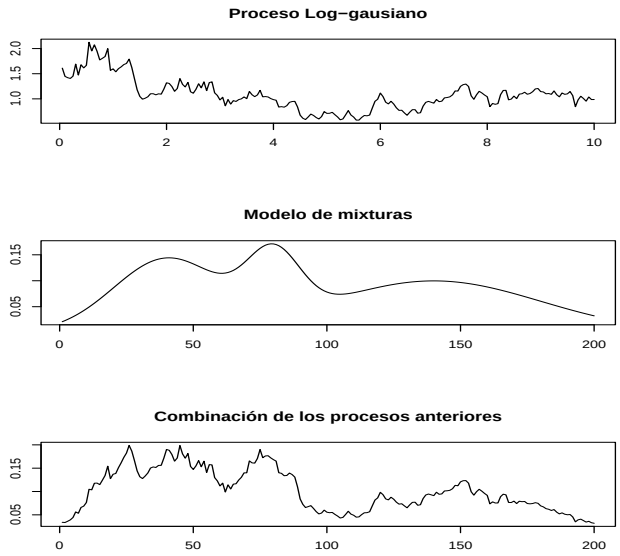

Figura 4.1: Proceso log-gaussiano con $\rho = 1, \sigma^2 = 0{,}25$; Modelo de mixturas $0{,}35 \cdot \mathcal{N}(2,1) + 0{,}15 \cdot \mathcal{N}(4, 0{,}25) + 0{,}5 \cdot \mathcal{N}(7,4)$; y combinación de ambos procesos.

otro lado, tenemos el modelo de mixturas que plasma las elevaciones en el riesgo alrededor de ciertas localizaciones concretas, las fuentes de riesgo. Por último, la propuesta que presentamos, que combina ambos procesos, es capaz de adaptarse a los dos mecanismos de agregación contemplados, el general y el individual.

Comenzaremos estableciendo ciertos criterios de notación previos a la definición del modelo. Denotaremos por $X = \{x_1, x_2, ...\}$ las localizaciones de las observaciones, o casos del brote. C será el soporte o dominio del proceso puntual y supondremos $C = \bigcup_{k=1}^{K} C_k$ una partición de dicho dominio sobre la que se definirá el proceso log-gaussiano. Notar que C habrá de ser

necesariamente finito ya que de otra forma no se podría definir un proceso log-gaussiano sobre él, al requerirse una partición finita del dominio para acoger este tipo de procesos. Denotaremos por $f(t)$ una función escalonada sobre C de forma que para cada celda C_k, f tomará un valor constante sobre ella. Así, para cada $t \in C_k$, $f(t)$ tomará un valor constante que denotaremos por $f(C_k)$. Por último, dado un valor $t \in C$, denotaremos por $C(t)$ la celda de la partición a la que pertenezca t. Así, para cualquier $t \in C_k$ para cierto valor de k se tiene $f(t) = f(C(t)) = f(C_k)$.

Concretamente, proponemos utilizar el producto de un proceso de Cox log-gaussiano y una superficie de intensidad con forma de mixtura. Además plantearemos el modelo de Cox como un modelo jerárquico bayesiano de forma que le podamos dar la estructura que se considere más oportuna a sus distintos niveles. La primera capa del modelo jerárquico expresará que el modelo que planteamos se corresponde con un modelo de Cox. Así, ésta se define de la siguiente manera

$$x_i|\lambda(t) \sim \prod_{x_i \in X} \lambda(x_i)\exp\left(-\int_C \lambda(t)dt\right) \ . \tag{4.4}$$

Por tanto, la función de intensidad del proceso viene definida por λ, una función real positiva que va a tomar la siguiente expresión

$$\lambda(t|Z, \mu, \tau, f) = \exp(f(t)) \sum_{j=1}^{m} w_j \mathcal{N}(t|\mu_j, \tau_j^{-1}) \ \ \forall t \in C \ , \tag{4.5}$$

donde, tal y como hemos comentado, $f(t)$ será una función escalonada sobre la partición que se ha escogido para C y el término del sumatorio se corresponde con un modelo de mixturas. De esta forma integraremos una componente de agregación individual, la mixtura, y una componente de agregación general, el término log-gaussiano. Respecto a los parámetros de

la mixtura, les vamos a dar las mismas distribuciones iniciales que se proponían en *RG*:

$$\mu_j \sim \mathcal{N}(\xi, \kappa) \;\; j = 1, ..., m \; ,$$

$$\tau_j|\beta \sim \Gamma(\alpha, \beta) \;\; j = 1, ..., m \; ,$$

$$w \sim \text{Dir}(\delta, ..., \delta) \; ,$$

$$\beta \sim \Gamma(g, h) \; ,$$

$$m \sim \mathcal{U}(0, M) \; .$$

De esta forma, la parte de clustering individual sigue las especificaciones de una propuesta que hasta la fecha está teniendo una gran aceptación en la literatura de los modelos de mixturas. Así, también tendremos un referente con el que comparar los resultados proporcionados nuestra propuesta. Respecto a la distribución inicial de los parámetros de la parte de clustering general vamos a proponer que el logaritmo de f tome una distribución normal multivariante sobre las celdas que integran la región de estudio. Por lo que la distribución de este mecanismo de agregación seguirá una distribución log-gaussiana. Así pues, definimos la distribución inicial de los parámetros de esta componente como

$$\log(f(\{C_1, ...C_K\}))|\phi, \Sigma \sim \mathcal{N}_K(\phi, \Sigma^2) \; ,$$

$$\phi \sim \mathcal{U}(-\infty, \infty) \; .$$

El término ϕ hace la función del término ρ que Lawson y Clark empleaban en su propuesta. ϕ modeliza el logaritmo del número de casos esperado por el proceso y, al igual que en la propuesta de Lawson y Clark, se podría haber incluido como un término más del modelo separado del término log-gaussiano y la mixtura. Sin embargo, hemos considerado que la inclusión de éste en el término log-gaussiano simplificaba en cierta medida

la formulación del modelo. Por último, para terminar de definir la componente log-gaussiana tendremos que precisar la estructura de la matriz de varianza-covarianza de ésta. Dicha matriz se ha definido como una función exponencial de la distancia entre las celdas, en concreto si d_{ij} es la distancia entre los centroides de las celdas C_i y C_j, se define la covarianza entre los valores $log(f(C_i))$ y $log(f(C_j))$ como

$$\Sigma_{ij}^2(\sigma, \rho) = \sigma^2 \exp(-\rho d_{ij}) \ .$$

Observar que σ será la desviación típica del logaritmo del efecto aleatorio $f(\cdot)$, mientras que ρ controlará la correlación entre las distintas celdas de la región de estudio. Concretamente, las localizaciones que disten entre sí una distancia de $3/\rho$ presentan una correlación de aproximadamente 0.05, por lo que dicho valor se suele interpretar como la distancia máxima a la que dos puntos presentarán una correlación apreciable. Existen otras posibles parametrizaciones de la matriz de varianza-covarianza como función de la distancia entre puntos, es más, en Möller et al. (1998) [69] se describen las condiciones necesarias para que una parametrización defina una matriz de covarianza válida. Sin embargo, se ha evitado elegir otras parametrizaciones que dependan de un número de variables mayor, ya que en ese caso la estimación de los parámetros resultaría más inestable. Respecto a la distribución inicial de los parámetros de la matriz de covarianza se han utilizado las siguientes distribuciones

$$\sigma \sim \mathcal{U}(0, b_\sigma) \ ,$$

$$\rho^{-1} \sim \mathcal{U}(a_\rho, b_\rho) \ .$$

La elección de la distribución inicial de la desviación típica se ajusta a las directrices propuestas entre otros por Gelman et al. (2004) [48], Gelman (2004) [47], Spiegelhalter et al. (2004) [89], para la varianza de efectos aleatorios. En Möller y Waagepetersen (2004) [70] se propone la utilización de una distribución $\Gamma(0, 10^{-5})$ impropia sobre σ^{-2}, mientras que Berger et al. (2001)

[11] utilizan como distribución inicial para la varianza una $Gamma(0, 0)$, también impropia. Nuestra elección garantiza que la distribución final de esta variable será propia, al tiempo que es defendida también por otros autores.

En cuanto a la distribución inicial de ρ, Berger et al. (2001) [11] establecen ciertas condiciones sobre ella para que la distribución posterior sea propia. En dicho trabajo se propone la distribución de referencia sobre este parámetro como la mejor alternativa no informativa y que garantiza que la distribución final será propia. Sin embargo, el cálculo de dicha distribución de referencia resulta ciertamente complejo, por lo que recurriremos a otras distribuciones alternativas. En Thomas et al. (2004) [96] se propone la utilización de una distribución inicial uniforme para ρ sobre un rango de valores acotado, aunque por otra parte Möller y Waagepetersen (2004) [70] proponen una distribución inicial sobre el logaritmo de ρ también sobre un rango de valores acotado. En ambas propuestas la acotación de la distribución uniforme resulta necesaria ya que en caso contrario la posterior resultaría impropia tal y como se expone en Berger et al. (2001) [11]. Por último, Diggle et al. (1998) [41] proponen utilizar una distribución inicial uniforme sobre ρ^{-1}. En principio no tenemos ningún argumento para optar por ninguna de estas dos propuestas, por lo que habremos de realizar un estudio más profundo de las implicaciones que conlleva la elección de una u otra distribución inicial para este parámetro.

En la figura 4.2 se ha realizado una representación de la función de correlación del proceso, $\Sigma^2(1, \rho)$, para ρ tomando valores enteros ente 1 y 10. La función de correlación que toma valores superiores corresponde a la curva definida por un valor de $\rho = 1$, conforme ρ toma valores más grandes las funciones de correlación disminuyen hasta $\rho = 10$ que corresponde a la curva inferior. En dicha representación se puede observar que conforme aumenta

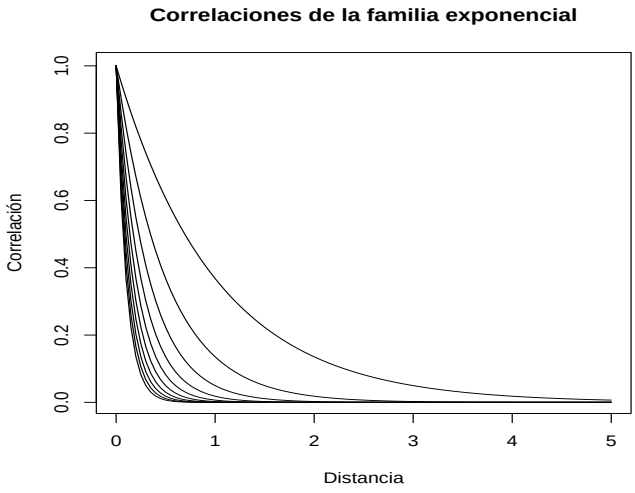

Figura 4.2: Función de correlación exponencial con ρ tomando valores entre 1 y 10. $\rho = 1$ corresponde con la función de correlación superior.

ρ las curvas son más y más parecidas, en concreto, para los valores más altos de ρ las funciones de correlación apenas cambian. En consecuencia, una distribución uniforme sobre ρ acumularía una gran masa en aquellas funciones de correlación que decaen más rápidamente, resultando esta elección ciertamente informativa. Así, sería deseable que la distribución inicial que establezcamos sobre ρ concediera probabilidades similares tanto a las funciones de correlación que decaen más rápido como aquellas que decaen más lentamente. Para ello habremos de definir una distribución inicial para ρ que ponga mayor masa en aquellos valores donde la función de correlación varíe de forma más rápida al variar ρ, mientras que la masa de la distribución inicial habrá de ser menor en aquellos valores donde una variación de este parámetro produzca una leve variación en la función de correlación. Como la función de correlación disminuye en todos los puntos de su dominio al aumentar el valor de ρ, podemos valorar la variación que se produce en dicha función al variar este parámetro en términos de su integral. Así, desde este punto de vista, la distribución inicial sobre ρ se podría definir proporcional a la siguiente expresión

$$P(\rho) \propto \left| \frac{\partial}{\partial \rho} \left(\int_0^\infty \exp(-\rho \cdot D) dD \right) \right| .$$

De esta forma, la masa de la distribución inicial será mayor en aquellos valores de ρ que produzcan mayor variación en la función de correlación, tal y como deseábamos. Así la distribución de ρ resulta

$$P(\rho) \propto \left| \frac{\partial}{\partial \rho} \left(\frac{\exp(-\rho \cdot D)}{-\rho} \Big|_0^\infty \right) \right| = \left| \frac{\partial}{\partial \rho} \rho^{-1} \right| = \rho^{-2} ,$$

o, lo que sería equivalente, considerar una distribución uniforme sobre ρ^{-1} tal y como se propone en Diggle et al. (1998) [41]. Es por este motivo por el que nos hemos decantado por dicha elección para definir la distribución inicial de ρ.

Respecto a los límites de la distribución inicial de ρ^{-1}, hemos fijado el límite superior de la distribución de forma que dadas dos celdas contiguas de la partición de la región de estudio, la correlación entre ellas sea superior a 0.05. De no poner esta restricción se producirían problemas computacionales al contemplar el modelo situaciones que no es capaz de discriminar. Básicamente, el modelo no puede distinguir si resulta preferible una correlación de 0.05 entre celdas contiguas y casi nula entre el resto de celdas o una correlación de 0.01 entre celdas contiguas y casi nula también para el resto de celdas, al no aportar los datos casi información para distinguir entre ambas situaciones. Es por ello que se debe acotar ese rango de valores ya que si el proceso de simulación entra en él le costará salir, produciéndose los citados problemas computacionales. Por otra parte, el límite inferior de esta distribución se ha fijado de forma que la correlación entre la celda central de la región de estudio y las celdas más extremas sea también inferior a 0.05. En el caso de no acotar dicha correlación se producirían problemas para identificar la media del proceso log-gaussiano al estar todas las observaciones muy correlacionadas.

A partir de este momento cuando nos refiramos a nuestra propuesta lo haremos como *propuesta semiparamétrica* para diferenciarla del modelo *RG*, que incluye únicamente la parte de mixturas. Llamaremos semiparamétrica a nuestra propuesta ya que integra dos mecanismos de agregación diferentes, un modelo de mixturas con funciones de distribución paramétricos y un proceso log-gausiano de mayor flexibilidad. Es más, según Bernardo y Smith (1994) [12] los procesos log-gausianos se pueden considerar modelos no paramétricos, desde un punto de vista bayesiano, ya que constan de un número infinito de parámetros (aunque en la práctica se reducen a un número finito para que sean computacionalmente abordables). Es por esto que denotamos nuestra propuesta como semiparamétrica, ya que integra dos procesos, uno paramétrico, el modelo de mixturas, y otro no paramétrico,

el proceso log-gaussiano. Por último señalamos que emplearemos el término semiparamétrico tanto para el caso unidimensional como el bidimensional.

4.2.1. Formulación del modelo completo

Hasta el momento, hemos definido el proceso puntual que entendemos puede modelizar de forma apropiada los procesos de agregación general e individual, ante un problema como el que se nos plantea. Sin embargo, en la propuesta *RG* se planteaba el modelo de mixturas haciendo uso de una *formulación completa* que incluía variables auxiliares que asignaban cada individuo a una componente de la mixtura. Dicha formulación completa, según pudimos comprobar, conllevaba distintas ventajas computacionales que hacían aconsejable su utilización. Por ello nos plantemos también una formulación completa para el modelo propuesto en el apartado anterior.

En esta propuesta vamos a utilizar también una variable auxiliar para cada individuo, de la misma forma que se utilizaba en los modelos de mixturas con un número indeterminado de componentes. Así, la función de verosimilitud de la formulación completa resultará

$$X|\lambda(t) \sim \prod_{x_i \in X} \exp(f(x_i))\mathcal{N}(x_i|\mu_{Z_i}, \tau_{Z_i}^{-1})\exp\left(-\int_C \lambda(t)dt\right) , \qquad (4.6)$$

donde

$$\lambda(t) = \exp(f(t)) \sum_{j=1}^{m} w_j \mathcal{N}(t|\mu_j, \tau_j^{-1})$$

y las variables auxiliares toman como distribución

$$Z_i \sim \text{Mn}(w) \;\; i = 1, ..., |X| \; .$$

Si en (4.6) se integra Z_1 resulta la siguiente función de verosimilitud

$$X|\lambda(t) \sim \left(\exp(f(x_1))\sum_{j=1}^{m} w_j \mathcal{N}(x_1|\mu_j, \tau_j^{-1})\right) \times$$

$$\times \left(\prod_{x_i \in X \setminus x_1} \exp(f(x_i))\mathcal{N}(x_i|\mu_{Z_i}, \tau_{Z_i}^{-1})\right) \exp\left(-\int_C \lambda(t)dt\right) .$$

De la misma forma, si integramos el resto de variables auxiliares $\{Z_2, ..., Z_n\}$ la función de verosimilitud resulta

$$X|\lambda(t) \sim \left(\prod_{x_i \in X} \exp(f(x_i))\sum_{j=1}^{m} w_j \mathcal{N}(x_i|\mu_j, \tau_j^{-1})\right) \exp\left(-\int_C \lambda(t)dt\right) =$$

$$= \left(\prod_{i=1}^{n} \lambda(x_i)\right) \exp\left(-\int_C \lambda(t)dt\right) .$$

Por tanto, resulta obvia la equivalencia entre la formulación del modelo completo y la del modelo original. Así, la formulación final del modelo completo queda de la siguiente forma:

$$\begin{gathered}
X|Z, \mu, \tau, w, f, \lambda(t) \sim \left(\textstyle\prod_{x_i \in X} \exp(f(x_i))\mathcal{N}(x_i|\mu_{Z_i}, \tau_{Z_i}^{-1})\right) \exp(-\textstyle\int_C \lambda(t)dt) \\
Z_i \sim \mathrm{Mn}(w) \;\; i = 1, ..., |X| \\
\lambda(t) = \exp(f(t)) \textstyle\sum_{j=1}^{m} w_j \mathcal{N}(t|\mu_j, \tau_j^{-1}) \\
\mu_j \sim \mathcal{N}(\xi, \kappa^{-1}) \;\; j = 1, ..., m \\
\tau_j \sim \Gamma(\alpha, \beta) \;\; j = 1, ..., m \\
w \sim \mathrm{Dir}(\delta, ..., \delta) \\
m \sim \mathcal{U}(0, M) \\
f(\{C_1, ...C_K\})|\phi, \Sigma \sim \mathcal{N}_K(\phi, \Sigma^2) \\
\phi \sim \mathcal{U}(-\infty, \infty) \\
\Sigma^2(i, j) = \sigma^2 \exp(-\rho d(C_i, C_j)) \\
\sigma \sim \mathcal{U}(0, b_\sigma) \\
\rho^{-1} \sim \mathcal{U}(a_\rho, b_\rho) \; .
\end{gathered} \tag{4.7}$$

Se ha de señalar que el modelo ampliado no corresponde estrictamente hablando a un proceso de Poisson ya que su función de verosimilitud no es de la forma

$$X|\lambda(t) \sim \prod_{x_i \in X} \lambda(x_i) \exp\left(-\int_C \lambda(t)dt\right) .$$

Sin embargo, como hemos comprobado que ambas formulaciones son equivalentes, y teniendo en cuenta que el planteamiento ampliado presenta ventajas computacionales, a partir de ahora utilizaremos esta formulación. No obstante, conviene destacar que otros autores como Cappé et al. (2003) [24] hacen inferencia desde el marco bayesiano sin recurrir a la formulación ampliada del modelo, por lo que la utilización de variables auxiliares no resulta estrictamente necesaria en este tipo de formulaciones.

A diferencia de la formulación de Lawson y Clark, nuestra propuesta tiene en cuenta el término log-gaussiano en la integral del proceso de Poisson. Estos autores ignoran el papel de los efectos aleatorios sobre la integral de la función de intensidad, por lo que en su trabajo dicha integral coincide con la integral de la mixtura únicamente. En nuestra opinión, el término de heterogeneidad espacial no habría de ignorarse a la hora de evaluar la integral del proceso de Poisson puesto que influye en el riesgo atribuible a cada localización de la región de estudio y por tanto sobre la superficie de riesgo.

4.3. Simulación MCMC de la propuesta univariante

Una vez hemos modelizado para el caso univariante, procedemos a describir el algoritmo de simulación que proponemos para llevar a cabo su inferencia. Dicho algoritmo estará basado en cadenas de Markov, como sue-

le ser habitual en la inferencia en modelos jerárquicos. Además habremos de hacer uso de métodos de simulación trans-dimensional para simular los parámetros de la mixtura, al ser su número de componentes indeterminado.

A partir de aquí denotaremos por $\lambda_j(t) = \exp(f(t)) \cdot \mathcal{N}(t|\mu_j, \tau_j^{-1})$ a la combinación de la j-ésima componente de la mixtura y la parte log-gaussiana del modelo. Literalmente, la función de intensidad se habría de denotar $\lambda_j(t|\mu, \tau, f(.))$, ya que ésta depende de todos estos parámetros. Sin embargo, para facilitar la notación, generalmente optaremos por la expresión abreviada, recurriendo a la notación extendida sólo donde se considere necesario.

4.3.1. Simulación de las variables de la mixtura

En primer lugar describimos la simulación de los parámetros de la mixtura, comenzando por los movimientos de tipo intra-dimensional en los que se simula las distribuciones de las medias, las precisiones y los pesos de cada componente, las variables auxiliares del modelo completo y el parámetro β. Posteriormente, se describirán los movimientos de tipo trans-dimensional de nacimiento-muerte y división-combinación que se utilizarán también en la simulación del modelo.

Simulación de μ_j

En primer lugar vamos a obtener la expresión de la distribución posterior de μ_j:

$$P(\mu_j|...) \propto P(x_i|\mu_j, ...)P(\mu_j)$$

$$\propto \left(\prod_{\{x_i : Z_{x_i}=j\}} \lambda_j(x_i) \right) \exp\left(-\sum_{k=1}^{m} w_k \int_C \lambda_k(t)dt \right) \cdot \mathcal{N}(\mu_j|\xi,\kappa)$$

$$\propto \mathcal{N}\left(\mu_j | \frac{\tau_j \sum_{Z_i=j} x_i + \kappa\xi}{\tau_j n_j + \kappa}, (\tau_j n_j + \kappa)^{-1} \right) \exp\left(-\sum_{k=1}^{m} w_k \int_C \lambda_k(t)dt \right) .$$

Observar que las distribuciones posteriores de estos parámetros constan de dos términos, el primero coincide con la función de densidad que se empleaba en el algoritmo *RG* para muestrear este parámetro mediante Gibbs Sampling, el segundo valora la interacción del término de mixtura y el log-gaussiano sobre la integral de la función de intensidad. Este segundo término impide que podamos utilizar Gibbs Sampling en nuestro proceso de simulación ya que dificulta la generación de valores directamente de la distribución posterior. Sin embargo, la primera parte de la expresión anterior nos puede valer como función de propuesta para generar valores en un mecanismo de muestreo mediante Metropolis-Hastings. Dicha función se utilizaba en el algoritmo de Richardson y Green produciendo unos parámetros de convergencia bastante aceptables. El segundo término de la distribución posterior de este parámetro se utilizará para el cálculo de la probabilidad de aceptación del nuevo valor propuesto. En el caso que las probabilidades de aceptación sean razonablemente altas el proceso de Metropolis-Hastings utilizado se comportará de forma similar al proceso de Gibbs Sampling empleado en *RG*.

Así, si denotamos por $\mu^* = (\mu_1, ..., \mu_{k-1}, \mu_k^*, \mu_{k+1}, ...\mu_m)$ el vector de medias donde la k-ésima componente ha sido generada de la forma que acabamos de describir y el resto de componentes toma el mismo valor que en el paso anterior de la cadena de Markov, la probabilidad de aceptación en un movimiento de Metropolis-Hastings, resulta el mínimo entre 1 y la

siguiente expresión:

$$\frac{P(\mu^*|...)\mathcal{N}(\mu_j|\frac{\tau_j\sum_{Z_i=j}x_i+\kappa\xi}{\tau_j n_j+\kappa},(\tau_j n_j+\kappa)^{-1})}{P(\mu|...)\mathcal{N}(\mu_j^*|\frac{\tau_j\sum_{Z_i=j}x_i+\kappa\xi}{\tau_j n_j+\kappa},(\tau_j n_j+\kappa)^{-1})}=$$

$$\frac{\exp(-\sum_{k=1}^m w_k\int_C\lambda_k(t|\mu^*,...)dt)}{\exp(-\sum_{k=1}^m w_k\int_C\lambda_k(t|\mu,...)dt)}=\exp\left(-w_j\int_C(\lambda_j(t|\mu_j^*,...)-\lambda_j(t|\mu_j,...))dt\right)\ .$$

Simulación de τ_j

La distribución posterior de este parámetro toma una expresión similar a la obtenida para las medias de las componentes de la mixtura. Ésta es

$$P(\tau_j|...)\propto\exp\left(-\sum_{k=1}^m w_k\int_C\lambda_k(t|\tau)dt\right)\Gamma\left(\alpha+\frac{n_j}{2},\beta+\frac{\sum_{\{x_i:Z_{x_i}=j\}}(x_i-\mu_j)^2}{2}\right)\ .$$

Al igual que en el muestreo de las medias, utilizaremos la segunda componente de la distribución posterior como función de propuesta, en ese caso la probabilidad de aceptación del nuevo valor resulta:

$$\exp\left(-w_j\int_C(\lambda_j(t|\tau_j^*,...)-\lambda_j(t|\tau_j,...))dt\right)\ .$$

Simulación de Z_i

En este caso la distribución posterior viene dada por

$$P(Z_i=j|...)\propto P(X|Z_i=j,...)P(Z_i=j)\propto\mathcal{N}(x_i|\mu_j,\tau_j^{-1})w_j\ ,$$

la misma expresión que en el modelo *RG*, por lo que podremos muestrear esta variable de la misma forma que se hacía en aquel, mediante Gibbs sampling. Concretamente, generaremos los nuevos valores a partir de una distribución multinomial donde los parámetros de dicha distribución vienen dados por la expresión anterior.

Simulación de w

Al igual que para las variables auxiliares Z, resulta sencillo comprobar que la distribución posterior de w es la misma que la utilizada en el modelo *RG*. Así pues podremos muestrear este parámetro de la misma forma que se hacía en el modelo anterior, mediante Gibbs Sampling generando valores a partir de

$$w \sim \text{Dir}(n_1 + \delta, ..., n_2 + \delta) \ ,$$

donde n_j es el número de observaciones asignadas a la j-ésima componente de la mixtura.

Simulación de β

En este caso también se podrá muestrear mediante Gibbs Sampling de la misma forma que proponían Richardson y Green en su algoritmo, es decir:

$$\beta \sim \Gamma\left(g + m\alpha, h + \sum_{i=1}^{m} \tau_i\right) \ .$$

Movimiento de división-combinación

Para el movimiento de división elegiremos al azar una de las componentes de la mixtura y la dividiremos de la misma forma que se hace en el algoritmo *RG*. Por tanto si $w_{j^*}, \mu_{j^*}, \tau_{j^*}$ son los parámetros correspondientes a la componente que se va a dividir propondremos

$$w_{j1} = w_{j^*} u_1 \ , \qquad w_{j2} = w_{j^*}(1 - u_1) \ ,$$

$$\mu_{j1} = \mu_{j^*} - u_2 \sigma_{j^*} \sqrt{\frac{w_{j2}}{w_{j1}}} \ , \qquad \mu_{j2} = \mu_{j^*} + u_2 \sigma_{j^*} \sqrt{\frac{w_{j1}}{w_{j2}}} \ ,$$

$$\sigma_{j1}^2 = u_3(1-u_2^2)\sigma_{j^*}^2\frac{w_{j*}}{w_{j1}}\,, \qquad \sigma_{j2}^2 = (1-u_3)(1-u_2^2)\sigma_{j^*}^2\frac{w_{j*}}{w_{j2}}\,,$$

como parámetros de las nuevas componentes de la mixtura. Las variables auxiliares empleadas en las expresiones anteriores se muestrean de las siguientes distribuciones $u_1 \sim \mathcal{B}(2,2)$, $u_2 \sim \mathcal{B}(2,2)$, $u_3 \sim \mathcal{B}(1,1)$. La probabilidad de aceptación de la nueva propuesta coincide exactamente con la probabilidad de aceptación del algoritmo *RG*, salvo el término del cociente de verosimilitudes que sí que cambiará en este caso. Por lo demás el cociente de distribuciones iniciales, el cociente de propuestas y el jacobiano de las transformaciones (son las mismas transformaciones en ambos algoritmos) no varían del algoritmo *RG* al que estamos proponiendo. Respecto al cociente de verosimilitudes tenemos

$$\frac{\prod_{x_i\in X}\lambda(x_i|w^*,\mu^*,\tau^*)\exp(-\int_C\lambda(t|w^*,\mu^*,\tau^*)dt)}{\prod_{x_i\in X}\lambda(x_i|w,\mu,\tau)\exp(-\int_C\lambda(t|w,\mu,\tau)dt)} =$$

$$= \frac{\prod_{x_i\in X}\mathrm{mix}(x_i|w^*,\mu^*,\tau^*)\exp(-\int_C\lambda(t|w^*,\mu^*,\tau^*)dt)}{\prod_{x_i\in X}\mathrm{mix}(x_i|w,\mu,\tau)\exp(-\int_C\lambda(t|w,\mu,\tau)dt)}\,,$$

donde *mix* es la parte correspondiente a la mixtura dentro de la función de intensidad. En la expresión final del cociente de verosimilitudes ha desaparecido la parte log-gaussiana de la función de intensidad ya que ésta se anula entre el numerador y el denominador. El cociente de mixturas coincide con el cociente de verosimilitudes del modelo de Green, por tanto para el cálculo de la probabilidad de aceptación del nuevo algoritmo únicamente tendremos que multiplicar la probabilidad de aceptación del modelo *RG* por

$$\exp\left(-\int_C(\lambda(t|w^*,\mu^*,\tau^*)-\lambda(t|w,\mu,\tau))dt\right)\,.$$

Resulta reseñable el paralelismo entre el proceso de simulación de las medias y precisiones de cada componente de la mixtura y el movimiento de nacimiento y muerte. En los tres procedimientos la función de propuesta es la misma que la propuesta en el algoritmo *RG* y, también en los tres, la probabilidad de aceptación viene dada por la probabilidad de ese movimiento

en el modelo sin componente log-gaussiana multiplicada por la exponencial de la resta entre la nueva integral y la antigua de la función de intensidad.

Respecto al movimiento de combinación, procederemos de la misma forma que se procedía en el algoritmo de *RG*. Elegimos 2 componentes al azar y las combinamos según las igualdades (3.10). De la misma forma que para el proceso de división, la probabilidad de aceptación de este movimiento será la misma que en el algoritmo de Richardson y Green multiplicada por

$$\exp\left(-\int_C(\lambda(t|w^*,\mu^*,\tau^*)-\lambda(t|w,\mu,\tau))dt\right) .$$

Movimiento de Nacimiento-Muerte

Para este movimiento también se procede de forma análoga a la propuesta en *RG*. En el caso de un nacimiento generaremos los valores de la nueva componente según

$$w_{j^*}\sim\mathcal{B}(1,k)\quad \mu_{j^*}\sim\mathcal{N}(\xi,\kappa^{-1})\quad \sigma_{j^*}^{-2}\sim\Gamma(\alpha,\beta) ,$$

y reescalaremos las demás componentes de w según

$$w_j'=w_j(1-w_{j^*}) .$$

En este caso la probabilidad de aceptación coincidirá exactamente con el algoritmo de Richardson y Green, ya que en dicha ocasión no influía el cociente de verosimilitudes sobre la probabilidad y tanto el cociente de distribuciones iniciales, el cociente de probabilidades de propuesta y el jacobiano de la transformación coinciden en ambos modelos. El movimiento de muerte de una componente también se realiza de manera análoga a la del modelo *RG*.

4.3.2. Simulación del proceso log-gaussiano

En la presente sección detallaremos el mecanismo de simulación de las variables del proceso log-gaussiano sobre las celdas del grid en las que se divide la región de estudio. La simulación de cada una de estas variables por separado conlleva una convergencia pobre del proceso, ya que para la aceptación de cada una de ellas la verosimilitud sólo varía ligeramente, por lo que se aceptará en general casi todos los valores propuestos. Así pues tendremos que recurrir a algún mecanismo de simulación conjunta para todas las celdas que componen el grid con el objetivo de mejorar los parámetros de convergencia de la simulación. Para ello haremos uso de simulación mediante el método de Langevin-Hastings (Roberts y Tweedie, 1996 [84]; Besag, 1994 [13]), tal y como se propone en Möller et al. (1998) [69].

La simulación mediante el método de Langevin-Hastings se basa en la simulación conjunta mediante Metrópolis-Hastings de un vector de parámetros. Así, los nuevos valores de la cadena de Markov son generados mediante un proceso de aceptación-rechazo en el que la función de propuesta de los nuevos valores toma la siguiente expresión

$$\mathcal{N}_d(\gamma + (h/2)\nabla(\log(p(\gamma|...)), hI_d) \ ,$$

donde γ es el valor actual de la cadena que se desea muestrear, d es la dimensión de γ, es decir, el número de variables que se muestrean simultáneamente, y h un parámetro de movilidad de la cadena fijado por el usuario para conseguir el umbral de aceptación deseado. En la expresión anterior ∇ denota el gradiente del vector al que precede.

El problema que tiene la simulación conjunta de las variables del proceso log-gaussiano es que la propuesta de nuevos valores, con probabilidades de aceptación razonables, no resulta sencillo. Sin embargo, el mecanismo de

Langevin-Hastings proporciona un criterio de generación de valores guiado por el gradiente de la función de verosimilitud, de forma que la probabilidad de aceptación de los valores propuestos sea alta. Este mecanismo de simulación resulta más eficiente cuanto mayor es d, la dimensión del conjunto de parámetros simulado (Christensen et al., 2001 [29]), por lo que resulta especialmente apropiado para nuestro caso, en el que el número de variables que se generan conjuntamente suele ser bastante elevado. Además, en Möller et al. (1998) [69] se demuestra que para la simulación de procesos log-gaussianos la versión truncada del algoritmo de Langevin-Hastings es geométricamente ergódica (propiedad deseable de la cadena de Markov que asegura su convergencia a la distribución posterior como función geométrica del número de iteraciones). Es por ello que creemos que este tipo de simulación también será apropiada para la simulación del modelo propuesto.

De aquí en adelante vamos a considerar la siguiente descomposición del vector de variables del proceso log-gaussiano asociado a cada celda

$$f_i = (\phi + \Sigma\Gamma)_i \ ,$$

donde Γ es un vector con tantas variables como se desee generar, que sigue una distribución normal de media 0 con todos los términos independientes y de varianza 1 para todas sus componentes. Mientras, Σ será la raíz cuadrada de la matriz de varianzas-covarianzas del proceso, es decir $\Sigma^2 = \Sigma \cdot \Sigma$. Resulta obvio que $f = \{f_1, ..., f_{|C|}\}$ definido de esta forma tiene una distribución $\mathcal{N}_{|C|}(\phi, \Sigma^2)$ tal y como se había establecido en el modelo. Consideraremos un modelo exponencial para la matriz de varianzas-covarianzas, aunque resulta posible la elección de otras funciones de correlación. Por tanto la matriz de varianzas-covarianzas, Σ^2 y su raíz cuadrada Σ dependerán de σ y ρ los parámetros del modelo exponencial.

Dada esta descomposición del vector f_i la formulación del modelo am-

pliado (4.7) resulta:

$$\begin{gathered} X|Z,\mu,\tau,w,f,\lambda(t) \sim \left(\prod_{x_i\in X}\exp(f(x_i))\mathcal{N}(x_i|\mu_{Z_i},\tau_{Z_i}^{-1})\right)\exp(-\textstyle\int_C \lambda(t)dt) \\ Z_i \sim \mathrm{Mn}(w) \;\; i=1,...,|X| \\ \lambda(t) = \exp(f(t))\textstyle\sum_{j=1}^m w_j\mathcal{N}(t|\mu_j,\tau_j^{-1}) \\ \mu_j \sim \mathcal{N}(\xi,\kappa^{-1}) \;\; j=1,...,m \\ \tau_j \sim \Gamma(\alpha,\beta) \;\; j=1,...,m \\ w \sim \mathrm{Dir}(\delta,...,\delta) \\ m \sim \mathcal{U}(0,M) \end{gathered} \tag{4.8}$$

$$\begin{gathered} f(\{C_1,...C_K\}) = \phi + \Sigma(\sigma,\rho)\Gamma \\ \phi \sim \mathcal{U}(-\infty,\infty) \\ \Sigma^2_{\{i,j\}}(\sigma,\rho) = \sigma^2\exp(-\rho d(C_i,C_j)) \\ \sigma \sim \mathcal{U}(0,b_\sigma) \\ \rho^{-1} \sim \mathcal{U}(a_\rho,b_\rho) \\ \Gamma_i \sim \mathcal{N}(0,1) \;\; i=1,...,|C| \; . \end{gathered} \tag{4.9}$$

Pasamos seguidamente a describir el proceso de simulación de los parámetros que componen f_i.

Simulación de ϕ

La distribución posterior de este parámetro toma la siguiente expresión:

$$P(\phi|...) = P(Y|\phi,...)P(\phi) \propto \left(\prod_{x_i\in X}\exp(\phi)\right)\exp\left(-\int_C \lambda(t|\phi)dt\right) =$$

$$= \exp\left(n\phi - \int_C \lambda(t|\phi)dt\right) \; .$$

Para la simulación de este parámetro hemos optado por un algoritmo de caminata aleatoria mediante Metrópolis-Hastings. La función de propuesta

elegida para dicha caminata es una normal centrada en el estado actual de la cadena. En este caso, la probabilidad de aceptación del nuevo valor vendrá dada por:

$$\exp\left(n(\phi^* - \phi) - \int_C (\lambda(t|\phi^*) - \lambda(t|\phi))dt\right) .$$

Simulación de Γ

Para la simulación de Γ proponemos usar el algoritmo de Langevin-Hastings descrito anteriormente. En primer lugar, calcularemos la distribución posterior de este vector. Dicha función será útil tanto para el cálculo de la probabilidad de aceptación del nuevo vector propuesto como para la realización de la propuesta, ya que ésta se basa en el gradiente del logaritmo de esta distribución posterior. Notar que Σ es una matriz de dimensión $|C| \times |C|$ y que el producto de su i-ésima fila por Γ determina el valor de la i-ésima celda. Denotaremos por Σ_x a la fila de Σ correspondiente a la celda que contiene al valor x.

La distribución posterior de Γ tiene la siguiente expresión:

$$p(\Gamma|...) \propto p(X|\Gamma, ...)p(\Gamma) \propto \left(\prod_{x_i \in X} \lambda(x_i)\right) \exp\left(-\int_C \lambda(t)dt\right) \exp(-\|\Gamma\|^2/2) \propto$$

$$\prod_{x_i \in X} \exp(\Sigma_{x_i}\Gamma) \exp\left(-\int_C \lambda(t)dt - \|\Gamma\|^2/2\right) ,$$

teniendo en cuenta que $\Sigma_{x_i} = e_{x_i}\Sigma$ donde e_{x_i} es un vector en el que todas sus casillas valen 0, salvo la que contiene a x_i que toma el valor 1, la expresión anterior resulta:

$$\exp\left(\sum_{i=1}^{n} e_{x_i}\Sigma\Gamma - \|\Gamma\|^2/2 - \int_C \lambda(t)dt\right)$$

$$= \exp\left(\hat{n}\Sigma\Gamma - \|\Gamma\|^2/2 - \int_C \lambda(t)dt\right) , \tag{4.10}$$

donde $\hat{n}$ será un vector en el que cada componente cuenta el número de observaciones que se han dado dentro de cada celda. Una vez se ha determinado la probabilidad posterior de Γ podemos calcular el término que guiará la nueva propuesta en el algoritmo de Langevin. Para el cálculo de dicho factor se ha de tener en cuenta que $\lambda(t)$ depende de Γ de la siguiente forma:

$$\lambda(t) = \exp(\phi + \Sigma_t\Gamma)\sum_{j=1}^{m} w_j \mathcal{N}(t|\mu_j, \tau_j^{-1}) ,$$

por tanto

$$\int_C \lambda(t) = \sum_{j=1}^{m} w_j \sum_{k=1}^{|C|} \int_{C_k} \lambda_j(t)dt =$$

$$= \sum_{k=1}^{|C|} \exp(\phi + \Sigma_k\Gamma)\left(\sum_{j=1}^{m} w_j \int_{C_k} \mathcal{N}(t|\mu_j, \tau_j^{-1})dt\right) ,$$

ya que para cada celda C_i el proceso log-gaussiano toma un valor constante $\exp(\phi + \Sigma_i\Gamma)$. Así la derivada respecto a Γ_i de la integral de la superficie de intensidad resulta:

$$\frac{\partial \int_C \lambda(t)dt}{\partial \Gamma_i} = \sum_{k=1}^{|C|} \exp(\phi + \Sigma_k\Gamma)\left(\sum_{j=1}^{m} w_j \int_{C_k} \mathcal{N}(t|\mu_j, \tau_j^{-1})dt\right)\Sigma_{ki} .$$

Teniendo en cuenta este hecho el jacobiano de la distribución tiene la forma:

$$(\nabla ln(p(\Gamma|y)))_i = -\Gamma_i + (\hat{n}\Sigma)_i - \sum_{j=1}^{m} w_j \sum_{k=1}^{|C|}\left(\int_{C_k} \lambda_j(t)dt\right)\Sigma_{ki} =$$

$$= -\Gamma_i + (\hat{n}\Sigma)_i - \sum_{k=1}^{|C|}\left(\int_{C_k} \lambda(t)dt\right)\Sigma_{ki} ,$$

por lo que, si denotamos por

$$e = \left(\widehat{n_k} - \int_{C_k} \lambda(t)dt\right)_{k=1}^{|C|} ,$$

al vector de residuos del ajuste de la mixtura en cada celda del grid, entonces podemos expresar el jacobiano de la función de propuesta de la siguiente forma:

$$\nabla ln(p(\Gamma|y)) = -\Gamma + e\Sigma \ .$$

De esta manera, la función de propuesta del algoritmo de Langevin-Hastings vendrá dada por la siguiente expresión:

$$\mathcal{N}_{|C|}(\Gamma^*|(1-h/2)\Gamma + (h/2)e\Sigma, hI_{|C|}) \ ,$$

donde $I_{|C|}$ denota la matriz identidad de dimensión $|C| \times |C|$, y la probabilidad de aceptación del valor propuesto vendrá dado por:

$$\frac{P(\Gamma^*|...)\mathcal{N}_{|C|}(\Gamma|(1-h/2)\Gamma^* + (h/2)e^*\Sigma, hI_{|C|})}{P(\Gamma|...)\mathcal{N}_{|C|}(\Gamma^*|(1-h/2)\Gamma + (h/2)e\Sigma, hI_{|C|})} \ ,$$

donde $P(\Gamma^*|...)$ toma la expresión (4.10).

Simulación de Σ

Según se ha comentado Σ es la raíz cuadrada de la matriz de varianzas-covarianzas a la que se ha dado estructura exponencial. Dicha matriz dependerá de ρ y σ, por lo que para la simulación de Σ se habrán de muestrear estas dos variables, a partir de las cuales se calculará la matriz de covarianzas entre las celdas y su raíz cuadrada. El cálculo de la raíz cuadrada de una matriz de las dimensiones que manejamos es un procedimiento relativamente costoso, sobre todo si se ha de repetir un gran número de veces en un proceso de simulación MCMC.

Para hacer este cálculo computacionalmente menos costoso precalcularemos Σ para 100 valores distintos de ρ y consideraremos una distribución inicial uniforme discreta para este parámetro. De esta forma, nos evitaremos

realizar la descomposición de Cholesky de la matriz de varianzas-covarianzas para el cálculo de Σ en cadena iteración del algoritmo. En Möller et al. (1998) [69] y Möller y Waagepetersen (2004) [70] se propone la utilización del método de "*Circulant Embedding*" para el cálculo de la raíz cuadrada de la matriz de covarianzas en cadena iteración. Dicho método hace uso de transformadas de Fourier y la representación espectral de la función de correlación entre las celdas. Se puede encontrar más información sobre este algoritmo, por ejemplo en Schlather (1999) [86]. Hemos optado por no utilizar dicho método en nuestra propuesta por no añadir más complejidad a su formulación, pero reconocemos que nuestro algoritmo consiste únicamente en una aproximación discreta del resultado que se obtendría mediante *Circulant embedding*.

Procedemos a detallar la simulación de ρ y σ que tal y como se ha descrito definen la matriz Σ.

Simulación de σ

Se ha tomado como distribución inicial para este parámetro una distribución uniforme entre 0 y b_σ, resultando la distribución posterior:

$$P(\sigma|...) \propto P(X|\sigma,....)P(\sigma) = \left(\prod_{x_i \in X} \lambda(x_i)\right) \exp\left(-\int_C \lambda(t|\sigma)dt\right) \cdot 1_{[0,b_\sigma]}(\sigma) \propto$$

$$\propto \exp\left(\sigma \hat{n} \exp(-\rho D)^{1/2}\Gamma - \int_C \lambda(t|\sigma)dt\right) \cdot 1_{[0,b_\sigma]}(\sigma) \ .$$

Donde D es la matriz de distancias entre las celdas del grid en las que se define el proceso log-gaussiano. Como no resulta sencillo muestrear valores directamente de la distribución posterior, utilizaremos Metropolis-Hastings para simular esta variable. Así, si se utiliza una caminata aleatoria con

función de propuesta gaussiana, la probabilidad de aceptación vendrá dada por la siguiente expresión

$$\exp\left((\sigma^* - \sigma)\left(\hat{n}\exp(-\rho D)^{1/2}\Gamma\right) - \int_C (\lambda(t|\sigma^*) - \lambda(t|\sigma))dt\right) 1_{[0,b_\sigma]}(\sigma^*) \ ,$$

ya que el cociente de las probabilidades de la nueva propuesta partiendo del estado actual y viceversa se cancelan entre sí.

Simulación de ρ

Tal y como se ha descrito propondremos una distribución inicial uniforme discreta entre -1 y 1 para ρ. En ese caso la distribución posterior tomará la siguiente expresión:

$$P(\rho|...) \propto P(X|\rho, ...)P(\rho) = \left(\prod_{x_i \in X} \lambda(x_i)\right) \exp\left(-\int_C \lambda(t|\rho)dt\right) 1_{[-1,1]}(\rho) \propto$$

$$\propto \exp\left(\sigma\hat{n}\exp(-\rho D)^{1/2}\Gamma - \int_C \lambda(t|\rho)dt\right) 1_{[-1,1]}(\rho) \ .$$

Para la simulación de esta variable recurriremos nuevamente al procedimiento de Metropolis-Hastings. Para ello utilizaremos como función de propuesta también una distribución uniforme discreta sobre el conjunto de valores contemplado para ρ. En ese caso, la probabilidad de aceptación resulta:

$$\exp\left(\sigma\hat{n}(\exp(-\rho^* D)^{1/2} - \exp(-\rho D)^{1/2})\Gamma - \int_C (\lambda(t|\rho^*) - \lambda(t|\rho))dt\right) \ ,$$

donde $\exp(-\rho^* D)^{1/2}$ corresponde a una de las 100 matrices precalculadas de antemano.

4.4. Modelización bidimensional del proceso

Una vez hemos descrito la versión unidimensional de la modelización pasamos a describir la versión bidimensional. Esta versión supone una sencilla adaptación de las ideas expuestas en el caso unidimensional, por lo que no nos vamos a detener en exceso en su desarrollo. En Stephens (1999) [91], Pérez y Berger (1999) [76] y Dellaportas y Papageorgiou (2004) [35], entre otros, se han utilizado modelos de mixturas multidimensionales con un número indeterminado de componentes. La primera de estas referencias desarrolla una adaptación bidimensional de la propuesta de Richardson y Green [80]. Haremos uso de dicha adaptación en nuestra propuesta aunque el algoritmo de simulación que vamos a emplear no será el mismo. Stephens utiliza un algoritmo de simulación trans-dimensional basado en un proceso continuo de nacimiento-muerte, mientras que nosotros haremos uso de simulación de salto reversible.

Respecto a la componente log-gaussiana, en el trabajo seminal de esta técnica, Möller et al. (1998) [69] ya establecen la versión bidimensional de este tipo de modelización. Por tanto, la inclusión de esta componente en nuestra propuesta seguirá las líneas propuestas en dicho trabajo y que ya han sido utilizadas en el caso unidimensional.

En cuanto a la notación empleada en esta versión, denotaremos por X el conjunto de observaciones del proceso, así cada observación x^* constará de 2 componentes (x_1^*, x_2^*). Nuevamente consideraremos el proceso puntual que ha generado las observaciones como un proceso de Poisson, por lo que la función de verosimilitud de las observaciones vendrá dada por

$$X|Z,\mu,\Lambda,w,f,\lambda(t) \sim \left(\prod_{x_i\in X} \exp(f(x_i))\mathcal{N}_2(x_i|\mu_{Z_i},\Lambda^2_{Z_i})\right)\exp\left(-\int_C \lambda(t)dt\right) ,$$

donde λ es una superficie de intensidad sobre un espacio bidimensional compuesta por el producto de un proceso de mixturas y otro log-gaussiano:

$$\lambda(t) = \exp(f(t)) \sum_{j=1}^{m} w_j \mathcal{N}_2(t|\mu_j, \Lambda_j^2) \ .$$

Nuevamente haremos uso de la formulación completa del modelo de mixturas, así Z_i serán las variables auxiliares que asignan cada observación a una componente de la mixtura

$$Z_i \sim \text{Mn}(w) \ \ i = 1, ..., |X| \ .$$

En cuanto a la distribución inicial de la media de cada componente de la mixtura, utilizaremos una distribución normal bivariante de forma similar al caso unidimensional:

$$\mu_j \sim \mathcal{N}_2(\xi, \kappa^{-1}) \ \ j = 1, ..., m \ ,$$

donde ξ se define como el baricentro de los eventos del patrón puntual y κ^{-1} en este caso es una matriz de varianzas-covarianzas diagonal. Los valores que hemos definido para dicha diagonal se corresponden con el cuadrado del rango de las coordenadas, en la primera y segunda dimensión respectivamente, de los eventos del patrón.

Definiremos la distribución de las matrices de precisión como la generalización bivariante de las distribuciones Gamma que empleábamos en el caso unidimensional. Por tanto vamos a emplear una distribución Wishart bivariante. En este caso, también se hará uso de una matriz de precisiones común β como valor promedio de todas las matrices, así la distribución de las matrices de precisión resulta:

$$\Lambda_j^{-2}|\beta \sim \mathcal{W}(2\alpha, (2\beta)^{-1}) \ j = 1, ..., m \ ,$$

donde el escalar α controla la similitud, a priori, que guardan las matrices de precisión de las distintas componentes. En Stephens (1999) [91] se propone

un valor para α de 3 que es el que tomaremos nosotros también. A su vez, la matriz β también se considera una variable en el proceso de inferencia y como tal tendrá su propia distribución inicial. Nuevamente vamos a seguir la propuesta de Stephens para esta variable y consideraremos que se distribuye según:

$$\beta \sim \mathcal{W}(2g, (2h)^{-1}) \ ,$$

donde se propone un valor para g de 0.3 y para h se propone

$$\begin{pmatrix} \frac{100g}{\alpha R_1^2} & 0 \\ 0 & \frac{100g}{\alpha R_2^2} \end{pmatrix} .$$

Los valores propuestos para g, h y α son ligeramente superiores a los utilizados en el caso unidimensional. Este hecho se debe a las recomendaciones de Stephens, que sugiere la utilización de valores superiores en el caso bidimensional, debido a la necesidad de imponer más intensamente el que las matrices de covarianza de las distintas componentes sean parecidas.

Por último, respecto a los pesos de las componentes de la mixtura y el número de éstas, hemos empleado la misma distribución que en la propuesta univariante, es decir

$$w \sim \mathrm{Dir}(\delta, ..., \delta) \ ,$$

$$m \sim \mathcal{U}(0, M) \ ,$$

donde hemos tomado también $\delta = 1$, al igual que en todos los artículos que conocemos que utilizan la modelización mediante mixturas con un número indeterminado de componentes.

Nuevamente para la definición del proceso log-gaussiano se requerirá la utilización de una partición de la región de estudio. Denotaremos por $C = \{C_{1,1}, .., C_{1,ny}, C_{2,1}, ..., C_{nx,ny}\}$ el grid a utilizar, donde nx será el número de

divisiones de C en la primera dimensión del espacio y ny el número de particiones en la segunda dimensión. La formulación del modelo log-gaussiano bidimensional y las distribuciones iniciales empleadas serán iguales que las empleadas en el caso unidimensional, concretamente:

$$\begin{gathered} f(C) = \phi + \Sigma(\sigma, \rho)\Gamma \\ \phi \sim \mathcal{U}(-\infty, \infty) \\ \Sigma^2_{\{i,j\}}(\sigma, \rho) = \sigma^2 \exp(-\rho d(C_i, C_j)) \\ \sigma \sim \mathcal{U}(0, b_\sigma) \\ \rho^{-1} \sim \mathcal{U}(a_\rho, b_\rho) \\ \Gamma_i \sim \mathcal{N}(0,1) \;\; i = 1, ..., |C| \;, \end{gathered} \tag{4.11}$$

donde a_ρ y b_ρ se han definido de la misma forma que en el caso unidimensional, garantizando que la correlación entre las celdas más cercanas es superior a 0.05 y la correlación entre las celdas centradas de C y las más extremas son inferiores a este valor.

4.4.1. Efecto frontera

Uno de los problemas que nos encontramos en el caso bidimensional es el del efecto frontera. Éste es consecuencia de que el modelo de mixturas está definido sobre todo el plano mientras que el proceso log-gaussiano, por limitaciones computacionales, únicamente puede ser definido sobre un conjunto finito del plano real. Así la adaptación de ambos procesos en un único marco requiere algunas precauciones que vamos a detallar a continuación.

Como solución a los problemas derivados de los diferentes dominios entre ambos procesos hemos utilizado distintas propuestas. La primera de

ellas consiste en definir el proceso de mixturas sobre el mismo dominio que el proceso log-gaussiano. En este caso hemos considerado una distribución uniforme sobre dicho dominio como distribución inicial de las medias de la mixtura. Sin embargo, dicha propuesta presenta distintos problemas. El primero es que sobre dicho dominio cualquier propuesta de nacimiento de una nueva componente de la mixtura tiene una verosimilitud elevada, por lo que el modelo ajusta un número de componentes en la mixtura bastante alto. Es decir, si ceñimos el dominio del modelo de mixturas al del proceso log-gaussiano, estaremos empleando una distribución inicial muy informativa sobre la ubicación de las medias, ya que estaremos obligando a que éstas se sitúen cerca de los datos aumentando el número de componentes de la mixtura. Pero éste no es el único problema que presenta esta propuesta ya que en ciertos patrones proporciona soluciones degeneradas, consecuencia de haber considerado un dominio finito. Esta solución degenerada se produce cuando una considerable proporción de los puntos del patrón se sitúan de forma más o menos alineada en torno a una recta que atraviesa la región de estudio. En esas ocasiones el modelo tiende a ajustar tal línea mediante una componente normal bivariante de forma muy alargada y de gran varianza. Dicha propuesta resulta muy verosímil al modelo ya que gran parte de dicha componente recae fuera del dominio del proceso log-gaussiano, por lo que su integral sobre C es pequeña resultando una elevada verosimilitud. Por otro lado, el resto de componentes de la mixtura tienden a imitar este comportamiento ya que de esta forma todas las matrices de varianzas-covarianzas del modelo son similares tal y como exige el modelo. Así, la solución degenerada de la que hablábamos consta de muchas componentes en la mixtura con matrices de varianza-covarianza similares y muy alargadas que ajustan el patrón puntual como un conjunto de bandas que atraviesan el dominio del proceso log-gaussiano. Esta solución resultará epidemiológicamente poco sostenible en la mayoría de los problemas planteados.

En consecuencia, la restricción del dominio del proceso de mixturas al dominio del proceso log-gaussiano presenta distintos problemas debidos a la utilización de un dominio finito. Por ello, hemos considerado más apropiado decantarnos por un dominio infinito para el modelo de mixturas y utilizar la misma distribución inicial normal para las medias de sus componentes que en el modelo *RG*. Sin embargo, dicha propuesta también presenta ciertos problemas y es que cuando algunas de las componentes queda vacía se actualiza en función exclusivamente de su distribución inicial, por lo que resulta habitual el que se propongan nuevas localizaciones bastante alejadas de los datos. Estas propuestas suelen ser aceptadas ya que de esa forma la componente vacía se traslada a zonas en las que la integral sobre el dominio del proceso log-gaussiano es prácticamente nula, por lo que la nueva localización resulta atractiva aunque no obedezca en absoluto al comportamiento de los datos. De esta forma, la simulación de este modelo producirá un gran número de componentes alejadas del dominio del proceso log-gaussiano y, por tanto, de los datos. Como resultado se generará una bolsa de componentes alejadas de los datos que producirá una sobreestimación del número de componentes de la mixtura.

Para solucionar este comportamiento vamos a considerar que la parte de clustering general del modelo propuesto está definida en todo el plano real: dentro del grid C seguirá un proceso log-gaussiano, tal y como habíamos considerado hasta este momento, mientras que fuera de este grid el proceso tomará un valor constante ω. En la figura 4.3 se ilustra esta idea. En ella se puede observar cómo a partir del patrón puntual se ha delimitado un grid de celdas C, determinado por el cuadrado central, en el que esta parte de la función de intensidad seguirá un proceso log-gaussiano mientras que fuera de dicho grid el proceso toma un valor constante.

Respecto a ω, un valor demasiado pequeño presentaría los mismos pro-

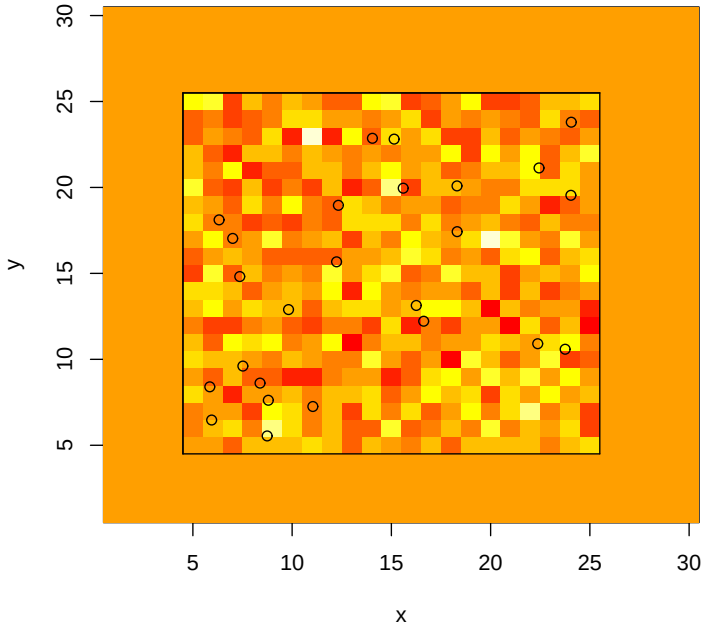

Figura 4.3: Proceso no paramétrico dentro y fuera del grid C.

blemas que se daban en el caso de considerar el proceso no paramétrico definido únicamente sobre C. Por el contrario, un valor demasiado grande haría que ninguna de las componentes de la mixtura se situara fuera de C ya que su integral penalizaría en exceso la función de verosimilitud. En ese caso observaríamos problemas similares a los que se daban en el caso de considerar una distribución inicial uniforme sobre C para las componentes de la mixtura. Además habría un problema añadido: las componentes también tenderían a situarse alejadas de la frontera para evitar que parte de su masa recaiga fuera de ésta (lo que aumentaría el valor de su integral) y esta repulsión respecto a la frontera podría condicionar la distribución posterior de la localización de las componentes. Por tanto, la elección del parámetro ω es de gran importancia para el correcto funcionamiento del algoritmo.

Así pues, resulta natural proponer un valor "neutro" para ω acorde con los valores que toma la función de intensidad en el interior del grid C. Teniendo en cuenta que la media de una distribución log-gaussiana coincide con la exponencial de la media del proceso gaussiano que la genera más la mitad de su varianza (ver, por ejemplo, Gelman et al., 2004 [48]), hemos utilizado como propuesta para ω

$$\omega = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \ .$$

De esta forma hacemos coincidir el valor promedio del proceso log-gaussiano dentro de C con el valor que toma ese proceso fuera del recinto. Así, si una componente de la mixtura tiene parte de su masa dentro de C y parte fuera, los valores esperados de sus integrales dentro y fuera de C serán proporcionales a la masa de la componente dentro y fuera de esta región, respectivamente. Con todo ello, esperamos evitar que se generen bolsas de componentes no explicativas fuera de C, ya que la integral de cualquier componente fuera de ella penalizará este tipo de situaciones.

Al ampliar el dominio del proceso puntual a todo el plano deberemos de tener la precaución de asegurar que la integral de la función de intensidad sigue siendo finita, ya que, en caso contrario, el modelo utilizado estaría mal formulado al depender su primera capa de esta integral. El proceso log-gaussiano utilizado en la función de intensidad produce únicamente valores finitos al estar definido sobre una región compacta, por lo que existirá un valor M tal que el valor de dicho proceso en todas las celdas será menor que M. La función de intensidad combina el modelo de mixturas con el proceso log-gaussiano, por tanto su integral cumplirá

$$\int \lambda(t)dt = \int \exp(f(t)) \sum_{j=1}^{m} w_j \mathcal{N}_2(t|\mu_j, \Sigma_j^2)dt \leq M \int \sum_{j=1}^{m} w_j \mathcal{N}_2(t|\mu_j, \Sigma_j^2)dt = M \ .$$

Así pues, la integral de la función de intensidad sobre todo el plano real es finita y por tanto el modelo utilizado está bien definido.

A partir de ahora cuando tengamos que evaluar la integral de la función de intensidad no nos deberemos ceñir exclusivamente a la integral sobre C sino que tendremos que tener en cuenta el valor de dicha integral en el exterior de esta región. La nueva integral tendrá la siguiente expresión:

$$\int \lambda(t)dt = \int_C \lambda(t)dt + \exp(\mu + \sigma^2/2) \int_{R^2 \setminus C} \sum_{j=1}^{m} w_j \mathcal{N}_2(t|\mu_j, \Sigma_j^2)dt \ .$$

Tal y como se podrá comprobar en las simulaciones realizadas en el próximo capítulo, parece que en el caso unidimensional no resulta tan problemático el efecto frontera. Pensamos que se puede deber a que en la versión unidimensional todas las componentes tienen una integral apreciable sobre C aunque se desplacen a regiones de poca verosimilitud a diferencia del caso bidimensional. Por tanto, en el caso unidimensional no hemos incorporado la integral de la función de intensidad fuera de C ya que en esta ocasión no hemos observado la necesidad de corregir este efecto.

4.5. Simulación del proceso bidimensional

En general la simulación del proceso bidimensional sigue las mismas líneas que la versión unidimensional. La diferencia más sustancial entre ambos algoritmos se da en los movimientos de simulación trans-dimensional. En el caso bidimensional únicamente utilizaremos el movimiento de nacimiento-muerte, ya que el movimiento de combinación-división conlleva una gran complejidad. Además, el movimiento de combinación-división no mejora sustancialmente la convergencia del algoritmo según se señala en Cappé et al. (2003) [24].

Pasamos seguidamente a describir el proceso de simulación de las variables del caso bidimensional. Conviene aclarar que ofreceremos menos detalles que en el caso unidimensional ya que los desarrollos en ambas situaciones son muy similares. La diferencia principal entre ambos algoritmos consiste en que en el caso unidimensional la integral de la superficie de intensidad no se va a ceñir únicamente a la región C sino que ésta se evaluará sobre todo el plano real.

4.5.1. Simulación de las variables de la mixtura

Simulación de μ_j

La distribución posterior de esta variable toma la siguiente expresión

$$P(\mu_j|...) \propto P(X|\mu_j,...)P(\mu_j) \propto \left(\prod_{\{x_i:Z_{x_i}=j\}} \mathcal{N}_2(x_i|\mu_j,\Lambda_j^2)\right) \times$$

$$\times \exp\left(-w_j \int \lambda_j(t)dt\right) \mathcal{N}_2(\mu_j|\xi,\kappa^{-1}) =$$

$$= \mathcal{N}_2 \left(\mu_j | \left(n_j \Lambda_j^{-2} + \kappa \right)^{-1} \cdot \left(\Lambda_j^{-2} (\sum_{\{i:Z_i=j\}} x_i) + \kappa \xi \right), \left(n_j \Lambda_j^{-2} + \kappa \right)^{-1} \right) \times$$

$$\times \exp \left(-w_j \int \lambda_j(t) dt \right) .$$

Así, al igual que para el caso unidimensional utilizaremos la primera parte de la expresión anterior para muestrear nuevos valores mientras que la segunda parte nos proporcionará el valor de la probabilidad de aceptación del valor muestreado, según un proceso de Metropolis-Hastings. Una vez muestreado el nuevo valor, la probabilidad de aceptación resulta:

$$\frac{\exp(-w_j \int \lambda_j(t|\mu_j^*) dt)}{\exp(-w_j \int \lambda_j(t|\mu_j) dt)} = \exp \left(-w_j \int (\lambda_j(t|\mu_j^*) - \lambda_j(t|\mu_j)) dt \right) .$$

La simulación de esta variable se realiza de forma muy similar al caso unidimensional, con la salvedad de la función de propuesta que ha sido adaptada al caso bidimensional.

Simulación de Λ_j^2

Procediendo de la misma forma que para μ_j utilizaremos como función de propuesta para Λ_j^{-2} parte de la distribución posterior de este parámetro, en concreto muestrearemos la matriz de precisión a partir de:

$$\mathcal{W} \left(2\alpha + n_j, (2\beta + \sum_{\{i:Z_i=j\}} (x_i - \mu_j)^2)^{-1} \right) .$$

En este caso, la probabilidad de aceptación también vendrá dada por:

$$\exp \left(-w_j \int (\lambda_j(t|(\Lambda_j^2)^*) - \lambda_j(t|\Lambda_j^2)) dt \right) .$$

Simulación de Z_i

La distribución posterior de este parámetro coincide con la distribución del caso unidimensional

$$P(Z_i = j|...) \propto P(X|Z_i = j, ...)P(Z_i = j) \propto \mathcal{N}(x_i|\mu_j, \Lambda_j^2)w_j \ .$$

Por tanto, dicha distribución sigue una multinomial donde la probabilidad de pertenecer a cada una de las m categorías viene dada por la expresión anterior normalizada, con j variando entre 1 y m. Así, podremos muestrear valores de esta variable mediante Gibbs Sampling.

Simulación de β

La distribución posterior de esta variable viene dada por la siguiente expresión

$$P(\beta|...) \propto \prod_{j=1}^{m} P(\Lambda_j^{-2}|\alpha, \beta)P(\beta) \ .$$

Teniendo en cuenta que todas las distribuciones implicadas en la distribución anterior son Wishart, tras algunos cálculos tenemos que la distribución posterior de β fijado el resto de parámetros sigue una distribución:

$$\beta \sim \Gamma\left(2g + 2k\alpha, (2h + 2\sum_i \Lambda_i^{-2})^{-1}\right) \ ,$$

por lo que también podremos emplear Gibbs Sampling también para muestrear esta variable.

Simulación de w

La distribución posterior de estas variables coincide con la del algoritmo unidimensional, por lo que podremos simularla de la misma forma que hacíamos en ese caso. En consecuencia, la simulación de esta variable se realizará mediante Gibbs Sampling muestreando nuevos valores a partir de

$$w \sim \text{Dir}(n_1 + \delta, ..., n_2 + \delta) \ ,$$

donde nuevamente n_j es el número de observaciones asignadas a la j-ésima componente de la mixtura.

Movimiento de nacimiento-muerte

En cuanto a la simulación del número de componentes de la mixtura, en el caso bidimensional vamos a emplear un único tipo de movimiento, el de nacimiento-muerte mediante simulación de salto reversible. En este sentido vamos a seguir las líneas propuestas en Cappé et al. (2003) [24] donde se propone no utilizar el movimiento trans-dimensional de división-combinación ya que aumenta la complejidad y "no mejora significativamente la precisión de los resultados". Se ha de señalar que otros autores (Pérez y Berger, 1999 [76]; Dellaportas y Papageorgiou, 2004 [35]; o Stephens, 1999 [91]) han propuesto otras alternativas de simulación para problemas similares al que nos ocupa. Sin embargo hemos adoptado la simulación mediante movimientos de nacimiento-muerte exclusivamente ya que nos ha parecido la alternativa más sencilla. Además, en los estudios computacionales de Cappé et al. (2003) [24] presentaba un comportamiento computacional incluso superior a la propuesta que incluye el movimiento de combinación-división.

Sin embargo la propuesta de movimiento de nacimiento-muerte propues-

ta por Richardson y Green (1997) [80] no resulta eficiente como único movimiento trans-dimensional ya que ésta genera componentes vacías y elimina únicamente este tipo de componentes. Así, este movimiento o bien genera componentes de poca relevancia, ya que mejorarán poco el ajuste al estar vacías, o bien habrá de esperar a que alguna componente quede vacía para poder eliminarla. Por estos motivos, el movimiento de nacimiento-muerte que vamos a utilizar en esta ocasión no va a funcionar exactamente de la misma forma que en el algoritmo de *RG*.

Para realizar el movimiento de muerte se elegirá al azar una de las componentes de la mixtura j^*, independientemente de que tenga individuos asignados o no, y se eliminará de la función de intensidad. Tras eliminar dicha componente, el peso de las restantes se reescalarán diviendolas por $(1-w_{j^*})$ de forma que tras el movimiento el peso de las componentes vuelva a sumar 1. Se ha de señalar que no es estrictamente necesario el que los individuos que pertenecían a la componente eliminada sean reasignados para calcular la probabilidad de aceptación del movimiento, de hecho no se han reasignado. Dicha probabilidad puede ser calculada atendiendo a la expresión de la función de intensidad que no hace uso de la versión ampliada del modelo (4.3), por tanto dicha función de intensidad no depende de las asignaciones de los individuos a las componentes. No obstante cualquier movimiento de nacimiento-muerte deberá preceder a un muestreo de dichas asignaciones ya que éstas si que intervienen en gran parte de los movimientos del proceso MCMC. Notar que en Cappé et al. (2003) [24] el algoritmo utilizado no hace uso en ningún momento de las variables auxiliares $\{Z_i : i = 1, ..., n\}$.

La probabilidad de aceptación del movimiento de muerte viene dado por el mínimo entre 1 y la siguiente expresión:

$$A = \frac{L(X|(w_j, \mu_j, \Lambda_j^2) \backslash (w_{j^*}, \mu_{j^*}, \Lambda_{j^*}^2)) P((w_j, \mu_j, \Lambda_j^2) \backslash (w_{j^*}, \mu_{j^*}, \Lambda_{j^*}^2))}{L(X|(w_j, \mu_j, \Lambda_j^2)) P((w_j, \mu_j, \Lambda_j^2))} \times$$

$$\times \frac{b((w_j, \mu_j, \Lambda_j^2) \backslash (w_{j^*}, \mu_{j^*}, \Lambda_{j^*}^2))}{d((w_j, \mu_j, \Lambda_j^2))} \times \frac{h(w_{j^*}, \mu_{j^*}, \Lambda_{j^*}^2)}{(1 - w_{j^*})^{(k-1)}} \ ,$$

donde el primer término corresponde al cociente de las distribuciones posteriores antes y después del movimiento. El segundo término representa el cociente entre la probabilidad de propuesta de un nacimiento tras la muerte y la probabilidad de haber propuesto una muerte en el actual movimiento. Por último el denominador del tercer término representa el jacobiano de la transformación practicada sobre los pesos de la mixtura, mientras que el numerador representa la función de densidad de los términos generados para llevar a cabo el movimiento inverso, el nacimiento de una nueva componente.

Para llevar a cabo el nacimiento de una nueva componente generaremos la media y la matriz de varianzas-covarianzas de ésta a partir de su distribución inicial, mientras que generaremos el peso de esta componente a partir de una distribución $\mathcal{B}(1, k)$. Tras la realización de un nacimiento, de la misma forma que tras una muerte, se llevará a cabo el muestreo de las asignaciones de individuos a componentes para que puedan ser utilizadas en el resto de pasos del algoritmo. La probabilidad de aceptación al realizar un nacimiento vendrá dada por el mínimo entre 1 y la siguiente expresión:

$$A^{-1} = \frac{L(X|(w, \mu, \Lambda^2) \bigcup (w^*, \mu^*, (\Lambda^2)^*)) P((w, \mu, \Lambda^2) \bigcup (w^*, \mu^*, (\Lambda^2)^*))}{L(X|(w, \mu, \Lambda^2)) P((w, \mu, \Lambda^2))} \times$$

$$\times \frac{d((w, \mu, \Lambda^2) \bigcup (w^*, \mu^*, (\Lambda^2)^*))}{b((w, \mu, \Lambda^2))} \times \frac{(1 - w^*)^{(k-1)}}{h(w^*, \mu^*, (\Lambda^2)^*)} \ .$$

Es decir, el inverso de la expresión utilizada para el movimiento de eliminación de una componente.

4.5.2. Simulación del proceso log-gaussiano

Al igual que en el caso unidimensional, haremos uso de la descomposición de la variable log-gaussiana según la siguiente expresión:

$$f(C_i) = (\phi + \Sigma(\rho, \sigma)\Gamma)_i \ ,$$

por lo que tendremos que simular las variables $\phi, \rho, \sigma, \Gamma$ como parte del proceso log-gaussiano. Pasamos seguidamente a describir el proceso de simulación de dichas variables.

Simulación de ϕ

La distribución posterior de este parámetro toma la misma forma que en el caso unidimensional, por lo que propondremos simularla de la misma manera. Para el algoritmo de Metropolis-Hastings utilizaremos una función de propuesta normal centrada en el estado actual de la cadena y la desviación típica se sintonizará de forma que la probabilidad de aceptación tome un valor que se considere adecuado. Tras proponer el nuevo valor para ϕ, la probabilidad de aceptación para éste viene dada por:

$$\exp\left(n(\phi^* - \phi) - \int(\lambda(t|\phi^*) - \lambda(t|\phi))dt\right) \ .$$

Simulación de σ

Nuevamente la distribución posterior de este parámetro coincide con la del caso unidimensional, así podremos simular esta variable de la misma forma. Utilizaremos el algoritmo de Metropolis-Hastings con función de propuesta gaussiana centrada en el estado actual. En ese caso la probabilidad

de aceptación viene dada por:

$$\exp\left(\left(\hat{n}\exp(-\rho D)^{1/2}\Gamma\right)(\sigma^* - \sigma) - \int(\lambda(t|\sigma^*) - \lambda(t|\sigma))dt\right) 1_{[0,b_\sigma]}(\sigma^*) \ .$$

Simulación de ρ

Para la simulación de ρ en el caso bidimensional se hace más necesaria todavía la precomputación de la raíz cuadrada de la matriz de varianzas-covarianzas $\left(\exp(-\rho D)^{1/2}\right)$, ya que en este caso los grids utilizados suelen tener un número mayor de celdas que en el caso unidimensional. También realizaremos la simulación de esta variable de la misma forma que en el caso unidimensional, es decir, utilizaremos como función de propuesta una distribución uniforme discreta sobre el rango [-1,1] y la probabilidad de aceptación del nuevo valor vendrá dado por la siguiente expresión:

$$\exp\left(\sigma\hat{n}(\exp(-\rho^* D)^{1/2} - \exp(-\rho D)^{1/2})\Gamma - \int(\lambda(t|\rho^*) - \lambda(t|\rho))dt\right) \ .$$

Simulación de Γ

La simulación de Γ también seguirá las mismas líneas que el caso unidimensional, ya que su distribución posterior coincide en ambas ocasiones. Por tanto, emplearemos también la simulación de Langevin-Hastings en este caso. La función de propuesta vendrá dada por:

$$\mathcal{N}_{|C|}(\Gamma^*|(1 - h/2)\Gamma + (h/2)e\Sigma, hI_{|C|}) \ ,$$

donde h es un parámetro de sintonización del algoritmo, I denota la matriz identidad y e toma la siguiente expresión

$$e = \left(\hat{n_k} - \int_{C_k} \lambda(t)dt\right)_{k=1}^{|C|} \ .$$

La probabilidad de aceptación del nuevo valor resulta:

$$\frac{P(\Gamma^*|...)\mathcal{N}_{|C|}(\Gamma|(1-h/2)\Gamma^* + (h/2)e^*\Sigma, hI_{|C|})}{P(\Gamma|...)\mathcal{N}_{|C|}(\Gamma^*|(1-h/2)\Gamma + (h/2)e\Sigma, hI_{|C|})} ,$$

donde

$$P(\Gamma|...) \propto \prod_{x_i \in X} \exp(\Sigma_{x_i}\Gamma)\exp\left(-\int \lambda(t)dt - \|\Gamma\|^2/2\right)$$

y Σ_{x_i} será la fila de la matriz Σ asociada a la celda que contiene a la observación x_i.

Capítulo 5

Valoración numérica de las propuestas

En este capítulo nos disponemos a aplicar los algoritmos propuestos en el capítulo anterior. Utilizaremos distintos bancos de datos simulados para estudiar el comportamiento del modelo propuesto sobre datos en los que conocemos el número de agrupaciones reales que los han generado. Las pruebas numéricas que vamos a efectuar se llevarán a cabo exclusivamente sobre el algoritmo unidimensional, ya que éste es más eficiente computacionalmente y el abanico de situaciones que comprende no es tan amplio como el del caso bidimensional. Las pruebas numéricas que vamos a realizar se basarán fundamentalmente en la comparación del algoritmo semiparamétrico que hemos propuesto en el capítulo anterior, que incluye una mixtura y un proceso log-gaussiano, frente a la propuesta mediante mixturas de Richardson y Green [80] que no incluye el término log-gaussiano.

Respecto a los algoritmos de simulación estudiados, tanto el modelo RG como nuestra propuesta han sido programados utilizando un procedimiento de simulación *fixed scan*, es decir, los distintos movimientos de actualización de las variables se suceden repetidamente con un orden sistemático, al

contrario de los algoritmos *random scan* donde los distintos movimientos se suceden según un orden aleatorio. El orden seguido en el algoritmo *RG* unidimensional es el siguiente:

- Actualización de w.
- Actualización de μ.
- Actualización de τ.
- Actualización de Z.
- Actualización de β.
- División o combinación de componentes de la mixtura.
- Nacimiento o muerte de componentes de la mixtura.

Mientras, para el modelo semiparamétrico la ordenación utilizada ha sido:

- Actualización de w.
- Actualización de μ.
- Actualización de τ.
- Actualización de Z.
- Actualización de β.
- División o combinación de componentes de la mixtura.
- Nacimiento o muerte de componentes de la mixtura.
- Actualización de ϕ.

- Actualización de ρ.
- Actualización de σ.
- Actualización de Γ.

Para el caso bidimensional el orden que hemos utilizado en el algoritmo RG ha sido el mismo que en el caso unidimensional, salvo el movimiento de *división-combinación* que tal y como se ha descrito en el capítulo anterior, no se utilizará en esta ocasión. En la versión bidimensional de nuestra propuesta tampoco se hará uso del movimiento de *división-combinación* y el resto de movimientos se reproducen también en el mismo orden. Señalamos que en la versión bidimensional de ambos algoritmos, Z deberá ser simulado también tras el movimiento de *nacimiento-muerte*, ya que estas variables no se actualizan en dicho movimiento e intervienen en otros momentos de la simulación, por tanto resulta necesario su actualización tras dicho movimiento.

Los hiperparámetros utilizados en el modelo de mixturas unidimensional han sido los siguientes. El hiperparámetro de la distribución de las precisiones toma valor $\alpha = 2$, y los hiperparámetros de β se han tomado como g=0.2, h=$10/R^2$, donde R es el rango de los datos simulados. Por último, el límite superior del número de componentes de las mixturas se ha determinado como $M = 15$. Estos parámetros coinciden con los sugeridos en el trabajo de Stephens (1999) [91]. En cuanto al caso semiparamétrico se han utilizado los valores que acabamos de señalar para el modelo de mixturas más los siguientes. Para el límite superior de la desviación típica del efecto log-gaussiano se ha determinado $b_\sigma = 2$, cuyo valor parece razonable teniendo en cuenta que corresponde a una escala logarítmica. Los límites de la distribución de la correlación para este efecto aleatorio se han establecido en los siguientes valores: $a_\rho = 3/max(d)$, $b_\rho = 3/min(d\backslash\{0\})$ donde d

es el conjunto de distancias entre las celdas del grid utilizado. La primera de estas elecciones garantiza que entre las dos celdas más distanciadas la correlación será inferior a 0.05, mientras que la segunda garantiza que la correlación entre las celdas más cercanas ha de ser superior a 0.05. La no acotación de estos valores conlleva problemas de convergencia del algoritmo propuesto tal y como se ha expuesto en el capítulo anterior.

Respecto al grid escogido para la estimación del proceso log-gaussiano, en casi todas las simulaciones que se van a efectuar en el presente capítulo se va a utilizar un grid con 100 celdas. En Benes et al. (2003) [10] se discute el efecto del tamaño del grid sobre la solución final y se concluye que un aumento indefinido del número de celdas en el grid no mejora en exceso la solución obtenida, sino que puede llegar a producir inestabilidad numérica en ésta. En todos los casos estudiados, el límite superior del grid utilizado vendrá dado por la siguiente expresión:

$$max(X) + 0{,}1 \cdot R \, ,$$

donde X es el conjunto de observaciones del proceso. Es decir, el extremo superior del grid excederá a la mayor de las observaciones en un 10 % del rango de los datos. A su vez, el extremo inferior del grid se determina de forma análoga, es decir, dicho valor será:

$$min(X) - 0{,}1 \cdot R \, .$$

Para la valoración de la convergencia de las cadenas simuladas nos hemos centrado fundamentalmente en aquellas variables que aparecen en todas las iteraciones, independientemente del número de componentes de la mixtura que tenga el proceso de simulación en cada momento. Así, no se ha valorado directamente la convergencia de las medias o las precisiones de las componentes de la mixtura ya que el número de variables de este tipo

que se disponen varía conforme avanza el proceso de simulación. Este hecho dificulta la valoración de la convergencia en estas variables aunque se han desarrollado algunos procedimientos para ello (Brooks y Giudici, 2000 [22]). Sin embargo, hasta donde nosotros conocemos no existe ningún paquete estadístico que traiga implementados dichos métodos y se ha considerado que la valoración de la convergencia realizada proporciona unas garantías razonables de que la calidad de la simulación es adecuada. En el algoritmo de mixturas RG se ha valorado la convergencia directamente sobre las siguientes cadenas: la deviance del modelo, el número de componentes de la mixtura m y el parámetro de la distribución inicial de las precisiones de las componentes β. Para la propuesta semiparamétrica se ha valorado también la convergencia de las cadenas para la variable log-gausiana f en un subconjunto de 5 celdas de la región de estudio elegidas al azar, y los parámetros que determinan el proceso log-gaussiano: σ, ϕ y ρ.

Para determinar la convergencia de los resultados en primer lugar hemos realizado una inspección visual de cada una de las variables. En el caso que dicha inspección visual resulta satisfactoria calculamos el estadístico de Brooks-Gelman-Rubin (Brooks y Gelman, 1998 [21]) para cada variable y si en todas ellas este estadístico es inferior a 1.1 se considera la simulación efectuada como adecuada. Además, hemos utilizado la autocorrelación de las cadenas generadas para cada variable como un indicador de la calidad de los resultados del proceso de simulación.

Hemos generado 2 cadenas independientes para la inferencia en cada banco de datos. Éstas constaban de 20.000 iteraciones cada una, de las que han sido descartadas las 10.000 primeras. Para el resto de simulaciones se ha guardado una de cada 20 iteraciones, por tanto el tamaño muestral en el que se basa la inferencia sobre cada variable del modelo es de 1.000 realizaciones. Se podría haber tomado un tamaño muestral superior en cada uno

de los análisis, pero dado el gran número de bancos de datos que nos disponemos a analizar y el elevado número de variables en cada uno de ellos -hay que tener en cuenta que el proceso log-gaussiano conlleva generalmente la simulación de un gran número de variables- resulta aconsejable no tomar un tamaño muestral excesivamente alto para evitar problemas computacionales de almacenamiento. Por último, en el caso que la simulación no cumpliera los criterios de convergencia que se han descrito en el párrafo anterior, la repetiríamos duplicando el número de iteraciones generadas, descartadas y el salto entre las distintas iteraciones guardadas.

En relación a la probabilidad de aceptación de los movimientos de Metropolis-Hastings y Langevin-Hastings, en Roberts y Tweedie (1996) [84] y en Roberts y Rosenthal (1998) [83] se determinan las probabilidades de aceptación óptimas de ambos algoritmos. Para el primero, la probabilidad de aceptación óptima corresponde a un 23 %, mientras que para el segundo dicho valor corresponde a un 57 %. Por tanto, sintonizaremos los parámetros del proceso de simulación de forma que se obtengan valores de aceptación similares a los señalados. Para las variables en las que se ha empleado como función de propuesta la distribución empleada en el algoritmo RG para simular mediante Gibbs Sampling no hemos realizado ningún tipo de sintonización. En esos casos, si la probabilidad de aceptación resulta superior a los valores que hemos precisado no nos debería preocupar, ya que dicha función de propuesta ofrecía resultados bastantes satisfactorios en el trabajo de Richardson y Green. Por tanto, si la probabilidad de aceptación es más alta tendremos un comportamiento parecido entre el algoritmo propuesto y el utilizado por estos autores. En ese caso esperamos que el comportamiento de nuestra propuesta sea bastante razonable ya que el del algoritmo RG así lo era.

La programación de los algoritmos propuestos en el capítulo anterior se

ha realizado en el entorno de programación estadístico R (http://cran.r-project.org/). Nos consta que este lenguaje de programación no es extremadamente eficiente para realizar procedimientos iterativos pero creemos que su "sencillez" de programación compensa esta carencia. La estrategia que se planteó originalmente fue la programación de todas las rutinas en R y posteriormente reprogramar en otro lenguaje más eficiente (se pensó en C) las partes que ralentizaran más el proceso de simulación. Sin embargo, una vez programados ambos algoritmos en R, se observó que los tiempos de computación que ofrecían eran bastante razonables por lo que no se consideró necesario exportar ninguna parte del proceso a otro lenguaje de programación.

5.1. Análisis de los datos de galaxias

En la presente sección vamos a presentar los resultados obtenidos de la aplicación del modelo de mixturas RG y de nuestra propuesta semiparamétrica al banco de datos de galaxias que aparece entre otros en Richardson y Green (1997) [80], Escobar y West (1995) [42], Stephens (2000) [93] o Cappé et al. (2003) [24]. De esta forma presentamos los resultados de nuestra propuesta sobre un conjunto de datos clásico analizado por distintos autores. Así, resultará posible comparar los resultados obtenidos en dichos trabajos con los que presentamos a continuación.

En primer lugar señalamos que los resultados obtenidos en nuestra aplicación del modelo de Richardson y Green son muy similares a los publicados por éstos en su artículo [80]. Este hecho no resulta sorprendente ya que la formulación empleada ha sido idéntica en ambas ocasiones y las únicas diferencias deberían referirse al error de Monte Carlo de la simulación. No obstante, vale la pena hacer esta apreciación ya que supone una validación

Modelo de Richardson Green						
Variable	k	β	logdeviance			
Gelman-Rubin	1.02	1.00	1.02			
Modelo semiparamétrico						
Variable	k	β	logdeviance	σ	ρ	ϕ
Gelman-Rubin	1.01	1.00	1.02	1.02	1.00	1.09
Variable	f[10]	f[24]	f[64]	f[78]	f[85]	
Gelman-Rubin	1.03	1.03	1.00	1.01	1.00	

Cuadro 5.1: Valores del estadístico de Brooks-Gelman-Rubin para distintas variables del modelo de Richardson-Green y el modelo semiparamétrico.

de las rutinas que hemos programado.

Respecto a las cadenas simuladas, hemos procedido a su observación visual, en las que no se ha observado ningún comportamiento que indicara una convergencia deficiente. En el cuadro 5.1 se describen los valores obtenidos del estadístico de Brooks-Gelman-Rubin para varias variables en el algoritmo *RG* y el semiparamétrico. Recordamos que valores del estadístico próximos a 1 sugieren que se ha alcanzado la convergencia y habíamos definido como criterio de validez de la simulación el que todos los estadísticos fueran inferiores a 1.10. Según se puede comprobar en dicho cuadro todos los estadísticos son inferiores a dicha cantidad, por lo que la simulación efectuada (10.000 iteraciones de calentamiento más 10.000 iteraciones para la inferencia) cumple los criterios que habíamos fijado. Notar que las variables comunes a ambos modelos tienen un comportamiento muy similar en relación al estadístico de Brooks-Gelman-Rubin. La única variable en la que encontramos algún indicio de convergencia ligeramente peor que en el resto es ϕ, sin embargo incluso en ésta se cumplen los criterios establecidos.

Modelo de Richardson Green						
Variable	k	β	logdeviance			
Autocorrelación	0.79	0.33	0.35			
Modelo semiparamétrico						
Variable	k	β	logdeviance	σ	ρ	ϕ
Autocorrelación	0.60	0.38	0.36	0.64	0.29	0.53
Variable	f[10]	f[24]	f[64]	f[78]	f[85]	
Autocorrelación	0.48	0.52	0.47	0.42	0.47	

Cuadro 5.2: Autocorrelación de orden 1 de las cadenas para distintas variables del modelo de Richardson-Green y el modelo semiparamétrico.

Respecto a la correlación de las cadenas de las variables anteriores, en el cuadro 5.2 se puede observar la autocorrelación de orden 1 de cada una de ellas. Las autocorrelaciones anteriores son todas inferiores a 0.8 y la autocorrelación de orden 5 de todas las variables (no se muestran) son todas casi nulas. Por tanto la calidad de las variables simuladas parece bastante aceptable, si bien algunas variables tienen peor comportamiento que otras, por ejemplo k, el número de componentes de la mixtura.

En la figura 5.1 se muestra la distribución posterior del número de componentes para cada modelo. En ella se puede observar que aunque el rango de valores probables según ambos modelos es similar, existen diferencias importantes entre ambas distribuciones. Según el modelo de mixturas la moda de dicha distribución se sitúa en 6, siendo [3,9] el intervalo del número de componentes en el que la probabilidad de la distribución posterior es superior a 0.05 para todos sus valores. En nuestra propuesta semiparamétrica la moda se sitúa en 3 y el intervalo anterior resulta [3,7] en esta ocasión. El número medio de componentes ajustadas por el modelo de mixturas es de 6.47, mientras que mediante nuestra propuesta dicho valor disminuye hasta

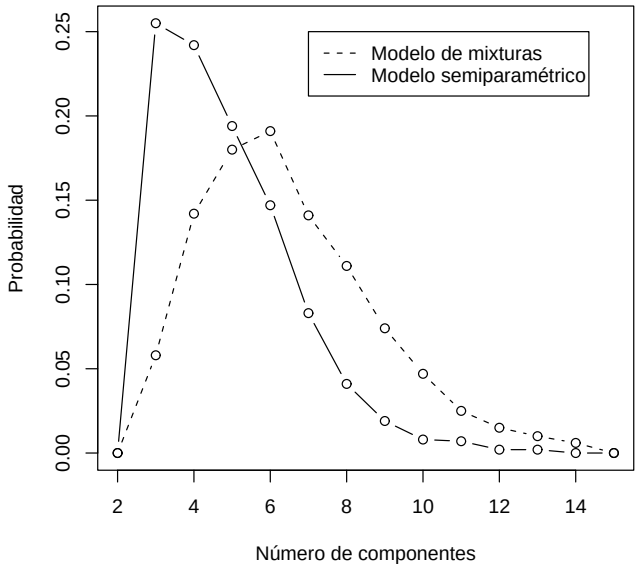

Figura 5.1: Distribución final del número de componentes según ambos modelos.

4.87. En el modelo semiparamétrico la moda de la distribución, 3, se sitúa contigua a un número de componentes con probabilidad 0. Por tanto parece que la introducción de la componente log-gausiana produce un desplazamiento de la masa de la distribución posterior hacia la izquierda hasta el número de componentes que el modelo ya no considera probable.

En la figura 5.2 se muestra la distribución de β resultante tanto del modelo de mixturas como de nuestra propuesta. Se puede observar que la distribución estimada de β según el modelo semiparamétrico sugiere valores superiores a los obtenidos a partir del modelo de mixturas. De hecho

la media estimada mediante nuestra propuesta, 3.27, resulta superior a la estimada por el modelo de mixturas 1.63. Teniendo en cuenta que la media de las precisiones de las componentes en ambos modelos es $\alpha/\beta = 2/\beta$, resulta que las precisiones del modelo semiparamétrico son en general inferiores que las del modelo de mixturas. Así pues, en este último modelo, en el que se emplean en general más componentes y de mayor precisión, el ajuste será menos parsimonioso y cualquier pequeña agregación en los datos se ajustará mediante la creación de una nueva componente. Sin embargo, el modelo semiparamétrico permite ajustar esas variaciones, quizás producto de heterogeneidad medioambiental o error aleatorio, con la componente log-gausiana. En consecuencia, este último modelo ante cualquier falta de ajuste no tiene que recurrir a emplear una nueva componente de la mixtura para compensarla, sino que dichas componentes se emplearan únicamente cuando se considere necesario para mejorar el ajuste y no ante cualquier carencia de éste.

Respecto a los parámetros del modelo log-gaussiano, la figura 5.3 muestra las distribuciones finales obtenidas para ϕ, σ, ρ de nuestra propuesta semiparamétrica. Se observa que el rango de valores obtenido para ϕ varía entre 3.15 y 5.09, concretamente la media de su distribución posterior se sitúa en 4.35. Dicha media conlleva que el numero medio de observaciones esperadas por el proceso es de alrededor de exp(4,35) = 77,5, muy próximo a los 82 observados. Por otro lado, la desviación típica del proceso toma valores entre 0 y 1.43, aunque su moda parece situarse alrededor de 0.4, magnitud considerable teniendo en cuenta que hablamos de una escala logarítmica. Por último, la distribución de la correlación del proceso es bastante uniforme sobre el rango de valores posible, apuntando ligeramente hacía los valores más bajos de ρ. Dichos valores sugieren la existencia de autocorrelación espacial en el proceso log-gaussiano, es más, dicha autocorrelación estaría presente hasta distancias relativamente grandes dentro del

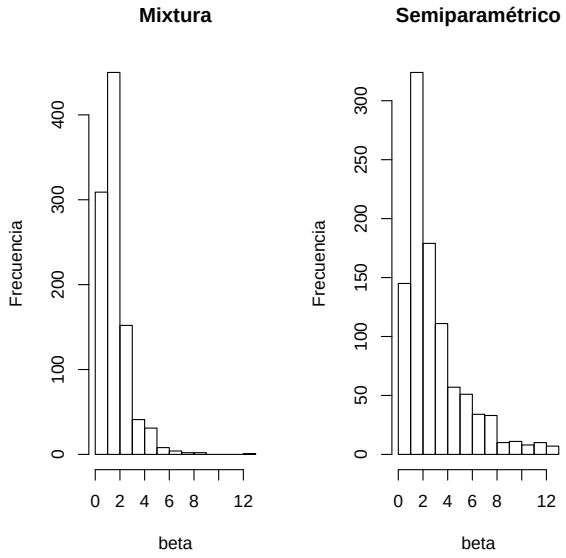

Figura 5.2: Distribución final de β según ambos modelos.

grid, o dicho de otra forma el proceso log-gaussiano se comportaría de forma bastante suave espacialmente. La forma tan llana de esta distribución no debería resultar extraña ya que, tal y como se señala en diversos trabajos (Möller et al., 1998 [69]; Möller y Waagepetersen, 2004 [70]) resulta difícil aprender de dicho parámetro a partir de los datos, por tanto resulta habitual obtener distribuciones finales bastante planas para esta variable.

En cuanto a la forma estimada del proceso log-gaussiano, en la figura 5.4 se ha representado éste junto a la mixtura ajustada por nuestra propuesta. Observamos que el proceso log-gaussiano varía ostensiblemente en aquellos lugares donde la mixtura toma valores más grandes. Por tanto, el proceso log-gaussiano matiza la forma ajustada por el modelo de mixturas allá donde la matización tiene sentido, donde hay presencia de observaciones. Dicha matización puede hacer que una localización llegue a tener doble probabilidad que otra, así vemos que alrededor de 20 hay localizaciones donde el proceso log-gaussiano toma valores próximos a 120 y otros inferiores a 60. La utilización del proceso log-gaussiano para ajustar variaciones en la función de intensidad evita abusar del número de componentes de la mixtura para que ésta se adecúe a los datos como parece suceder en el modelo de Richardson y Green.

En la figura 5.5 se ha representado en verde la función de densidad ajustada por el modelo de mixturas, mientras que la función de densidad del modelo semiparamétrico se ha representado en color negro. Se puede observar que ambos modelos proporcionan funciones de densidad bastante similares, siendo la diferencia más apreciable entre ambas el hecho de que nuestra propuesta parece ajustarse más a los datos al tener una estructura más flexible. Los datos observados también se han representado en la misma figura en la parte inferior. Por otra parte, en la figura 5.5 se ha representado en color rojo la distribución del modelo de mixturas del proceso

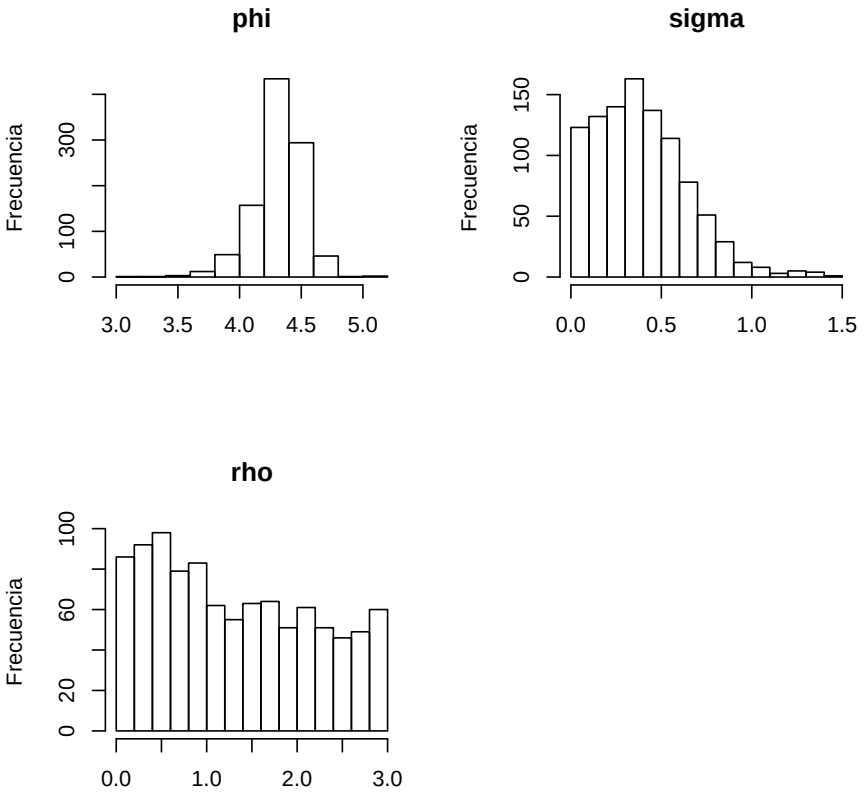

Figura 5.3: Distribución final de los parámetros del proceso log-gaussiano.

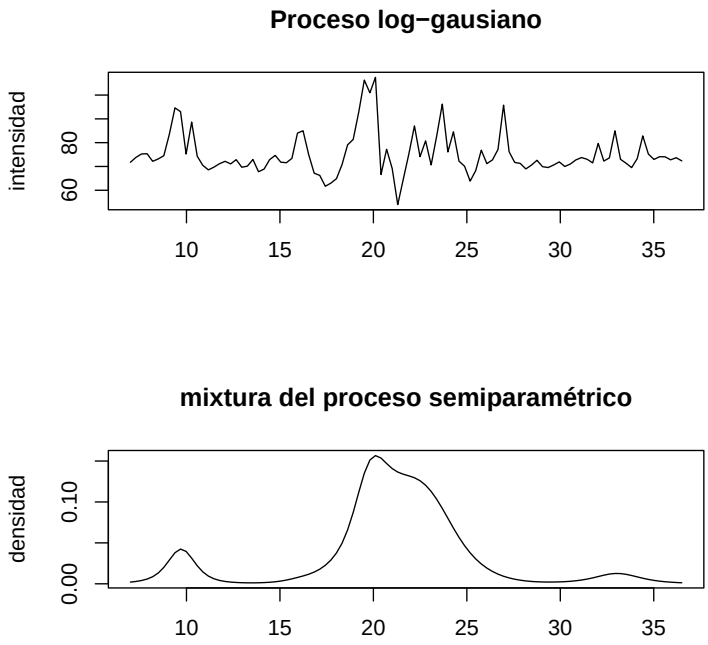

Figura 5.4: Componentes del modelo semiparamétrico. Proceso log-gaussiano y mixtura.

semiparamétrico, excluyendo el término log-gaussiano. Entre esta distribución y la resultante del modelo de mixturas de Richardson y Green sí que se pueden observar diferencias de mayor magnitud. Se aprecia que la mixtura de nuestra propuesta tiene un comportamiento bastante más parsimonioso, corroborando los comentarios de los párrafos anteriores. El ajuste de ambos modelos de mixturas presenta diferencias considerables y en principio no se tiene ningún criterio para decidir cual de los dos modelos proporciona mejor ajuste. Por ello, en la siguiente sección se va a realizar un estudio computacional sobre datos simulados donde se valorará el comportamiento de ambos modelos. En base a dichas pruebas se podrá evaluar cuales son los puntos fuertes y las carencias de cada propuesta y cual de ellas proporciona mejores resultados.

Por último, en cuanto a los tiempos de computación de ambas propuestas, el algoritmo de Richardson y Green sobre los datos de Galaxias ha tardado 523.1 segundos en realizar las 20.000 iteraciones programadas, mientras que el algoritmo semiparamétrico ha invertido 1390.1 segundos en realizar el mismo número de iteraciones. Por tanto el algoritmo de nuestra propuesta, aunque conlleva una gran complejidad computacional e incorpora un gran número de variables, no resulta inviable en términos de tiempo de computación. Ni siquiera habiendo programado las rutinas en R, que tal y como hemos comentado no es la opción computacionalmente más eficiente.

5.2. Valoración sobre datos simulados

Vamos a dividir la valoración de los dos algoritmos propuestos en dos partes. En la primera generaremos los datos directamente a partir de un modelo de mixturas y se realizará la inferencia según ambas propuestas, el modelo de mixturas y el semiparamétrico. En la segunda se generarán los

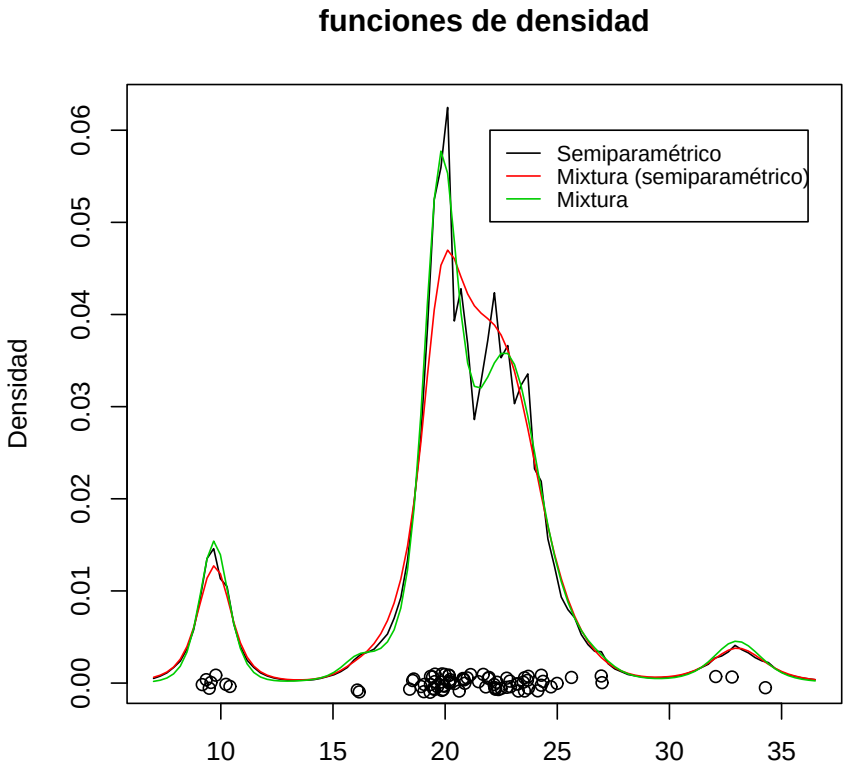

Figura 5.5: Función de intensidad ajustada por el proceso semiparamétrico (negro), mixtura del modelo semiparamétrico (rojo) y modelo de mixtura sin componente log-gausiana (verde).

datos a partir de distintos modelos de mixturas con ruido log-gaussiano. De esta forma se podrá valorar las mejoras que introduce nuestra propuesta en presencia de un proceso con los dos tipos de agregación, general e individual. A continuación describimos los resultados de ambas pruebas.

5.2.1. Valoración sobre datos de mixturas

Para la valoración del comportamiento del modelo de mixturas y nuestra propuesta semiparamétrica sobre datos generados a partir de una mixtura, hemos considerado 10 situaciones específicas y para cada una de ellas se han replicado 10 bancos de datos. Es decir, vamos a evaluar ambos modelos sobre un total de 100 *bancos de datos* que provienen de 10 modelos de mixturas distintos o *grupos*.

Las mixturas generadas son o bien de distribuciones normales o de distribuciones t, de esta forma se podrá valorar la influencia que puede tener la desviación respecto la hipótesis de normalidad sobre ambos modelos. En el cuadro 5.3 se describen los 10 grupos contemplados en función de los parámetros que los distinguen. Así pues, los grupos pueden variar según el número de componentes de la mixtura o el número de observaciones, las desviaciones típicas de las componentes, los pesos de la mixtura, el número de celdas del proceso log-gaussiano, el dominio en el que se ha definido o los grados de libertad en el caso que se trate de una mixtura de distribuciones t. En todos estos parámetros hay ciertas especificaciones que vamos a considerar por defecto, y en cada grupo se variará alguna de ellos para valorar su efecto. Las especificaciones por defecto se corresponden con las siguientes características: 50 observaciones generadas a partir de una mixtura de 3 distribuciones normales, todas ellas con el mismo peso y con desviación típica 0.25, las medias de las componentes de la mixtura se distribuyen de

manera uniforme entre 0 y 5. Por último, para la inferencia del proceso log-gaussiano consideraremos por defecto un grid de 100 celdas.

El primer grupo de datos generado seguirá todas las especificaciones que hemos establecido por defecto, así nos valdrá como grupo de referencia con el que podremos comparar el resto. El segundo grupo contendrá 5 componentes, conservando el resto de especificaciones. El tercero contendrá 100 observaciones, mientras que el cuarto contendrá una componente con desviación típica 0.5. El quinto grupo, además de considerar 5 componentes, cambia el rango en el que éstas son generadas, concretamente lo amplía hasta el intervalo $[0\ ,\ 5{\cdot}(5/3)]$. De esta forma se mantiene la relación de distancia entre componentes respecto al grupo de referencia, es decir una componente cada 5/3 unidades de distancia. El sexto grupo considera 7 componentes y también amplía el rango de la misma forma que se hacía con el quinto caso. El séptimo grupo contiene una componente con el doble de peso que el resto, mientras que el octavo contemplará 200 celdas para el proceso log-gaussiano. Por último, el noveno y décimo grupo serán mixturas de distribuciones t en lugar de distribuciones normales, en concreto el noveno grupo utilizará distribuciones t con 5 grados de libertad y el décimo con 10.

El tamaño de los bancos de datos simulados puede parecer pequeño y podría plantearse la elección de un tamaño más elevado para valorar con más detalle el comportamiento de las dos propuestas que pretendemos evaluar. Sin embargo, se ha de tener en cuenta el tipo de aplicación a la que se pretende dedicar los modelos examinados, la evaluación de brotes o datos epidémicos y que, por suerte, en esas ocasiones el número de observaciones no suele ser demasiado elevado. Por tanto, tenemos particular interés por valorar su comportamiento en problemas con pocas observaciones y es por ello por lo que se van a efectuar las pruebas numéricas con bancos de datos

Grupo	#Com.	#Obs.	Desviaciones	Pesos	#Cel.	L.Sup	GL
1	3	50	-	-	100	5	-
2	5	50	-	-	100	5	-
3	3	100	-	-	100	5	-
4	3	50	(0.25,0.5,0.25)	-	100	5	-
5	5	50	-	-	100	5*5/3	-
6	7	50	-	-	100	5*7/3	-
7	3	50	-	(1,2,1)	100	5	-
8	3	50	-	-	200	5	-
9	3	50	-	-	100	5	5
10	3	50	-	-	100	5	10

Cuadro 5.3: Descripción de los grupos de datos generados. Leyenda: *#Com.*, número de componentes de la mixtura. *#Obs.*, número de observaciones generadas. *Desviaciones*, desviaciones típicas de cada componente (por defecto todas son 0.25). *Pesos*, pesos de las componentes de las mixturas sin estandarizar (por defecto todos iguales). *#Cel.*, número de celdas del proceso log-gaussiano. *L. Sup*, límite superior del intervalo en el que se han generado los datos. *GL*, grados de libertad en el caso de que se haya empleado una distribución t para cada componente.

de tamaño muestral relativamente pequeño.

En la figura 5.6 se ha representado, a modo de ejemplo, el primer banco de datos generado para los 9 primeros grupos. En la parte inferior de cada gráfico se pueden observar los datos generados como círculos negros y en rojo la localización de las medias de cada componente. Además, la línea roja corresponde a la distribución de la mixtura que se ha generado, mientras que la línea roja corresponde a la estimación kernel de dicha distribución generada por *R*. De esta forma se ilustra la dificultad de estimación del número de modas que encuentran los métodos de estimación que no contemplan específicamente una forma de mixtura.

Tal y como se puede apreciar en la figura 5.6, en muchas ocasiones el número de modas de las mixturas es inferior al número de componentes de éstas. Este hecho resulta una dificultad añadida a la hora de la inferencia ya que en estos casos resultará todavía más difícil determinar el número de componentes.

En todos los bancos de datos analizados hemos realizado las validaciones de la convergencia detalladas al principio de la sección, es decir inspección visual de las cadenas, test de Brooks-Gelman-Rubin para la verificación del periodo de calentamiento y cálculo de la autocorrelación de las cadenas simuladas. Tras la valoración de la convergencia, 3 bancos de datos han tenido que ser vueltos a simular con 20.000 iteraciones de calentamiento y otras 20.000 simulaciones más para el algoritmo de Green y Richardson. Mientras que para el modelo semiparamétrico hemos tenido que volver a simular 6 bancos de datos. Teniendo en cuenta que las pruebas en el caso del modelo de mixturas se han realizado sobre 3 variables, mientras que en nuestra propuesta se han realizado para 11, parece que este último tiene un comportamiento similar o incluso mejor que el algoritmo de mixturas en

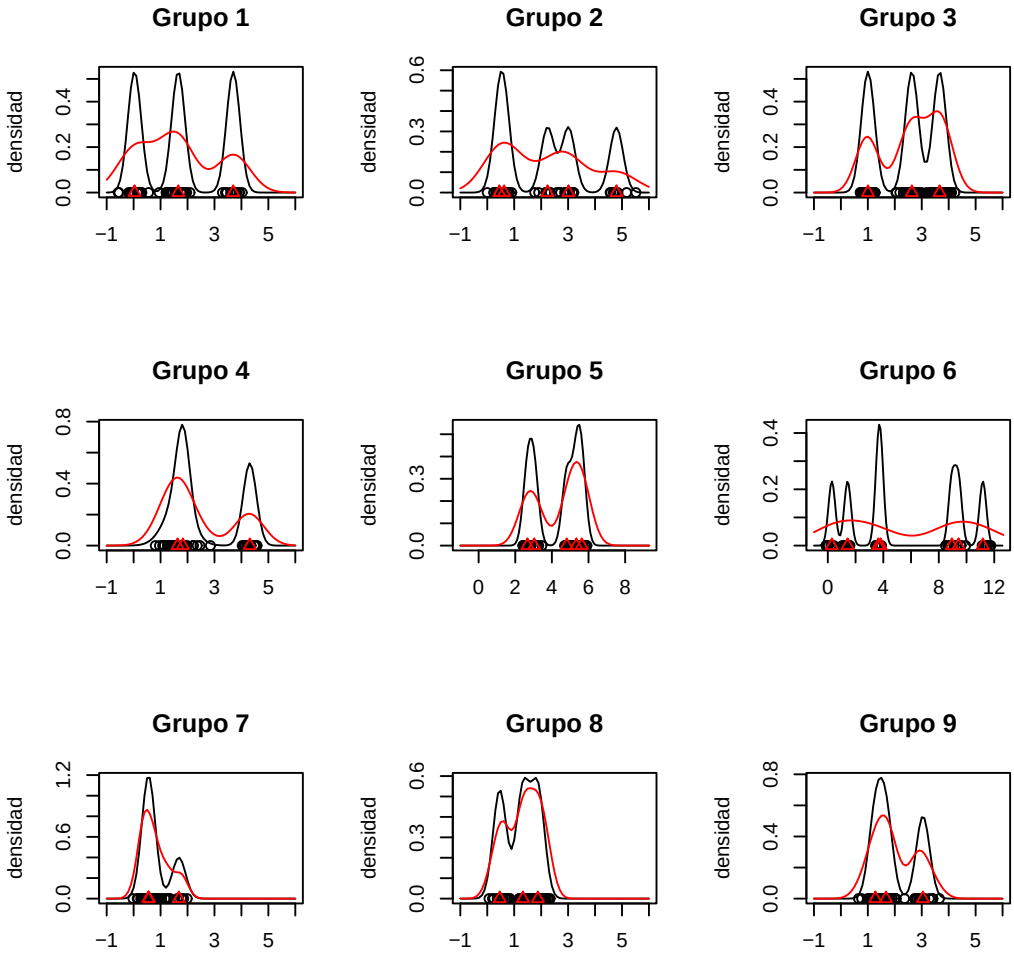

Figura 5.6: Primer banco de datos para los 9 primeros grupos de mixturas.

cuanto a convergencia.

Respecto a la autocorrelación de las cadenas monitorizadas no se observan diferencias entre los distintos grupos simulados. El parámetro con peores valores de autocorrelación de orden 1 es σ, en el que dicho dicho parámetro toma una valor promedio en todas las simulaciones de 0.47, en cualquier caso dicho valor no parece preocupante.

Respecto al tiempo de computación de cada uno de los grupos, en el cuadro 5.4 se presentan los tiempos medios (en minutos) empleados por cada modelo en cada conjunto de datos. Observamos que el algoritmo de mixturas ofrece tiempos de computación más altos únicamente cuando se aumenta el número de observaciones en los datos, es decir en el grupo número 3. Sin embargo los tiempos del proceso semiparamétrico se ven afectados también por el número de componentes de las mixturas (grupos 2, 5 y 6) además del número de observaciones de las que constan los bancos de datos. Volvemos a constatar, al igual que con los datos de galaxias, que nuestra propuesta no aumenta en exceso el coste computacional respecto al algoritmo de mixturas. Sobre todo si se tiene en cuenta la complejidad computacional que conlleva el proceso log-gaussiano al introducir tantas variables en el modelo. Además, sorprendentemente el aumento del número de celdas, y por tanto de variables, en el proceso log-gaussiano apenas incrementa los tiempos de cálculo del algoritmo (grupo 8). No obstante, estos resultados dependen de la programación efectuada y cualquier alteración en las rutinas o cambio en la programación alterarán los resultados que hemos obtenido.

En cuanto al número de componentes ajustadas por cada modelo, en el cuadro 5.5 se muestran distintos indicadores relacionados con la inferencia sobre este valor. Para cada grupo de datos se muestra la diferencia entre el número medio de componentes ajustadas y el número verdadero que se ha

grupo	1	2	3	4	5	6	7	8	9	10
mixtura	7.2	7.1	9.4	7.1	7.1	7.1	7.1	7.2	7.2	7.2
semipar.	19.5	22	22.2	19.5	22	22.8	19.7	19.4	18.9	19

Cuadro 5.4: Tiempo de computación medio (en minutos) de los modelos para los bancos de datos de cada grupo.

utilizado para cada grupo k^* y la varianza del número de componentes ajustado suponiendo conocido el número de componentes de cada grupo. Este último valor cuantificará la dispersión de las medias obtenidas alrededor del verdadero valor del número de componentes. Finalmente, en la última columna del cuadro se presenta el cociente de la probabilidad posterior del verdadero valor del número de componentes, entre el modelo de mixturas y el modelo semiparamétrico. Dicha probabilidad se ha calculado suponiendo una aproximación normal para la distribución final de este valor con media y varianza dada por los resultados de cada banco de datos. La última fila del cuadro corresponde al valor promedio de las 10 filas anteriores, de esta forma se resume el comportamiento general de cada modelo sobre el total de datos analizados.

En el cuadro 5.5 se puede observar que en general ambos algoritmos sobreestiman el número de componentes de las mixturas, sin embargo dicha sobreestimación es inferior en el modelo semiparamétrico. En los únicos casos en los que se infraestiman el número de componentes de las mixturas corresponde a los grupos 2 y 6. Esto se debe a que en dichos grupos hay un considerable número de medias que se sitúan muy cerca entre sí, por tanto en esas ocasiones la forma de la función de densidad parece apuntar a un número de componentes inferior a las que realmente existen. Este hecho se puede apreciar en el cuadro 5.6 donde para cada grupo se muestra el cociente entre el número total de modas y componentes de la mixtura. Resulta

Grupo	$\overline{k-k^*}$ *Green*	$\overline{k-k^*}$ *Semipar.*	$\overline{(k-k^*)^2}$ *Green*	$\overline{(k-k^*)^2}$ *Semipar.*	*Cociente D.final*
1	0.7262	0.4364	4.7752	3.3412	0.794
2	-0.1705	-0.6364	5.0489	4.4882	0.868
3	0.6536	0.4523	3.404	2.8489	0.851
4	0.9803	0.5235	5.2879	3.2811	0.785
5	0.4426	0.038	5.8634	4.8034	0.874
6	-1.2887	-1.7082	6.9207	7.184	1.145
7	0.8854	0.6647	4.309	3.2993	0.882
8	0.9354	0.5632	7.1242	4.6238	0.819
9	0.5299	0.1954	4.4247	3.1432	0.901
10	0.5817	0.3425	4.5901	3.5455	0.829
total	0.4276	0.0871	5.1748	4.055	0.875

Cuadro 5.5: Estadísticos del número de componentes para cada grupo.

Grupo	1	2	3	4	5	6	7	8	9	10
Cociente	0.76	0.72	0.83	0.80	0.72	0.67	0.90	0.83	0.76	0.73

Cuadro 5.6: Cociente entre el número de modas y componentes en cada grupo de datos

interesante apreciar como al aumentar el rango en el que se distribuyen las medias (grupo 5), es decir al espaciar las componentes de las mixturas, nuevamente ambos algoritmos sobreestiman el número de componentes. Por tanto a no ser que las componentes de la mixtura se solapen demasiado entre sí, ambos modelos sobreestimarán el número de componentes, aunque la sobreestimación será menor en nuestra propuesta.

Respecto a la variabilidad en la estimación del número de componentes, nuestra propuesta produce estimaciones en general con menor variabilidad en torno al verdadero valor. Dicha variabilidad sólo resulta mayor en el modelo semiparamétrico en el caso del grupo 7 que es el grupo con mayor sesgo en ambos algoritmos, por tanto nuestra propuesta corrige dicho sesgo aumentando la variabilidad en su predicción. Así el comportamiento de este algoritmo es bastante satisfactorio. Los dos modelos producen estimaciones más variables en los casos con mayor número de componentes, lo que también resulta razonable ya que cuanto mayor es dicho valor más difícil resultará determinar el número de componentes.

En cuanto a la probabilidad final del verdadero número de componentes en ambos modelos, observamos que en general el modelo semiparamétrico otorga mayor masa de probabilidad a este valor que el modelo de Green ya que el cociente de ambas es, en general, inferior a 1. El único caso en el que el modelo semiparamétrico ofrece peores resultados es el grupo 7, aunque ya hemos comentado que en dicha ocasión resulta difícil discriminar el número

Grupo	$P(k = k^*)$ *Green*	$P(k = k^*)$ *Semipar.*	$P(k = m^*)$ *Green*	$P(k = m^*)$ *Semipar.*
1	0.262	0.272	0.369	0.400
2	0.162	0.152	0.249	0.261
3	0.292	0.287	0.331	0.352
4	0.251	0.271	0.282	0.339
5	0.181	0.186	0.224	0.247
6	0.104	0.093	0.212	0.228
7	0.302	0.306	0.357	0.371
8	0.210	0.230	0.331	0.347
9	0.225	0.239	0.248	0.291
10	0.240	0.267	0.361	0.394
total	0.223	0.230	0.296	0.323

Cuadro 5.7: Probabilidades posteriores de que el número de componentes sea igual al número de componentes verdaderas (k^*) y al número de modas de cada conjunto de datos (m^*).

de componentes ya que en ese caso las componentes se solapan mucho entre sí. Por tanto nuestra propuesta parece tener un comportamiento predictivo mejor que el del modelo de Green en relación al número de componentes de cada banco de datos.

En el caso que no asumamos una aproximación normal para valorar la verosimilitud de k^*, podemos evaluar directamente la proporción de iteraciones que cada cadena ha visitado dicho valor. En el cuadro 5.7 se muestra dicha frecuencia para ambos modelos, así como la frecuencia de veces que el número de componentes ha coincidido con el número de modas de cada banco de datos m^*. Se observa que la probabilidad de que el número de componentes coincida con el valor verdadero es muy similar para los

dos modelos, si bien suele ser ligeramente superior para el modelo semiparamétrico. Sin embargo, dicho valor es superior en el modelo de Green para los grupos 2, 3 y 6, es decir, los modelos donde el número de componentes puede parecer menos claro debido al solapamiento de las componentes (grupos 2 y 6) y el grupo con mayor número de observaciones. Por otra parte, la probabilidad posterior de que el número de componentes coincida con el número de modas de cada banco de datos es superior, en todos los grupos, en nuestra propuesta. Además, las diferencias encontradas entre ambos modelos para esta probabilidad son en general superiores que para las probabilidades de que el número de componentes coincida con k^*. Por tanto, parece que el modelo semiparamétrico discrimina mejor que el modelo de Green el número de modas de cada banco de datos mientras que el número de componentes se predice de manera similar por ambas propuestas.

En el cuadro 5.8 se muestra para cada grupo el número de bancos de datos en que la moda de la distribución posterior del número de componentes coincide con el verdadero número de componentes o con su número de modas. Observamos que el número medio de veces que coincide el número de modas a posteriori con el número de componentes de la mixtura coincide para ambos modelos, en los dos casos éste asciende a un 30 %. No observamos ninguna característica especial en los grupos donde este valor difiere en ambos modelos. Sin embargo sí que se aprecian diferencias entre los modelos en relación al número de veces que la moda de la distribución posterior coincide con la verdadera moda para cada conjunto de datos. En general en el modelo semiparamétrico dichos valores coinciden con más frecuencia, un 68 % frente un 62 %. En el único caso en el que nuestra propuesta presenta peores valores que el modelo de mixturas es en el grupo 6. A tenor de estos resultados se puede observar nuevamente que ambos modelos predicen bastante mejor el número de modas en cada banco de datos que el número de componentes.

Grupo	$\#(\hat{k} = k^*)$ *Green*	$\#(\hat{k} = k^*)$ *Semipar.*	$\#(\hat{k} = m^*)$ *Green*	$\#(\hat{k} = m^*)$ *Semipar.*
1	4	3	7	8
2	3	2	5	6
3	6	6	7	8
4	2	3	6	7
5	2	2	4	4
6	0	0	5	4
7	5	5	8	8
8	4	4	7	7
9	2	2	4	6
10	3	4	9	10
Total	3.0	3.0	6.2	6.8

Cuadro 5.8: Número de ocasiones en los que la moda de la distribución posterior $\hat{k}$ coincide con el número de componentes verdaderas (k^*) o el número de modas de cada conjunto de datos (m^*) para cada modelo.

Grupo	1	2	3	4	5	6	7	8	9	10
Diverg.	0.11	-0.27	0.03	-0.05	0.02	0.02	-0.06	0.06	-0.09	0.04

Cuadro 5.9: Diferencia media entre las divergencias del modelo de mixturas y el modelo semiparamétrico respecto a las distribuciones que han generado los datos.

Por último, para cada conjunto de datos hemos calculado la divergencia entre la función de densidad original y la media de la función de densidad posterior sobre un grid de 100 puntos, según ambos modelos. De esta forma se pretende evaluar qué modelo proporciona mejores estimaciones de la función de densidad original. En el cuadro 5.9 se presenta la diferencia entre las funciones de divergencia para el modelo de mixturas y el modelo semiparamétrico para cada grupo de datos. Por tanto los valores positivos de dicho cuadro indican una discrepancia menor entre la distribución teórica y el resultado de nuestra propuesta que la discrepancia frente a la distribución resultante del modelo de mixturas. En sólo el 60 % de los grupos la divergencia ha sido positiva mientras que las divergencias negativas aparecen en una serie de grupos sin relación aparente. Por tanto parece que no se puede deducir que ninguno de los dos modelos presente mejor comportamiento que el otro en términos de divergencia respecto de la función original, en consecuencia ninguno de los dos modelos tiene un comportamiento destacado en cuanto al ajuste de la función de distribución.

A modo de resumen, a partir de los resultados obtenidos en esta sección podemos destacar los siguientes hechos:

- De los casos contemplados, los únicos que parecen tener un comportamiento particular son aquellos en los que el número de componentes superpuestas es más elevado. En estos casos el ajuste del número de componentes de ambos modelos resulta algo más pobre.

- En cuanto al ajuste de las componentes de la mixtura el modelo semiparamétrico ofrece en general mejores resultados. Dicha mejora se ha podido constatar en 5 aspectos: menor sesgo de la media de la distribución final, menor variabilidad de la predicción respecto al verdadero valor, mayor probabilidad a posteriori del número de componentes

original y el número de modas original. Por último, también se ha obtenido un porcentaje de coincidencia mayor entre la moda de la distribución posterior y el número de modas de la distribución original. Por el contrario, no se ha evidenciado ninguna mejora en cuanto a la coincidencia de la moda de la distribución posterior y el número de componentes originales.

- Se ha determinado también que en todos los casos estudiados el número de componentes ajustadas por el modelo semiparamétrico es menor que para el modelo de mixturas. Este hecho corrobora la hipótesis que establecíamos con los datos de galaxias en relación al ajuste más parsimonioso de nuestra propuesta.

- No se ha podido determinar un ajuste mejor por parte de ninguno de los modelos en términos de la divergencia respecto a la función de distribución original.

5.3. Valoración sobre datos de mixturas con ruido log-gaussiano

Una vez hemos comparado el comportamiento de ambos modelos sobre datos generados a partir de mixturas, vamos a repetir los análisis anteriores sobre datos generados a partir de un modelo de mixturas al que se le ha añadido ruido con estructura log-gausiana. Es decir, las funciones de intensidad a partir de la que se generarán los datos en esta sección tienen la siguiente forma:

$$\lambda(x) = \left(\sum_{j=1}^{k} w_j \mathcal{N}(x|\mu_j, \tau_j^{-1}) \right) \cdot \exp(f(x|\sigma, \rho)) \ ,$$

Grupo	σ	ρ
11	1	∞
12	0.5	∞
13	0.2	∞
14	0.5	8.57
15	0.5	4.28
16	0.5	2.14

Cuadro 5.10: Parámetros del proceso log-gaussiano para los grupos generados.

donde f sigue una distribución normal multivariante sobre un conjunto de celdas que cubren la región de estudio.

Para los análisis de esta sección se han generado 6 grupos de datos que se han numerado del 11 al 16 para no confundirlos con los de la sección anterior. Para cada uno de estos grupos nuevamente se han generado 10 bancos de datos distintos, por tanto para los análisis de esta sección hemos hecho uso de 60 bancos de datos. En el cuadro 5.10 se describen las características de estos grupos en función de los parámetros del proceso log-gaussiano, σ y ρ. Los valores finitos de ρ responden a aquellos para los que la correlación entre puntos separados entre sí 5, 10 y 20 celdas de distancia, respectivamente, vale 0.05. Por tanto, en esos casos podemos entender que las celdas que disten más de esos valores se distribuirán de forma "independiente" en el proceso log-gaussiano. En los casos en que ρ sea infinito todas las celdas del proceso log-gaussiano serán independientes entre sí o, dicho de otra forma, el error log-gaussiano no tiene estructura espacial. Respecto a la componente con forma de mixtura de dichos datos se han empleado las especificaciones del grupo 1 de la sección anterior, así dicho grupo nos valdrá como referencia también para comparar los resultados de la presente sección.

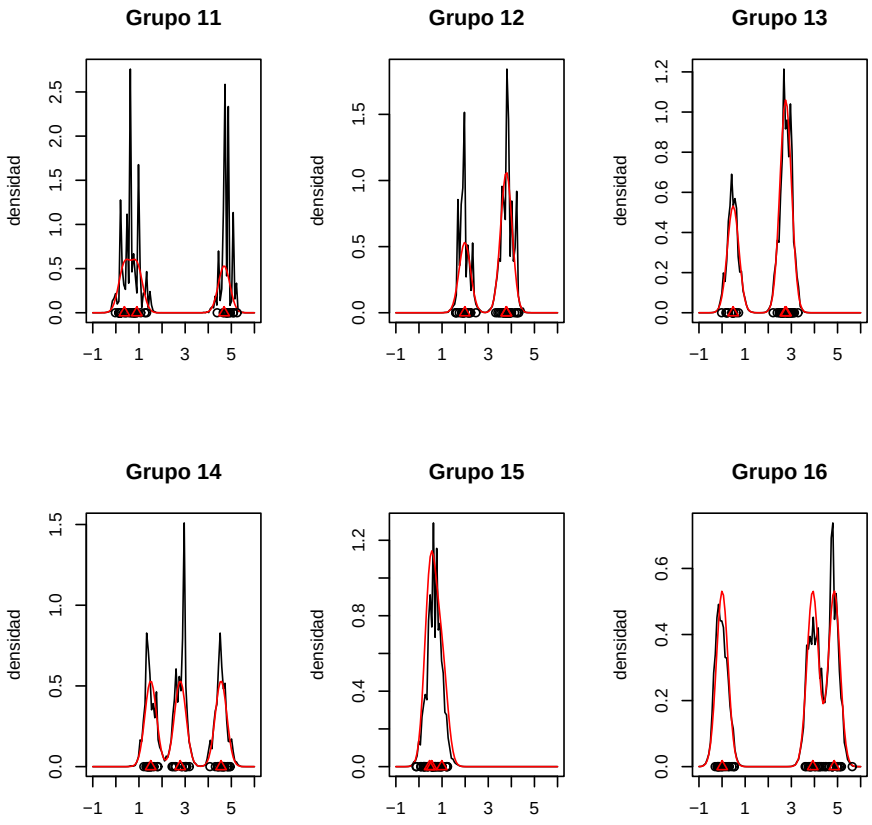

Figura 5.7: Primer banco de datos de cada grupo generado según un modelo de mixturas con ruido log-gaussiano.

En la figura 5.7 hemos representado también a modo de ejemplo el primer banco de datos de los grupos 11 a 16. Al pie de cada figura aparecen los datos generados (color negro) junto a la localización de la media de cada componente (color rojo). Además, se ha representado mediante una línea de color negro la función de densidad que ha sido utilizada para generar cada banco de datos. Mientras que la mixtura asociada a cada conjunto de datos se ha representado mediante una línea de color rojo. Se puede observar como la magnitud del ruido log-gaussiano disminuye de las figuras 11 a la 13 ya que la desviación típica del proceso log-gaussiano disminuye para cada uno de estos grupos. También se aprecia que el "ruido" tiene un comportamiento menos caótico en el grupo 16 que en el 15 y éste a su vez también menor que en el 14. Esto es consecuencia de que el ruido log-gaussiano en los últimos grupos guarda correlación espacial con un número mayor de celdas, por tanto en esos casos el ruido tiene un comportamiento más suavizado que en los grupos anteriores.

En cuanto a la validación de la convergencia y los tiempos de computación empleados, los valores obtenidos son muy similares a los de la sección anterior por tanto no nos vamos a extender más en este aspecto.

Respecto a la inferencia sobre los parámetros del proceso log-gaussiano, en la figura 5.8 se muestran los histogramas de la estimación de la desviación típica para el primer banco de datos de los grupos 1, 3, 11, 12, 13, 14, 15 y 16 junto con la media para cada uno de ellos. El grupo 1 se ha incluido como referencia y el grupo 3 para valorar el efecto del aumento de tamaño de los bancos de datos. Se aprecia que, en aquellos grupos en los que la desviación típica era mayor, las medias estimadas también toman valores mayores, sin embargo la estimación proporcionada no es demasiado precisa. Este hecho se debe a que el aprendizaje sobre este parámetro requiere un número de observaciones considerablemente elevado ya que los datos aportan poca

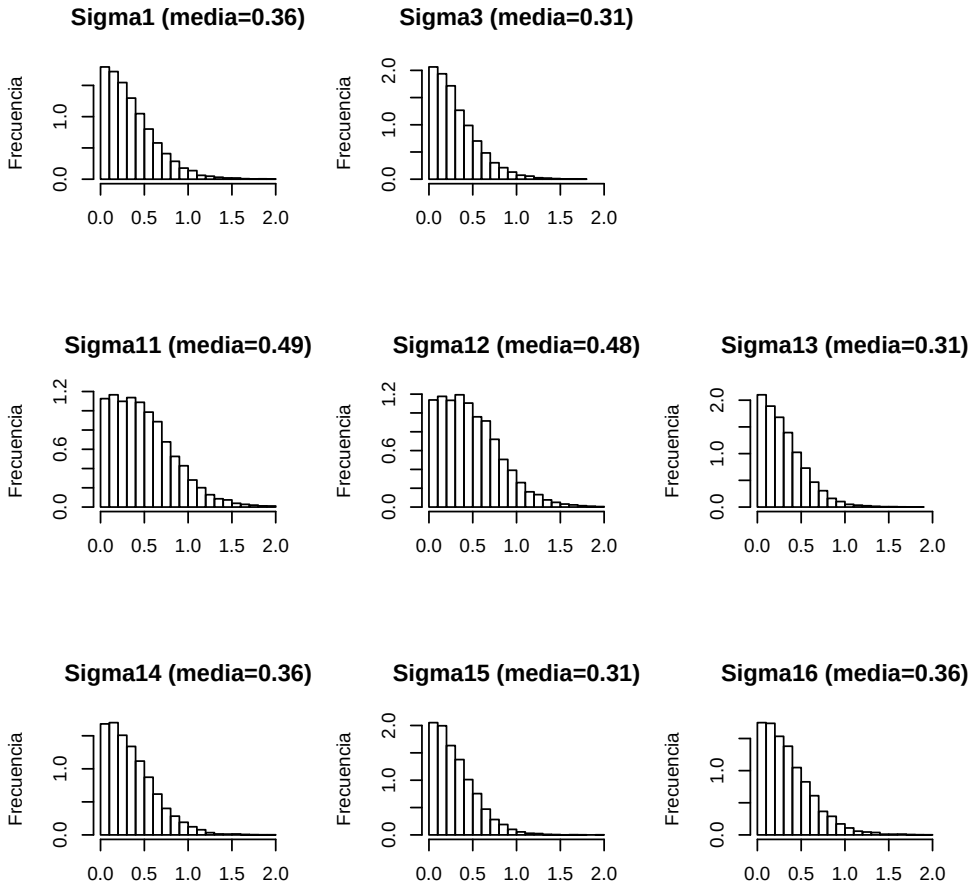

Figura 5.8: Histograma y media de σ para el primer banco de datos de los grupos 1, 3, 11, 12, 13, 14, 15, 16.

Grupo	$\overline{k-k^*}$ *Green*	$\overline{k-k^*}$ *Semipar.*	$\overline{(k-k^*)^2}$ *Green*	$\overline{(k-k^*)^2}$ *Semipar.*	*Cociente D.finales*
11	2.013	1.473	10.640	7.228	0.687
12	0.913	0.377	5.750	3.354	0.822
13	0.622	0.327	5.494	4.238	0.847
14	1.682	1.188	7.748	5.112	0.790
15	1.121	0.838	6.322	4.755	0.687
16	0.372	0.188	2.732	2.217	0.768
total	1.121	0.732	6.448	4.484	0.767

Cuadro 5.11: Estadísticos del número de componentes para cada grupo.

información sobre su valor. Señalamos que en el grupo 3, en el que los bancos de datos contienen 100 observaciones, la media estimada se sitúa más próxima a su verdadero valor que en el grupo 1 ilustrando que el aumento de tamaño puede proporcionar estimaciones más fiables de este parámetro.

En la figura 5.9 se muestran los histogramas de la distribución posterior estimada de ρ para los bancos de datos que acabamos de señalar. En casi todos ellos la distribución final de este parámetro tiene forma uniforme al igual que su distribución inicial. Por tanto, nuevamente, quedan patentes los problemas de aprendizaje sobre este parámetro tal y como se señala en Möller y Waagepetersen (2004) [70]. Los únicos casos en los que parece haber un leve aprendizaje sobre ρ son los grupos 11 y 12 en los que σ toma un valor alto y ρ se aleja levemente de los valores más pequeños.

En las dos primeras columnas del cuadro 5.11 se muestra la diferencia entre el número medio de componentes ajustadas por cada modelo y el verdadero número de componentes del término de mixtura de cada banco de datos. Se aprecia que el número de componentes es sobreestimado en ambas

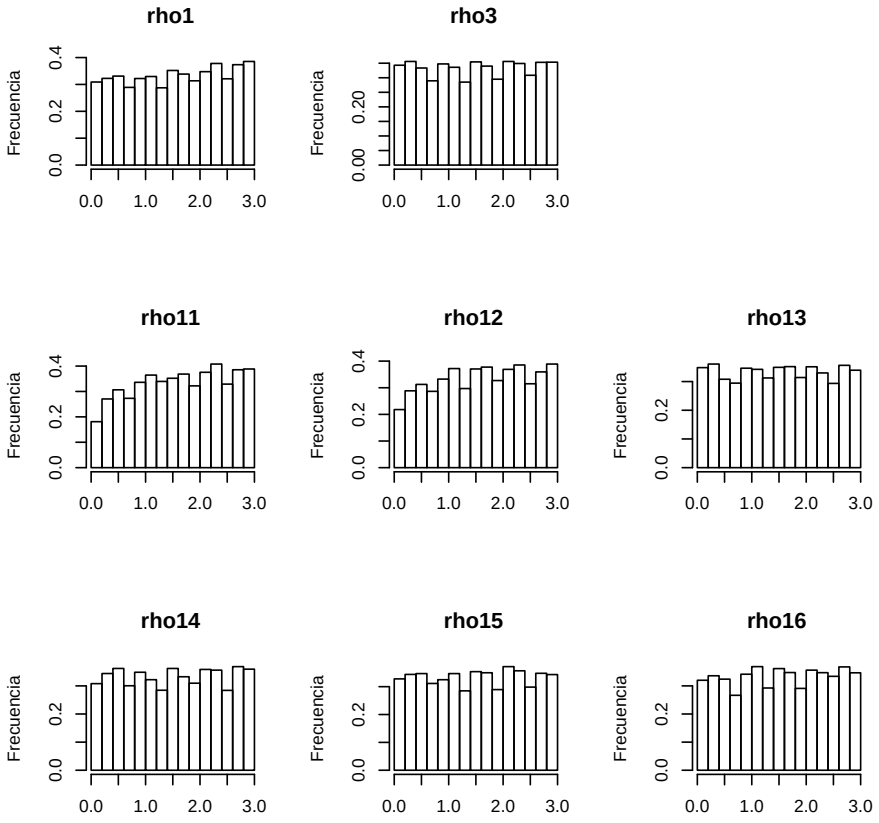

Figura 5.9: Histograma y media de ρ para el primer banco de datos de los grupos 1, 3, 11, 12, 13, 14, 15, 16.

Grupo	$P(k = k^*)$ *Green*	$P(k = k^*)$ *Semipar.*	$P(k = m^*)$ *Green*	$P(k = m^*)$ *Semipar.*
11	0.2298	0.249	0.2442	0.292
12	0.2409	0.2645	0.3177	0.3672
13	0.2077	0.2261	0.325	0.3429
14	0.2406	0.2934	0.2614	0.3114
15	0.2661	0.2841	0.3346	0.3724
16	0.2839	0.2928	0.3739	0.3952
total	0.245	0.268	0.309	0.346

Cuadro 5.12: Probabilidad posterior de que el número de componentes sea igual al número de componentes verdaderas (k^*) y al número de modas de cada conjunto de datos (m^*).

ocasiones, siendo dicha sobreestimación mayor en los procesos con mayores desviaciones típicas y con valores de ρ altos. Respecto a la variabilidad de la distribución final de k en torno a k^*, se puede observar en la tercera y cuarta columna del cuadro 5.11 que nuevamente la variabilidad es mayor en las mismas ocasiones que encontramos un sesgo superior para esta variable. Este comportamiento, en ambos algoritmos, resulta bastante razonable ya que de esta forma se compensa el sesgo con un aumento en la variabilidad de la distribución. Por último, respecto al cociente de distribuciones finales de k entre el modelo de mixturas y nuestra propuesta evaluadas en k^*, se observa que en todos los casos dicho valor tiene menor probabilidad en el modelo de mixturas que en el semiparamétrico. Además dicha diferencia entre ambos modelos se hace más patente en los grupos en que la variabilidad del proceso log-gaussiano es mayor. En esos casos, tal y como parece natural, resulta más apropiado nuestra propuesta al tener una presencia mayor el ruido log-gaussiano en la generación de los datos.

Grupo	$\#(\hat{k} = k^*)$ *Green*	$\#(\hat{k} = k^*)$ *Semipar.*	$\#(\hat{k} = m^*)$ *Green*	$\#(\hat{k} = m^*)$ *Semipar.*
11	4	5	5	6
12	3	2	6	7
13	1	2	6	5
14	4	4	6	6
15	5	5	8	8
16	3	3	6	6
total	3.33	3.50	6.16	6.33

Cuadro 5.13: Número de ocasiones en los que la moda de la distribución posterior, $\hat{k}$, coincide con el número de componentes verdaderas (k^*) o el número de modas de cada conjunto de datos (m^*) para cada modelo.

Por otro lado, en el cuadro 5.12 se puede apreciar que el modelo semiparamétrico ofrece mejores resultados en cuanto a la frecuencia en que la posterior coincide, bien con el número de componentes de la distribución que ha generado los datos, bien con el número de modas de ésta. Nuevamente se observa que ambos modelos ajustan mejor el número de modas de los datos que el número de componentes reales que los han generado. Además, también se puede observar que cuanto mayor es la variabilidad del proceso log-gaussiano y mayor es ρ, se predice de manera más deficiente el número de modas según ambos modelos. Respecto al número de componentes no se observan unos resultados tan concluyentes.

Respecto a la coincidencia entre la moda de la distribución final y el número de componentes o modas de la distribución original, en el cuadro 5.13 se muestran los resultados obtenidos. El único resultado que se desprende a partir de éstos es que la moda de la distribución posterior coincide en mayor medida con la moda de los datos más que su número de componentes.

Grupo	11	12	13	14	15	16
Divergencia	0.58	0.14	0.06	0.22	0.07	0.04

Cuadro 5.14: Diferencia media entre las divergencias del modelo de mixturas y el modelo semiparamétrico respecto a las distribuciones que han generado los datos.

El resto de hipótesis que podrían desprenderse no parecen ser demasiado concluyentes.

Por último, en el cuadro 5.14 se presenta la diferencia de las divergencias para el modelo de mixturas y nuestra propuesta. Para el cálculo de las divergencias se ha ignorado el término log-gaussiano tanto en la función de intensidad original como de las ajustadas. De esta forma se evalúa en que grado se asemejan las mixturas ajustadas por ambos modelos a la mixtura original, independientemente del ruido que tuviera ésta e independientemente del ruido estimado. En todos los grupos dicha diferencia es positiva, lo que indica que la mixtura original es ajustada de forma más satisfactorio por nuestra propuesta. Además, se observa que la diferencia toma valores superiores cuando la desviación típica del efecto log-gaussiano es mayor y para valores altos de ρ.

En resumen podemos destacar las siguientes conclusiones respecto a la modelización de datos de mixturas con ruido log-gaussiano.

- En cuanto al ajuste de las componentes de la mixtura el modelo semiparamétrico ofrece en general mejores resultados. Dicha mejora se ha podido constatar en 5 aspectos: menor sesgo de la media de la distribución final, menor variabilidad de la predicción respecto al verdadero valor, mayor probabilidad posterior de dicho valor y mayor probabili-

dad posterior del verdadero número de modas. Por el contrario no se ha evidenciado ninguna mejora en cuanto a la coincidencia de la moda de la distribución posterior con el número de componentes y modas verdaderas.

- Se ha determinado también que en todos los casos estudiados el número de componentes ajustadas por el modelo semiparamétrico es menor que para el modelo de mixturas. Este hecho corrobora la hipótesis que establecíamos con los datos de galaxias en relación al ajuste más parsimonioso de nuestra propuesta.

- En general el modelo semiparamétrico se muestra más apropiado cuanto mayor es el ruido log-gaussiano y para valores de ρ altos, es decir cuando el ruido tiene estructura espacial muy tenue.

- El análisis de la divergencia respecto a la mixtura original apunta hacia un mejor ajuste del modelo semiparamétrico.

A modo de conclusión del presente capítulo podríamos señalar que nuestra propuesta proporciona un ajuste más parsimonioso de datos provenientes de mixturas, con o sin ruido. El número de componentes ajustadas por el modelo semiparamétrico parece ajustarse más al valor original que la estimación proporcionada por el modelo de mixturas sin ruido. Además, incluso en términos de la divergencia respecto de la función de densidad original el modelo semiparamétrico tiene mejor comportamiento en presencia de ruido aleatorio. Por tanto, teniendo en cuenta que el comportamiento de este último modelo no resulta peor en ausencia de ruido encontramos particularmente aconsejable su uso siempre, ya que proporciona una estimación más parsimoniosa y más robusta ante la presencia de ruido o clustering general.

Capítulo 6

Aplicación a los brotes de Legionelosis de Alcoi

Tras describir en el capítulo 5 los resultados numéricos de la comparación del modelo de mixturas frente a nuestra propuesta, en el presente capítulo vamos a aplicar ambos modelos a los datos de los brotes de legionelosis de Alcoi introducidos en el capítulo 1. Esperamos que las ventajas que hemos encontrado en el capítulo anterior de nuestra propuesta se mantengan también sobre este caso real ya que esta propuesta ha sido ideada para aplicarse en problemas como el que ahora nos ocupa. No obstante vamos a aplicar tanto el modelo de mixturas como el modelo semiparamétrico en el estudio de todos los brotes para comparar los resultados de ambos.

La aplicación de los algoritmos descritos a los brotes de legionelosis y a cualquier otro problema epidemiológico similar requiere que contemplemos previamente cómo se distribuye geográficamente la población en el interior del núcleo urbano. Obviamente, la función de intensidad habrá de ser mayor allá donde haya mayor densidad de población susceptible de contraer la enfermedad. Para ello vamos a hacer uso de los controles que se tomaron tras

el primero de los brotes estudiados. Recordamos, tal y como se describió en el capítulo 1, que los controles son personas de características similares a los casos pero que se diferencian de estos últimos en que no presentan la enfermedad a estudio. Por tanto, la distribución geográfica de los controles nos puede servir para describir como se distribuye la población con similares características a los casos en el interior de la ciudad. El objetivo de nuestra aplicación será determinar la localización o localizaciones geográficas en las que se acumulan un número de casos anormalmente alto comparado con la distribución de los controles. Para ello el modelo semiparamétrico contemplará los dos tipos posibles de agregación descritos en este trabajo, las de tipo individual y general.

En la siguiente sección se va a describir brevemente la adaptación del algoritmo semiparamétrico bidimensional al caso que dispongamos de un conjunto de controles y que se quiera incorporar su distribución geográfica para la descripción de la distribución de los casos. Una vez se haya descrito dicha adaptación, en la siguiente sección se procederá a detallar los resultados del estudio de los distintos brotes de legionelosis con el algoritmo propuesto.

6.1. Incorporación de la información de los controles

Para la incorporación de información sobre los controles en el modelo semiparamétrico disponemos de 2 alternativas. La primera consiste en estimar la función de intensidad de los controles previamente a la modelización de los casos, e incorporar dicha función estimada al proceso de inferencia de los casos. La segunda alternativa es la estimación conjunta de la intensidad de los casos y los controles de forma simultánea. Esta última opción nos

parece más adecuada ya que incorporará la variabilidad de la estimación de la función de intensidad de los controles a la estimación de la intensidad de los casos, en caso contrario dicha variabilidad sería ignorada. Por tanto nos vamos a decantar por la segunda de las opciones expuestas.

En Lawson y Clark (1999) [62] se recurre también a la estimación simultánea de la función de intensidad de los casos y los controles. Estos autores proponen estimar la función de intensidad de los controles mediante una suavización kernel con una función núcleo normal bivariante. La estimación del ancho de banda de dicha función se realiza también dentro de la simulación MCMC, de esta forma se incorpora la incertidumbre sobre dicho parámetro al resto de parámetros del modelo. Sin embargo, en nuestra modelización hemos contemplado el uso de procesos log-gaussianos para la descripción no-paramétrica de la variabilidad en la función de intensidad de los casos. Por tanto vamos a hacer uso también de dicha propuesta para la descripción de la función de intensidad de los controles. Consideramos que el proceso log-gaussiano supone una alternativa más flexible que la estimación kernel, por tanto esperamos que proporcionará una estimación más fiable de esta función de intensidad.

Pasamos seguidamente a describir el modelo que vamos a emplear y que incorpora la información de la distribución de los controles. Si Y es el conjunto de controles muestreados de la población, consideraremos que el proceso puntual que describe la distribución de dichos controles viene dado por:

$$\begin{array}{c} Y \sim \prod_{i=1}^{|Y|} g(y_i)\exp(-\int_C g(t)dt) \\ \log(g(\{C_1,...,C_k\})) = \psi + \Sigma_g(\sigma_g,\rho_g)\Gamma_g \\ \psi \sim \mathcal{U}(-\infty,\infty) \end{array} \tag{6.1}$$

$$\begin{gathered}\Sigma_g^2(\sigma_g,\rho_g)=\sigma_g^2\exp\left(-\rho_g D\right)\\ \sigma_g\sim\mathcal{U}(0,b_\sigma)\\ \rho_g\sim\mathcal{U}(a_\rho,b_\rho)\\ (\Gamma_g)_i\sim\mathcal{N}(0,1)\ i=1,...,|C|\ ,\end{gathered}\tag{6.2}$$

por tanto el campo aleatorio g describe la función de intensidad de los casos, mediante un proceso log-gausiano. Este proceso resultará también necesario para describir geográficamente la aparición de los casos sobre la región de estudio. Por otro lado, el proceso puntual que describe la aparición de los casos viene dado por la siguiente función de intensidad:

$$\begin{gathered}X\sim\prod_{i=1}^{|X|}\left(f(x_i)\cdot g(x_i)\mathcal{N}_2(x_i|\mu_{Z_i},\Lambda_{Z_i}^2)\right)\exp(-\int\lambda(t)dt)\\ \lambda(t)=f(t)g(t)\left(\sum_{j=1}^m w_j\mathcal{N}_2(t|\mu_j,\Lambda_j^2)\right)\ ,\end{gathered}\tag{6.3}$$

donde $g(x)$ es la función que define el proceso log-gaussiano de los controles, $f(x)$ describirá las diferencias entre casos y controles debidas a procesos de agregación general, mientras que el tercero de los términos describe las agrupaciones individuales de casos alrededor de localizaciones concretas, aunque desconocidas. $f(x)$ seguirá un proceso log-gaussiano, de la misma forma que se ha descrito para el modelo semi-paramétrico bidimensional en el capítulo 4. Por tanto consideraremos que dicho proceso viene dado por:

$$\begin{gathered}\log(f(\{C_1,...,C_k\}))=\phi+\Sigma_f(\sigma_f,\rho_f)\Gamma_f\\ \phi\sim\mathcal{U}(-\infty,\infty)\\ \Sigma_f^2(\sigma_f,\rho_f)=\sigma_f^2\exp\left(-\rho_f D\right)\\ \sigma_f\sim\mathcal{U}(0,b_\sigma)\\ \rho_f\sim\mathcal{U}(a_\rho,b_\rho)\\ (\Gamma_f)_i\sim\mathcal{N}(0,1)\ i=1,...,|C|\ .\end{gathered}\tag{6.4}$$

Por último, la componente de agregación individual se define como una

mixtura de distribuciones normales bivariantes con un número indeterminado de componentes según las siguientes especificaciones:

$$\begin{gathered} Z_i \sim \mathrm{Mn}(w)\ i = 1, ..., n \\ w \sim \mathrm{Dir}(\delta, ..., \delta) \\ \mu_j \sim \mathcal{N}_2(\xi, \kappa^{-1})\ j = 1, ..., m \\ \Lambda_j^{-2} \sim \mathcal{W}(\nu, \beta^{-1})\ j = 1, ..., m \\ \beta \sim \mathcal{W}(g, h) \\ m \sim \mathcal{U}(0, M)\ . \end{gathered} \tag{6.5}$$

Las ecuaciones (6.1), (6.2), (6.3), (6.4) y (6.5) definen el modelo que vamos a utilizar. En él se describen las agrupaciones de casos en torno a ciertas localizaciones, controlando la distribución de la población en la región de estudio y las agregaciones de tipo general que podrían haber en la distribución de los casos. A continuación detallamos el algoritmo que hemos utilizado para realizar la inferencia sobre el modelo que acabamos de describir.

6.1.1. Proceso de inferencia

Tal y como resulta habitual en los modelos jerárquicos bayesianos vamos a realizar la inferencia del modelo mediante métodos de simulación MCMC. Intentaremos adaptar nuestro algoritmo bidimensional, expuesto en el capítulo 4, a la situación que se nos presenta, donde tenemos que incorporar la información de los controles para la estimación de la función de intensidad de los casos. La adaptación de dicho algoritmo al caso que nos ocupa resulta bastante sencilla, únicamente hemos de tener en cuenta 2 detalles que comentamos a continuación.

En primer lugar hemos de separar el proceso de inferencia de los casos

y el de los controles. Es decir, hemos de hacer que la inferencia sobre los controles aprenda única y exclusivamente de éstos e ignore las localizaciones donde se han observado los casos. De no ser así, la función de intensidad de los controles se adaptará a la información proporcionada por los casos. De esa forma, tal y como hemos podido comprobar, resulta muy difícil identificar las distintas componentes que componen la función de intensidad de los casos. Para ello vamos a simular el modelo (6.1) que nos proporcionará la estimación de la función de intensidad g basada únicamente en la información proporcionada por los controles. Simultáneamente, cada una de las simulaciones generadas de g se imputará sobre la expresión de la función de intensidad de los casos, (6.2), y condicionado al valor imputado se procede a estimar la intensidad λ. Así, la función de intensidad de los controles depende exclusivamente de éstos, pero a su vez estamos incorporando en la estimación de λ la variabilidad en la estimación de la intensidad de los controles. Consideramos que esta solución es bastante más satisfactoria que la imputación directa de la función de intensidad g ya que en ese caso se ignoraría la variabilidad de dicha función de intensidad proporcionando resultados excesivamente exactos. En Mwalili et al. (2005) [72] podemos encontrar un ejemplo similar en el que cada iteración de la inferencia de un modelo se imputa en la inferencia de un segundo modelo. Además el software WinBUGS en su última versión incluye una función específica para simular situaciones como la que se nos plantea.

Por otra parte, en el capítulo 4 comentábamos que, para evitar que se produjeran situaciones extrañas fuera de la región de estudio (efecto frontera), era preferible definir el proceso puntual de los casos sobre todo el plano real. Para ello, la función f que define el proceso log-gaussiano de los casos tomaba un valor distinto dentro de cada celda del grid C y un valor constante $(\exp(\phi + 0{,}5\sigma_f^2))$ fuera de él. Si queremos evitar los problemas que se comentaban en el capítulo 4 con el efecto frontera, habremos de

proceder de manera similar con la función g a la hora de estimar la función de intensidad de los casos. Es decir, g tomará un valor distinto para cada celda del grid C que cubre la región de estudio, dichos valores se estimarán a partir de la información proporcionada por los controles. Sin embargo, la función de intensidad de los casos, λ, se define sobre todo el plano real de la siguiente forma:

$$\lambda(t) = \begin{cases} \exp(g(C_t)) \cdot \exp(f(C_t)) \cdot \left(\sum_{j=1}^{m} w_j \mathcal{N}_2(t|\mu_j, \Lambda_j^2)\right) & \forall t \in C \\ \exp(\psi + \frac{1}{2}\sigma_g^2) \cdot \exp(\phi + \frac{1}{2}\sigma_f^2) \cdot \left(\sum_{j=1}^{m} w_j \mathcal{N}_2(t|\mu_j, \Lambda_j^2)\right) & \forall t \notin C \end{cases} ,$$

donde C_t denota la celda de la región C que contiene la localización t. De esta forma evitaremos los problemas ocasionados por la frontera de la región de estudio, de la misma forma que se ha conseguido en el capítulo 4 con el algoritmo que no incorpora la información de los controles.

Pasamos seguidamente a describir el proceso de simulación empleado para muestrear cada parámetro del modelo.

Simulación de ψ

La distribución posterior de este parámetro resulta:

$$P(\psi|...) = \exp\left(|Y| \cdot \psi - \int_C g(t|\psi, ...)dt\right) .$$

Muestrearemos de dicha distribución mediante el método de simulación de Metrópolis-Hastings, con función de propuesta normal. En ese caso la probabilidad de aceptación del nuevo valor, ψ^* viene dada por:

$$\exp\left(|Y| \cdot (\psi^* - \psi) - (\exp(\psi^* - \psi) - 1) \cdot \int_C g(t|\psi, ...)dt\right) .$$

Simulación de ϕ

La distribución posterior de este parámetro toma la siguiente expresión:

$$P(\phi|...) = \exp\left(|X| \cdot \phi - \int \lambda(t|\phi, ...)dt\right) .$$

Este parámetro también se puede muestrear mediante Metropolis-Hastings con una función de propuesta normal. La probabilidad de aceptación en ese caso será:

$$\exp\left(|X| \cdot (\phi^* - \phi) - (\exp(\phi^* - \phi) - 1) \cdot \int \lambda(t|\phi, ...)dt\right) .$$

Simulación de σ_g

La distribución posterior en este caso viene dada por:

$$P(\sigma_g|...) = \exp\left(\sigma_g n_y^t \cdot \exp(-\rho_g \cdot D)^{1/2} \cdot \Gamma_g - \int_C g(t|\sigma_g, ...)dt\right) 1_{[0,b_\sigma]} ,$$

donde n_y es un vector que contiene el número de controles contenidos en cada celda del grid en el que se define el proceso log-gaussiano de éstos. $I_{[0,b_\sigma]}$ valdrá 1 si σ_g pertenece al intervalo $[0, b_\sigma]$ y 0 en caso contrario. El muestreo de esta variable se realizará mediante Metropolis-Hastings con función de propuesta normal, en ese caso la probabilidad de aceptación valdrá:

$$\exp\left((\sigma_g^* - \sigma_g)n_y^t \cdot \exp(-\rho_g \cdot D)^{1/2} \cdot \Gamma_g - \int_C \left(g(t|\sigma_g^*, ...) - g(t|\sigma_g, ...)\right) dt\right) . \tag{6.6}$$

Simulación de σ_f

Su distribución posterior toma la siguiente expresión:

$$P(\sigma_f|...) = \exp\left(\sigma_f n_x^t \cdot \exp(-\rho_f \cdot D)^{1/2} \cdot \Gamma_f - \int \lambda(t|\sigma_f, ...)dt\right) 1_{[0,b_\sigma]} ,$$

donde nuevamente n_x es el número de casos en cada una de las celdas del grid de su proceso log-gaussiano. Empleamos también una caminata aleatoria normal para muestrear los nuevos valores mediante Metropolis-Hastings. La probabilidad de aceptación de σ_f^* es la siguiente:

$$\exp\left((\sigma_f^* - \sigma_f)n_x^t \cdot \exp(-\rho_f \cdot D)^{1/2} \cdot \Gamma_f - \int \lambda(t|\sigma_f^*, ...) - \lambda(t|\sigma_f, ...)dt\right) .$$

Simulación de ρ_g

La distribución posterior de este parámetro es la siguiente:

$$P(\rho_g|...) = \exp\left(\sigma_g n_y^t \cdot \exp(-\rho_g \cdot D)^{1/2} \cdot \Gamma_g - \int_C g(t|\rho_g, ...)dt\right) .$$

De la misma forma que para los algoritmos propuestos en el capítulo 4 precalcularemos la matriz $\exp(-\rho_g \cdot D)^{1/2}$ para un conjunto equiespaciado de 100 valores para ρ_g. Muestrearemos la distribución posterior de ρ_g mediante Metropolis-Hastings con función de propuesta uniforme sobre los valores precalculados. La expresión de la probabilidad de aceptación del nuevo valor ρ_g^* resulta:

$$\begin{array}{c} \exp(\sigma_g n_y^t \cdot (\exp(-\rho_g^* D)^{1/2} - \exp(-\rho_g D)^{1/2}) \cdot \Gamma_g \\ - \int_C (g(t|\rho_g^*, ...) - g(t|\rho_g, ...))dt) . \end{array} \tag{6.7}$$

Simulación de ρ_f

Su distribución posterior es la siguiente:

$$P(\sigma_f|...) = \exp\left(\sigma_f n_x^t \cdot \exp(-\rho_f \cdot D)^{1/2} \cdot \Gamma_f - \int \lambda(t|\rho_f, ...)dt\right) .$$

Al igual que para ρ_g, precalcularemos la raíz cuadrada de la matriz de correlaciones entre las celdas para 100 valores de ρ_f. Este parámetro también será muestreado mediante Metropolis-Hastings con distribución uniforme entre los 100 valores precalculados. La probabilidad de aceptación tiene la siguiente expresión:

$$\exp\Big(\sigma_f n_x^t \cdot (\exp(-\rho_f^* D)^{1/2} - \exp(-\rho_f D)^{1/2}) \cdot \Gamma_f - \int(\lambda(t|\rho_f^*, ...) - \lambda(t|\rho_f, ...))dt\Big) \ .$$

Simulación de Γ_g

La distribución posterior de este parámetro resulta:

$$P(\Gamma_g|...) = \exp\left(n_y^t \cdot \Sigma_g \cdot \Gamma_g - \frac{\|\Gamma_g\|^2}{2} - \int_C g(t)dt\right) \ ,$$

donde $\Sigma_g = \sigma_g \exp(\rho_g \cdot D)^{1/2}$. Al igual que en los algoritmos del capítulo 4 encontramos apropiado muestrear Γ_g mediante el método de Langevin-Hastings que permite la actualización conjunta de todas sus componentes con una probabilidad de aceptación razonable. Procediendo de la misma forma que se describió en el capítulo 4, la función de propuesta que habremos de utilizar toma la siguiente forma:

$$\mathcal{N}_{|C|}(\Gamma_g^*|(1 - h/2) \cdot \Gamma_g + (h/2)e_g^t \cdot \Sigma_g, hI_{|C|}) \ ,$$

donde e_g es un vector en el que la k-ésima componente tiene como expresión:

$$(n_y)_k - \left(\int_{C_k} g(t)dt\right) \ ,$$

y h es un parámetro de sintonización del proceso de simulación. En ese caso la probabilidad de aceptación de la propuesta anterior viene dada por:

$$\frac{P(\Gamma_g^*|...)\mathcal{N}_{|C|}(\Gamma_g|(1 - h/2) \cdot \Gamma_g^* + (h/2)(e_g^*)^t \cdot \Sigma_g, hI_{|C|})}{P(\Gamma_g|...)\mathcal{N}_{|C|}(\Gamma_g^*|(1 - h/2) \cdot \Gamma_g + (h/2)e_g^t \cdot \Sigma_g, hI_{|C|})} \ .$$

Simulación de Γ_f

La distribución posterior en esta ocasión tiene la siguiente forma:

$$P(\Gamma_g|...) = \exp\left(n_x^t \cdot \Sigma_f \cdot \Gamma_f - \frac{\|\Gamma_f\|^2}{2} - \int \lambda(t)dt\right) .$$

Γ_f también será muestreado mediante el método de Langevin-Hastings, así si

$$e_f = n_x - \left(\int_{C_k} \lambda(t)dt\right)_{k=1}^{|C|} ,$$

la función de propuesta para Γ_f vendrá dada por:

$$\mathcal{N}_{|C|}(\Gamma_f^*|(1-h/2)\cdot\Gamma_f + (h/2)e_f^t \cdot \Sigma_f, hI_{|C|}) .$$

En ese caso la probabilidad de aceptación viene dada por:

$$\frac{P(\Gamma_f^*|...)\mathcal{N}_{|C|}(\Gamma_f|(1-h/2)\cdot\Gamma_f^* + (h/2)(e_f^*)^t \cdot \Sigma_f, hI_{|C|})}{P(\Gamma_f|...)\mathcal{N}_{|C|}(\Gamma_f^*|(1-h/2)\cdot\Gamma_f + (h/2)e_f^t \cdot \Sigma_f, hI_{|C|})} .$$

Simulación de μ_i, Λ_i^2, Z_j, w, β, proceso de nacimiento-muerte

La simulación de los parámetros de la mixtura se realizan exactamente de la misma forma que para el modelo bidimensional del capítulo 4. La única precaución que se habrá de llevar es que, a la hora de calcular la integral de la función de intensidad para la simulación de μ_i, Λ_i^2 y el proceso de nacimiento-muerte, se habrá de tener en cuenta que ahora la función de intensidad depende del término $g(t)$, a diferencia del modelo que no incorporaba información sobre la distribución de los controles.

6.2. Resultados de la aplicación a los distintos brotes estudiados

Una vez programado el algoritmo que hemos descrito en la sección anterior, los hemos aplicado al estudio de los distintos brotes de legionelosis que tuvieron lugar en Alcoi desde septiembre de 1999 hasta noviembre de 2003. De los brotes que han tenido lugar en el periodo señalado, dos tienen una particular importancia por el número de casos que acumularon. El primero de los brotes se extiende desde el 20 de septiembre de 1999 al 27 de febrero de 2000 y constó de 36 casos. Por otro lado, el tercero cronológicamente comenzó el 16 de septiembre de 2000 y tuvo una duración de 2 meses y medio, en esta ocasión el número final de casos ascendió a 96. Este brote fue el más virulento de cuantos se dieron en este periodo de tiempo. El número total de casos acumulados durante todo el periodo considerado asciende a 212 casos de legionelosis.

Respecto a los controles escogidos, hemos tomado una única muestra de 65 individuos que intenta reproducir la distribución geográfica de la población con características similares a la de los casos. Los controles se muestrearon a partir de los ingresos hospitalarios que se registraron durante el primero de los brotes y que tuvieran motivo de ingreso distinto de neumonía. Cada uno de estos controles se tomaron apareados con un caso del primer brote según edad. Para cada caso se muestrearon 2 controles, salvo en 7 de ellos en los que únicamente se pudo muestrear un control. Cada control fue encuestado para determinar qué factores podrían diferenciarlos en mayor medida de los casos, sin embargo la información de la encuesta no se ha considerado relevante para el estudio mediante procesos puntuales. Aunque hubiera sido deseable el muestreo de controles para todos los brotes del periodo estudiado, consideramos que los controles disponibles son váli-

dos no sólamente para el primer brote sino para el resto. Esto se debe a que los controles no son más que una muestra de la población que describe su distribución en el núcleo urbano y ésta debería ser independiente del brote.

Respecto a la distribución de casos y controles en el interior del núcleo urbano, en la figura 6.1 hemos representado la localización de cada uno de ellos. En dicha representación hemos diferenciado entre los casos de los brotes 1 y 3 del resto, ya que según hemos comentado éstos van a ser estudiados por separado. En la figura 6.1 se pueden observar algunas localizaciones que acumulan más casos de los que parecería razonable si no existiera ningún mecanismo que los agregara, como por ejemplo en el suroeste de la ciudad donde se aprecia claramente un cluster de casos, pertenecientes en gran parte al tercer brote. Además también se observa que en la zona sureste de la ciudad la concentración de casos es bastante menor que en el resto de la ciudad. En los trabajos Abellán et al. (2002) [1] y Martínez-Beneito et al. (2005) [66] se puede encontrar más información sobre el carácter agregado de la distribución de los casos en los distintos brotes y la estimación no paramétrica de su función de intensidad. En la representación anterior también se puede observar la particular orografía de la ciudad, atravesada por distintos barrancos y el cauce del rio Serpis que hacen que la distribución de la población por el núcleo urbano no sea en absoluto uniforme.

A la hora de realizar el análisis mediante el algoritmo descrito en la sección anterior necesitaremos definir un grid de celdas sobre un territorio que cubra todo el núcleo urbano. La opción más sencilla corresponde a un grid con forma rectangular, sin embargo en la figura 6.1 se puede observar que Alcoi no tiene para nada dicha forma. Por tanto, si tomamos un grid rectangular que cubra Alcoi, un gran número de celdas corresponderá a territorio fuera del núcleo urbano, por lo que esa opción resultará muy ineficiente. Pa-

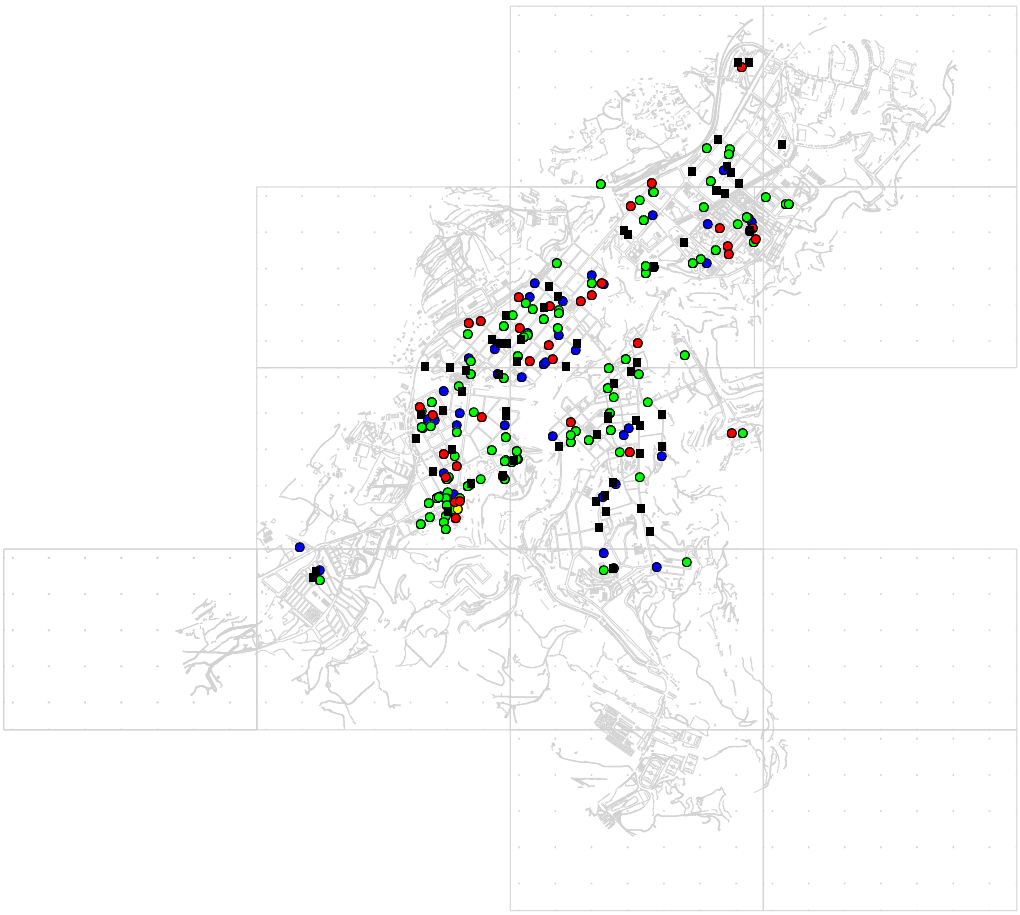

Figura 6.1: Distribución de casos y controles en la ciudad de Alcoi. Los círculos de color rojo corresponden a los casos del brote 1, los verdes al brote 3 y los azules el resto de casos observados en el periodo de estudio. Los cuadrados negros corresponden a los controles del estudio.

ra solucionar este aspecto vamos a rotar previamente las coordenadas de los casos y controles en torno a su baricentro de forma que tras dicha rotación los datos se adapten lo mejor posible a un rectángulo. Dicha rotación viene definida por la descomposición en valores singulares de las coordenadas de los casos y controles. En la figura 6.2 se pueden observar las coordenadas de los casos y controles tras la rotación efectuada. En el nuevo sistema de coordenadas el origen corresponde al baricentro de los casos y controles disponibles, mientras que la escala de los ejes corresponde a la escala real en kilómetros, es decir tanto la coordenada (1,0) como la (0,1) distan un kilómetro del baricentro de los casos y controles. El primero de los ejes coordenados en la representación 6.2 corresponde a una dirección que atravesaría la ciudad de sur-oeste a nor-este. Esta dirección es la que mayor varianza explica del conjunto de localizaciones disponible. La segunda de los ejes corresponde a una línea perpendicular a la anterior que atraviesa la ciudad de sur-este a nor-oeste. Tal y como se puede observar, la nube de puntos obtenida tras la rotación se adapta mejor a un rectángulo que la nube de puntos de la figura 6.1.

Respecto a los hiperparámetros que se han utilizado para aplicar el modelo a los brotes de legionelosis hemos empleado las siguientes especificaciones: los hiperparámetros del modelo de mixturas corresponden a las especificaciones propuestas en Stephens (1999) [91] para los modelos de mixturas normales en 2 dimensiones. Estos son una adaptación directa de los hiperparámetros propuestos en Richardson y Green (1997) [80] para el modelo unidimensional, con la única diferencia que en el modelo bidimensional los parámetros que controlan la variabilidad de las matrices de varianza-covarianza se han tomado ligeramente mayores. Así, en el caso bidimensional $\alpha = 3$ (cuando antes tomaba un valor de 2) y $g = 0{,}3$ (cuando antes tomaba un valor de 0.2). Stephens propone este cambio con la intención de ser ligeramente más informativo en el algoritmo bidimensional. Aunque

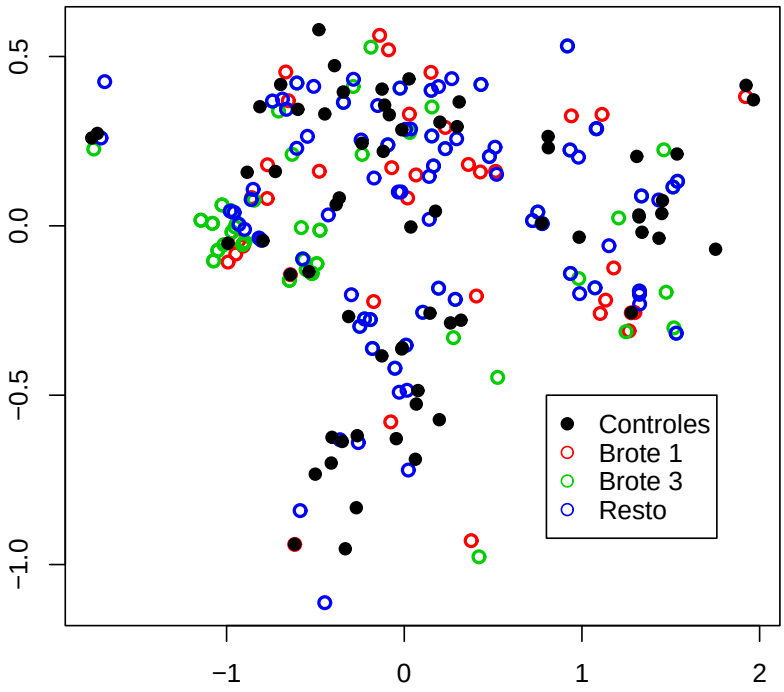

Figura 6.2: Distribución de casos y controles sobre el nuevo eje coordenado.

no ha sido comprobado directamente sobre nuestros datos, Stephens argumenta que este cambio mejora el comportamiento del modelo y por eso lo hemos incluido nosotros también. En relación a los hiperparámetros de los procesos log-gaussianos se ha tomado $b_\sigma = 4$ como límite superior para las desviaciones típicas de ambos procesos.

Para la simulación de las distribuciones posteriores se han generado en cada estudio 2 cadenas de 10.000 simulaciones, de las que se han excluido las 5.000 primeras como periodo de calentamiento. De las 5.000 iteraciones restantes hemos guardado una de cada 10 iteraciones para reducir los requerimientos de memoria, por tanto las muestras de las distribuciones posteriores de los parámetros constan de 1.000 valores. El estudio de la convergencia se ha efectuado de la misma forma que en el modelo unidimensional del capítulo 4. Es decir, en primer lugar se ha realizado una visualización de las cadenas simuladas y posteriormente se ha hecho uso del estadístico de Brooks-Gelman-Rubin sobre todas las variables del proceso de mixturas. También se ha calculado dicho estadístico para evaluar la correlación, desviación típica y una muestra de 10 efectos aleatorios de los procesos log-gaussianos. Hemos considerado que la simulación ha convergido si todos los estadísticos obtenidos eran inferiores a 1.1.

Hemos utilizado el mismo grid para definir ambos procesos log-gaussianos, el de la intensidad de los controles y el del proceso semiparamétrico que define la distribución de los casos. Si $R1$ es el rango de la primera coordenada de casos y controles ($\chi = \{X_1, ..., X_{212}, Y_1, ..., Y_{65}\}$) la primera dimensión del grid comprende el rango:

$$[min(\chi) - 0{,}1 \cdot R1, max(\chi) + 0{,}1 \cdot R1] \ .$$

El rango de la segunda dimensión del grid se ha establecido de la misma forma. Ambas dimensiones del grid se han dividido en 30 intervalos iguales,

por tanto se han utilizado 900 celdas en total para definir el proceso log-gaussiano. Cada una de estas celdas tiene aproximadamente 150 metros de longitud en la dirección del eje x, y alrededor de 60 metros de longitud sobre el eje y.

En todos los brotes estudiados se han implementado el modelo descrito en la sección anterior, (en adelante modelo semiparamétrico), así como el mismo modelo pero sin componente log-gaussiana para la distribución de los casos (en adelante modelo de mixturas). De esta forma, al igual que en el capítulo 4, vamos a comparar el comportamiento del modelo que incluye la componente log-gaussiana y la que no la incluye. Así podremos evaluar hasta que punto dicha componente influye o no sobre los resultados obtenidos.

Hemos aplicado el algoritmo propuesto en 3 casos diferentes. En primer lugar se ha aplicado al estudio del primer brote, que tal y como hemos señalado constaba de 36 casos. En segundo lugar se ha vuelto a aplicar al estudio del tercer brote (cronológicamente) así se estudian por separado los 2 brotes más virulentos de cuantos se han dado en Alcoi en el periodo estudiado. Por último, se ha aplicado el algoritmo al estudio de los 212 casos que se han recogido en todo el periodo. Este último análisis se debe a que resulta lógico pensar que algunas de las fuentes de infección ha podido actuar en más de uno de los brotes y por tanto el análisis conjunto de éstos proporcionará resultados más precisos sobre la localización de las fuentes de difusión de la bacteria.

6.2.1. Estudio del brote 1

A continuación presentamos los resultados obtenidos del estudio del primero de los brotes, que abarca el periodo desde el 20 de septiembre de 1999

hasta el 27 de febrero de 2000 y constó de 36 casos.

En primer lugar, figura 6.3, presentamos los resultados de la inferencia sobre los parámetros del proceso log-gaussiano que describe la intensidad de los controles en el modelo semiparamétrico. En la parte izquierda de la representación se muestra el comportamiento de las cadenas simuladas, mientras que en la parte derecha se muestra el histograma correspondiente a ambas simulaciones. En dicha representación se aprecia que ya se ha alcanzado la convergencia de ambos parámetros en el momento que empezamos a recoger los valores simulados. La autocorrelación de orden 1 de las cadenas para el parámetro de desviación típica es 0.764, mientras que para el parámetro de correlación espacial ésta vale 0.571. La distribución posterior de la desviación típica tiene como media 1.64 y su intervalo de credibilidad al 95 % varía entre 1.14 y 2.26. Tal y como se puede apreciar el límite superior establecido en la distribución inicial (4), no condiciona los valores obtenidos en la distribución final. Por otro lado el parámetro de correlación tiene una media de 2.24 y sus límites del intervalo de credibilidad al 95 % corresponden a 1.86 y 3.21. Para este parámetro se había establecido una distribución inicial uniforme entre 1.86 y 11.27, por tanto vemos que la distribución final ha aprendido en gran medida de los datos disponibles. Observamos que este hecho no se daba en el algoritmo unidimensional donde el parámetro de correlación apenas aprendía de los datos, aunque en ese caso la presencia de la mixtura puede hacer más difícil el aprendizaje sobre el parámetro de correlación espacial. El extremo inferior del intervalo de credibilidad para este parámetro, según podemos observar, toca el límite inferior de su distribución a priori, por tanto la distribución inicial condiciona la distribución posterior del parámetro. Sin embargo no encontramos alternativa al límite inferior que hemos considerado ya que éste obliga a que la correlación entre las celdas más alejadas del grid sea inferior a 0.01 y si no exigimos esta condición encontramos problemas para

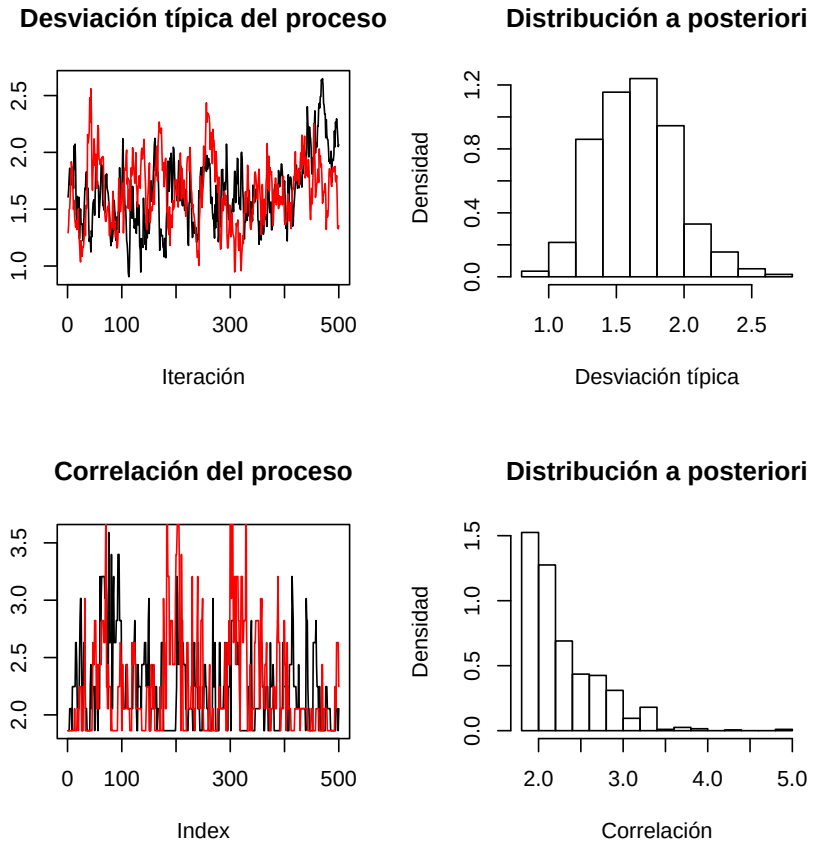

Figura 6.3: Análisis brote 1 mediante el modelo semiparamétrico. Inferencia sobre la desviación típica y la correlación del proceso log-gaussiano de los controles.

identificar la media del proceso log-gaussiano, al igual que en el problema unidimensional. Por tanto mantenemos la distribución inicial que habíamos planteado en un primer momento.

Los valores obtenidos del parámetro de correlación sugieren que la correlación espacial entre celdas es bastante alta, por tanto la estimación de la función de intensidad de los controles tendrá una forma muy suavizada. En la figura 6.4 podemos observar dicha estimación en la que las celdas de color amarillo indican mayor intensidad del proceso que genera los controles, mientras que las celdas rojas corresponden a zonas de menor intensidad. En la figura aparecen también los controles en color negro, mientras que los casos corresponden a los puntos de color azul. Tal y como se puede observar, la superficie de intensidad parece adaptarse bastante bien a la distribución de los controles, ya que las celdas más amarillas se corresponden con las zonas con mayor presencia de éstos. Además, comparando la figura 6.4 con la figura 6.1 se puede apreciar cómo la función de intensidad reproduce la orografía del Alcoi, ya que alrededor del cauce del rio, barrancos y puentes la estimación de dicha función suele ser menor, tal y como cabría esperar. También se puede observar cómo la función de intensidad se adapta a la localización de los controles pero no a la de los casos, ya que en la localización de éstos la función de intensidad no toma necesariamente valores altos. Valga como ejemplo la función de intensidad estimada en la celda que contiene al caso situado en la parte inferior y ligeramente desplazado a la derecha en la figura 6.4.

En caso de emplear el modelo de mixturas en lugar del semiparamétrico, la media a posteriori de la desviación típica del proceso que describe a los controles vale 1.65 y su intervalo de credibilidad al 95 % [1.06,2.33]. Para el parámetro de correlación, la media de la distribución posterior vale 2.25 y su intervalo de credibilidad al 95 % [1.86,3.59]. Estos valores son muy similares

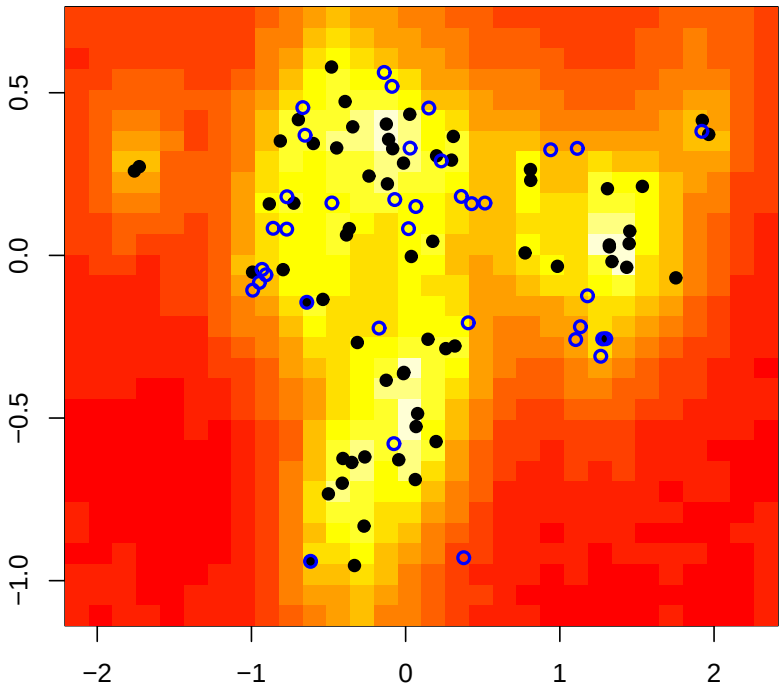

Figura 6.4: Análisis brote 1 mediante el modelo semiparamétrico. Media posterior estimada en cada celda de la función de intensidad de los controles. Los puntos negros representan la localización de los controles, los azules la localización de los casos

Componentes	1	2	3	4	5	6	7	8
Frecuencia	0.400	0.401	0.129	0.040	0.021	0.005	0.003	0.001

Cuadro 6.1: Análisis brote 1. Frecuencia en la distribución posterior del número de componentes del modelo semiparamétrico. Cadenas e histograma de la distribución posterior.

a los obtenidos en nuestra propuesta. De hecho las diferencias entre ambas estimaciones se deben única y exclusivamente al error de Monte Carlo, ya que la estimación de la intensidad de los controles depende sólamente de éstos y no depende para nada de los casos ni del modelo que se haya utilizado para la descripción de su intensidad. No incluimos la estimación de la función de intensidad de los controles según el modelo de mixturas, ya que tal y como hemos comentado ésta es muy similar a la presentada en la figura 6.4 y sus diferencias se deben exclusivamente a error de Monte Carlo.

A continuación nos vamos a centrar en los parámetros que describen la distribución de los casos en el modelo semiparamétrico. En la figura 6.5 se presentan las cadenas para el número de componentes de la mixtura, así como su histograma. La autocorrelación de orden 1 de dichas cadenas es de 0.644. La media de la distribución posterior de k es de 1.913, y su intervalo de credibilidad al 95 % varía entre 1 y 5. En el cuadro 6.1 se muestran las frecuencias relativas obtenidas de la simulación. Tal y como se puede observar el 80 % de la masa de la distribución posterior se sitúa entre 1 y 2 componentes. Por tanto parece que el modelo apunta hacia la existencia de a lo sumo 2 o 3 agrupaciones de casos.

Respecto a la estimación de la matriz β, el valor esperado de las matrices de varianza-covarianza de las componentes de la mixtura, la media de su

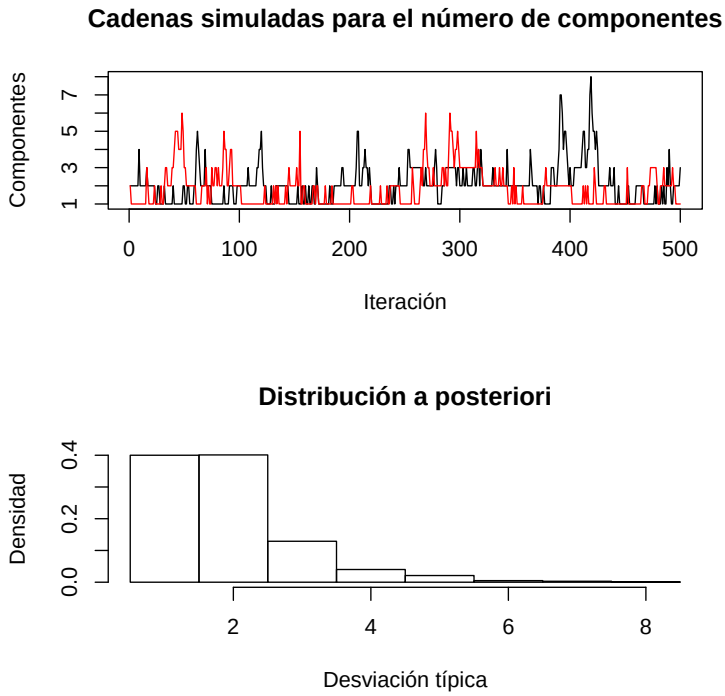

Figura 6.5: Análisis brote 1 mediante el modelo semiparamétrico. Distribución posterior del número de componentes de la mixtura en el modelo semiparamétrico.

distribución posterior coincide con

$$\begin{pmatrix} 1{,}36 & 0{,}07 \\ 0{,}07 & 0{,}24 \end{pmatrix},$$

Teniendo en cuenta que el hiperparámetro ν (los grados de libertad de las matrices de precisión de las componentes de la mixtura) vale 3 y que el valor esperado para dichas matrices es $\nu \cdot \beta$, el valor esperado de la desviación típica sobre el primer eje asciende a 0.67 kilómetros, sobre el segundo eje se mantiene en 0.28 kilómetros y una correlación entre ambos ejes de 0.13. La autocorrelación de las cadenas es de 0.12 para la varianza de la primera componente, 0.13 para el parámetro de covarianza y 0.24 para la segunda componente.

En la figura 6.6 se observa en color rojo los valores simulados de las medias de las componentes de la mixtura. Por tanto, en dicha representación se aprecia la distribución posterior de las medias, independientemente del número de componentes de la mixtura resultantes de la simulación. Parece que en la zona central existe una densidad mayor de componentes, mientras que alrededor del punto (1,-0.4) se apunta la existencia de la segunda componente que señalábamos anteriormente.

En la figura 6.7 se ha definido un grid de 50 componentes horizontales y 50 verticales sobre la región de estudio y se ha coloreado cada una de sus celdas según el número de componentes de la figura 6.6 que recaen en cada celda o en sus colindantes. Nos referiremos a partir de ahora a esta figura como *Intensidad suavizada de las componentes*. En esta representación las celdas de color amarillo corresponden a las localizaciones con mayor densidad de componentes. Así, se puede apreciar de forma más clara que en la figura 6.6 la distribución posterior de las componentes del modelo de mixturas.

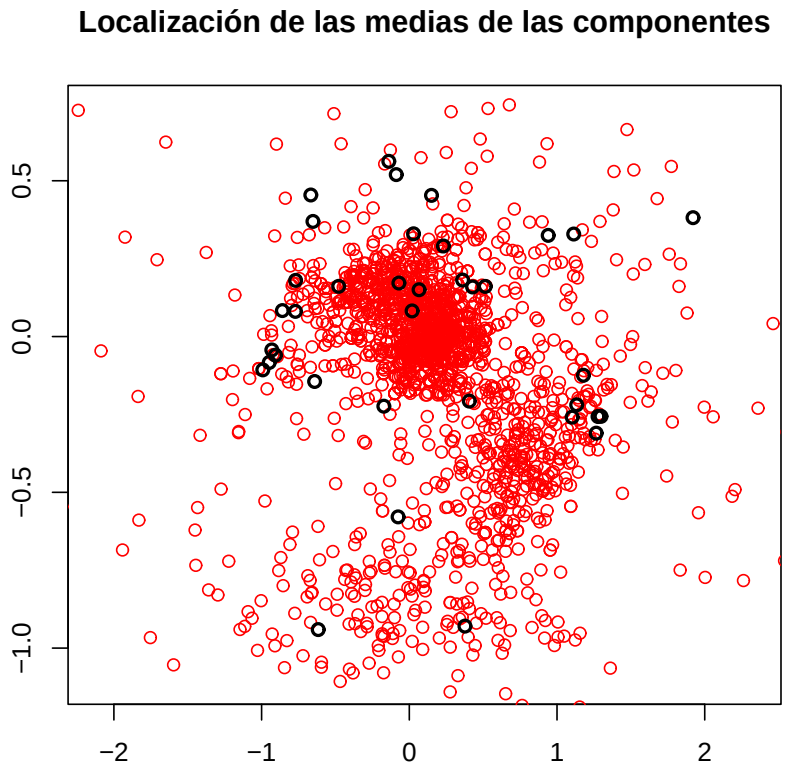

Figura 6.6: Análisis brote 1 mediante el modelo semiparamétrico. Distribución final de las medias de las componentes. En rojo las localizaciones de las medias resultantes de la simulación, en negro los casos del brote 1.

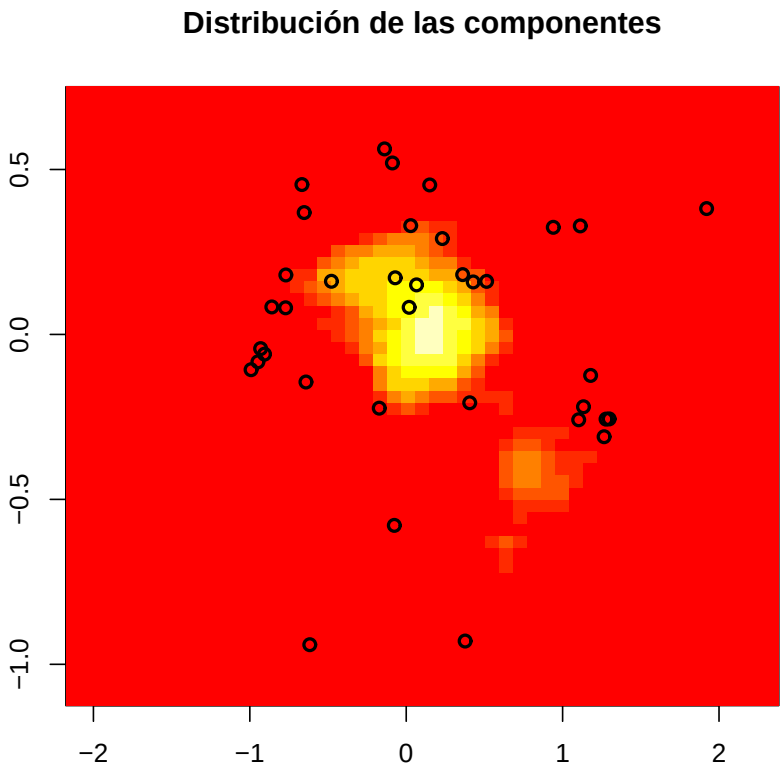

Figura 6.7: Análisis brote 1 mediante el modelo semiparamétrico. Distribución final de las medias de las componentes. El color de cada celda representa el porcentaje de simulaciones que ha recaído una componente en ella o en cualquier celda colindante.

En la figura 6.7 se aprecia que la mayor densidad de componentes de la mixtura se sitúa alrededor del punto (0,0), el baricentro de las localizaciones observadas. También se observa una intensidad más elevada en la parte superior izquierda de la moda situada en (0,0) y alrededor del punto (1,-0.4), aunque esta última componente parece extenderse hacia la parte inferior izquierda de la representación. Por tanto, desde un punto de vista epidemiológico, las instalaciones de riesgo situadas en las zonas con tonalidad amarilla deberían ser objeto de una especial atención y vigilancia.

Puede parecer un tanto extraño la alta probabilidad de observarse alguna componente alrededor de la localización (1,-0.4). En principio podría parecer más natural que dicha componente se situara alrededor de (1.1,-0.3) que es donde observamos una pequeña acumulación de casos. Sin embargo, teniendo en cuenta la distribución de los controles en la figura 6.2, observamos que la población de esa zona vive en las coordenadas del eje y superiores a -0.3 y que por debajo de esa coordenada no vive nadie en esa zona. El modelo propuesto sitúa la componente en la zona deshabitada, ya que si la situara directamente en (1.1,-0.3), esperaríamos haber observado más casos alrededor de las coordenadas (1.1,-0.2) o (1.1,0), por ejemplo. Por tanto, como este cluster tiene un comportamiento mucho más localizado que el otro (cuando el modelo penaliza que ambos sean muy diferentes), la solución que se propone es desplazar la localización del cluster al área deshabitada. Así se explica el que este segundo cluster sea bastante más pequeño, ya que de esta forma sería consecuencia del efecto parcial de una fuente de riesgo alejada de la población.

Respecto a la estimación de los parámetros del proceso log-gaussiano de los casos, en la figura 6.9 se observan las cadenas simuladas y los histogramas de la distribución posterior para la desviación típica y la correlación espacial. La autocorrelación de orden 1 para la desviación típica asciende a 0.82 y

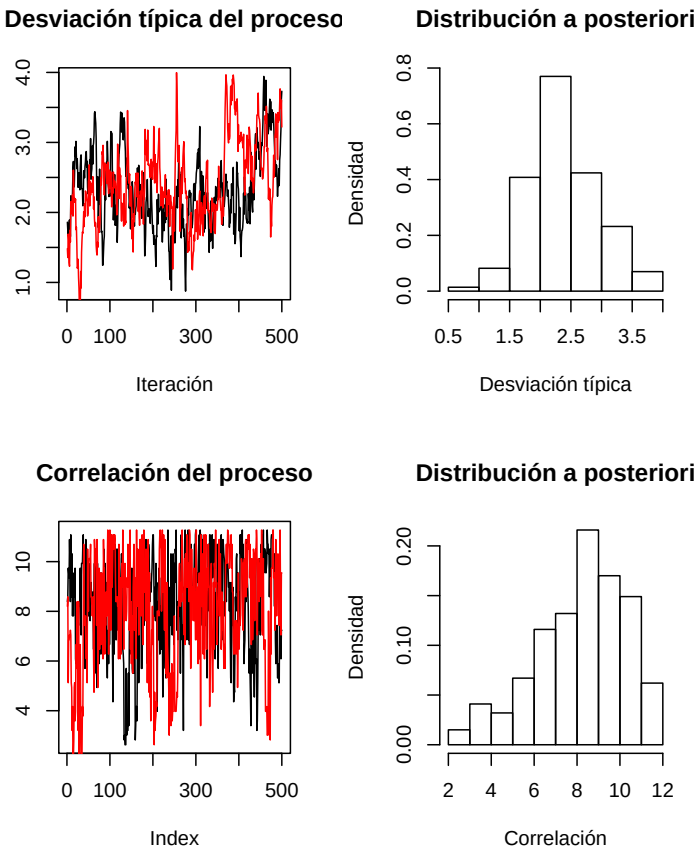

Figura 6.8: Análisis brote 1 mediante el modelo de mixturas. Inferencia sobre la desviación típica y la correlación del proceso log-gaussiano de los casos.

para el parámetro de correlación a 0.47. La media de la distribución posterior de la desviación típica es de 2.36 y su intervalo de credibilidad al 95 % [1.34,3.58]. Por otra parte el parámetro de correlación toma como media 8.17 e intervalo de credibilidad al 95 % [3.20,11.27]. Así, la correlación entre celdas en este proceso es mucho menor que para el proceso que describe la distribución de los controles, por tanto se espera una menor correlación espacial que en ese caso.

En la figura 6.9 se observan los valores que ha tomado el proceso log-gaussiano de los casos sobre las celdas del grid. Se aprecia cómo dicho proceso se adapta a las localizaciones donde se ha observado algún caso de legionelosis. De la misma forma que las componentes del modelo de mixturas no parecían adaptarse demasiado bien a la localización de los casos, en el proceso log-gaussiano ocurre justo lo contrario ya que su comportamiento parece que tiene un comportamiento muy poco parsimonioso. Dicho comportamiento se ve corroborado también por la estimación que hemos obtenido de la correlación del proceso log-gaussiano.

Parece que el modelo semi-paramétrico no encuentra demasiadas evidencias de agregación individual dada la configuración espacial de los casos, sino que por el contrario explica casi toda la variabilidad espacial mediante agregación de tipo general.

A continuación pasamos a describir los resultados obtenidos de la aplicación del modelo de mixturas a este brote. En la figura 6.10 se puede observar la distribución posterior del número de componentes de la mixtura. La autocorrelación de orden 1 de la muestra de la distribución posterior es de 0.74. La media a posteriori del número de componentes es 3.25, y su intervalo de credibilidad al 95 % va desde 1 hasta 9. En el cuadro 6.2 se presentan las frecuencias obtenidas para cada valor de la distribución posterior de k. En

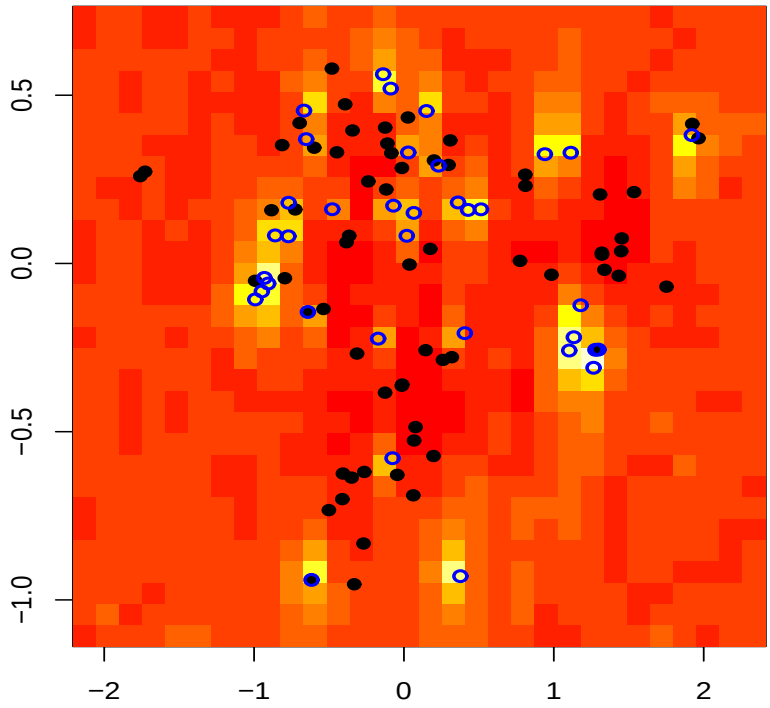

Figura 6.9: Análisis brote 1 mediante el modelo de mixturas. Media estimada del proceso log-gaussiano de los casos para cada celda. Los puntos negros representan la localización de los controles, los azules la localización de los casos.

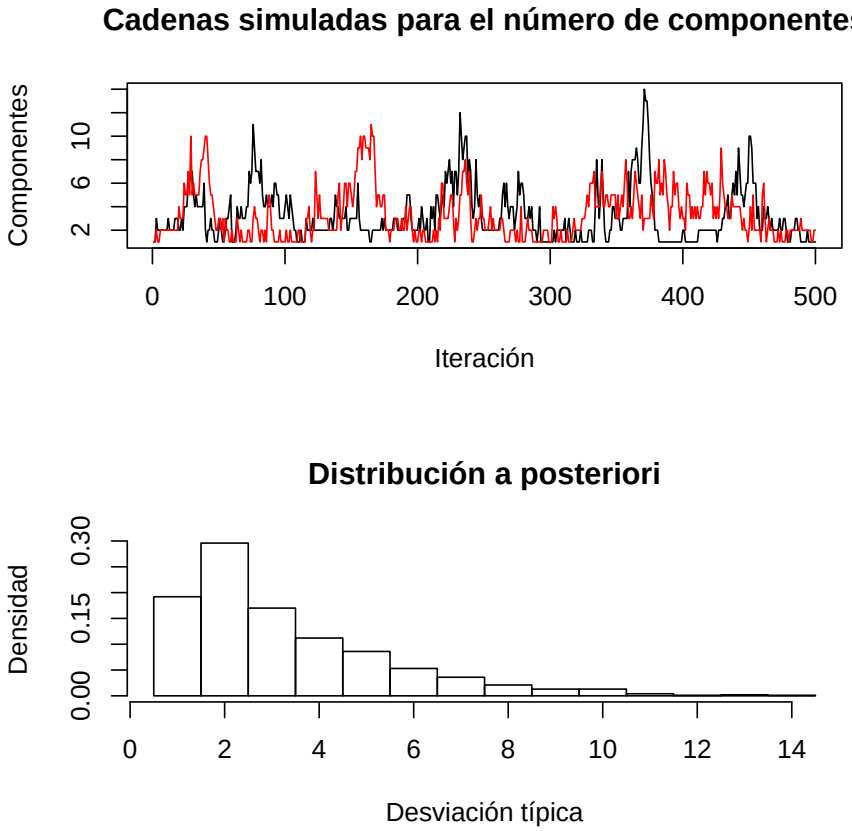

Figura 6.10: Análisis brote 1 mediante el modelo de mixturas. Distribución posterior del número de componentes de la mixtura.

Componentes	1	2	3	4	5	6	7
Frecuencia	0.192	0.296	0.170	0.112	0.086	0.053	0.036
Componentes	8	9	10	11	12	13	14
Frecuencia	0.021	0.013	0.013	0.004	0.001	0.002	0.001

Cuadro 6.2: Análisis brote 1 mediante el modelo de mixturas. Frecuencia en la distribución posterior del número de componentes del modelo de mixturas.

este caso la variabilidad es bastante superior a la del modelo semiparamétrico, aunque nuevamente la moda de la distribución recae en 2. Al igual que en el capítulo 5 podemos observar que el proceso log-gaussiano restringe la variabilidad en la función de intensidad debida a la presencia de casos y de esta forma evita el que se abuse de componentes de la mixtura para explicar la presencia de cada uno de éstos.

En cuanto a la matriz β, la media de su distribución posterior toma el siguiente valor:

$$\begin{pmatrix} 1{,}34 & 0{,}09 \\ 0{,}09 & 0{,}24 \end{pmatrix} .$$

En este caso el valor esperado de la desviación típica sobre el primer eje desciende a 0.67 kilómetros, sobre el segundo eje se mantiene en 0.29 kilómetros y la correlación entre ambos ejes toma 0.17 como valor. Aunque, tal y como se puede observar los valores obtenidos para la matriz β en ambos modelos son casi idénticos. La autocorrelación de las cadenas simuladas es de 0.28 para la varianza de la primera componente, 0.27 para la covarianza y 0.38 para la varianza de la segunda componente.

En la figura 6.11 se observan las localizaciones simuladas de las medias de las componentes de la mixtura. A la vista de la representación parece que

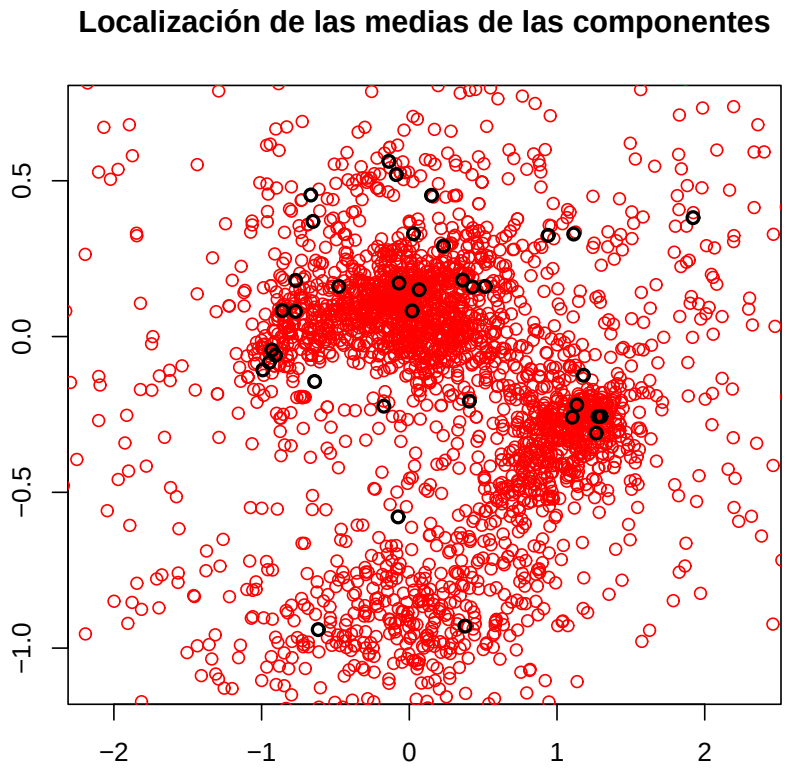

Figura 6.11: Análisis brote 1. Distribución final de las medias de las componentes para el modelo de mixturas. En rojo las localizaciones de las medias resultantes de la simulación, en negro los casos del brote 1.

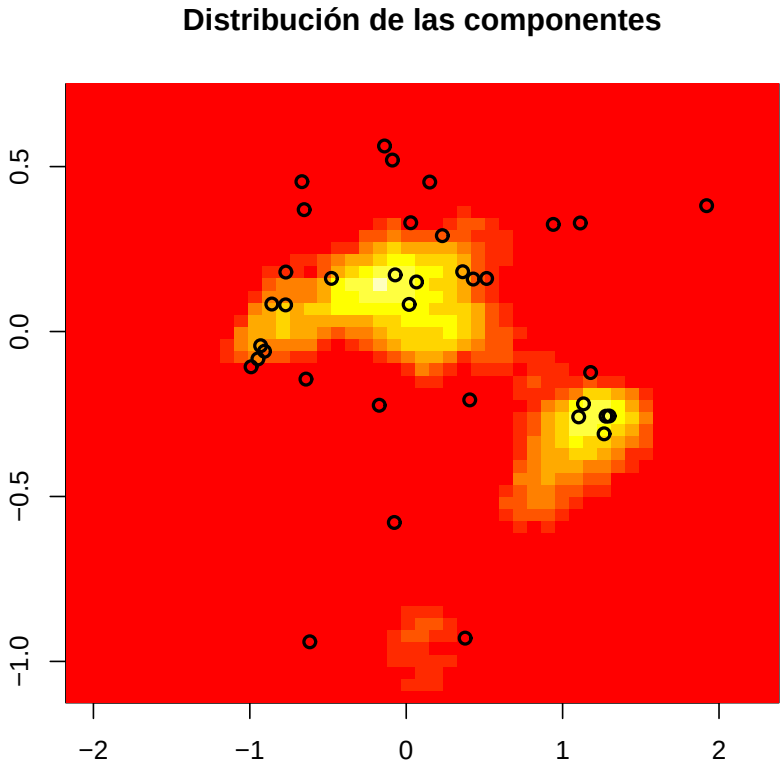

Figura 6.12: Análisis brote 1. Distribución final de las medias de las componentes para el modelo de mixturas. El color de cada celda representa el número de simulaciones que hay una componente en ella o en cualquiera de sus colindantes.

la distribución posterior de dichas componentes es muy similar a la obtenida con la propuesta semiparamétrica. No obstante, se puede apreciar que en esta ocasión las localizaciones de las componentes se adaptan más a las localizaciones de los casos. Este hecho se debe a que en el modelo de mixturas no se ha contemplado ninguna otra fuente de variabilidad alternativa, por tanto la presencia de cualquier caso sólo se puede explicar mediante la presencia de una componente relativamente próxima.

En la figura 6.12 se presenta la intensidad suavizada de las componentes para el modelo de mixturas. En ella se corroboran los comentarios del párrafo anterior ya que se puede observar que las manchas amarillas se adaptan más a la presencia de casos que en el modelo semiparamétrico. Este modelo apunta ligeramente hacia la presencia de una tercera componente alrededor de la coordenada (-0.9,0) o incluso una cuarta en la parte inferior de la representación.

6.2.2. Estudio del brote 3

El tercero de los brotes ocurridos acumuló un total de 96 casos. Este brote se extendió desde el 16 de septiembre de 2000 hasta el 1 de diciembre de ese mismo año. Este brote constó de 3 ondas epidémicos separadas entre sí por unos pocos días pero que vamos a estudiar conjuntamente.

Tanto para este estudio como para el análisis conjunto de todos el periodo se han utilizado los mismos controles que para el brote 1. Como la estimación de la función de intensidad de los controles se realiza independientemente de la estimación de los casos, dicha intensidad será la misma en los tres estudios contemplados. Así pues, ni en esta ocasión ni en el análisis conjunto vamos a presentar los resultados de la estimación de la intensidad

Modelo semiparamétrico							
Componentes	1	2	3	4	5	6	7
Frecuencia	0.120	0.162	0.322	0.267	0.082	0.030	0.006
Componentes	8	9	10	11	12	13	14
Frecuencia	0.010	0	0	0.001	0	0	0
Modelo de mixturas							
Componentes	1	2	3	4	5	6	7
Frecuencia	0	0.481	0.213	0.120	0.100	0.049	0.020
Componentes	8	9	10	11	12	13	14
Frecuencia	0.007	0.004	0.001	0.001	0.003	0.001	0

Cuadro 6.3: Análisis brote 3. Frecuencias de la distribución final del número de componentes de la mixtura para ambos modelos.

de los controles, ya que básicamente coinciden con los presentados en el estudio del brote 1.

Respecto a la estimación del número de componentes de la mixtura según el modelo semiparamétrico y el de mixturas, en la figura 6.13 se puede observar la distribución posterior de este parámetro en cada uno de ellos. Para nuestra propuesta la media a posteriori vale 3.20, y su intervalo de credibilidad al 95 % va de 1 hasta 6. En el caso del modelo de mixturas la media a posteriori cambia a 3.18, mientras que su intervalo de credibilidad al 95 % va de 2 a 8. En el cuadro 6.3 se presenta la distribución posterior del número de componentes para los dos modelos. Tal y como se puede observar, el modelo de mixturas apunta de manera más precisa hacia un brote bifocal, sin embargo la cola derecha de su distribución es mucho más amplia que la del modelo semiparamétrico. Por tanto, en esta ocasión parece que nuestra propuesta es capaz de evitar el exceso de componentes que se da ocasionalmente en el modelo de mixturas, aunque añade incertidumbre a

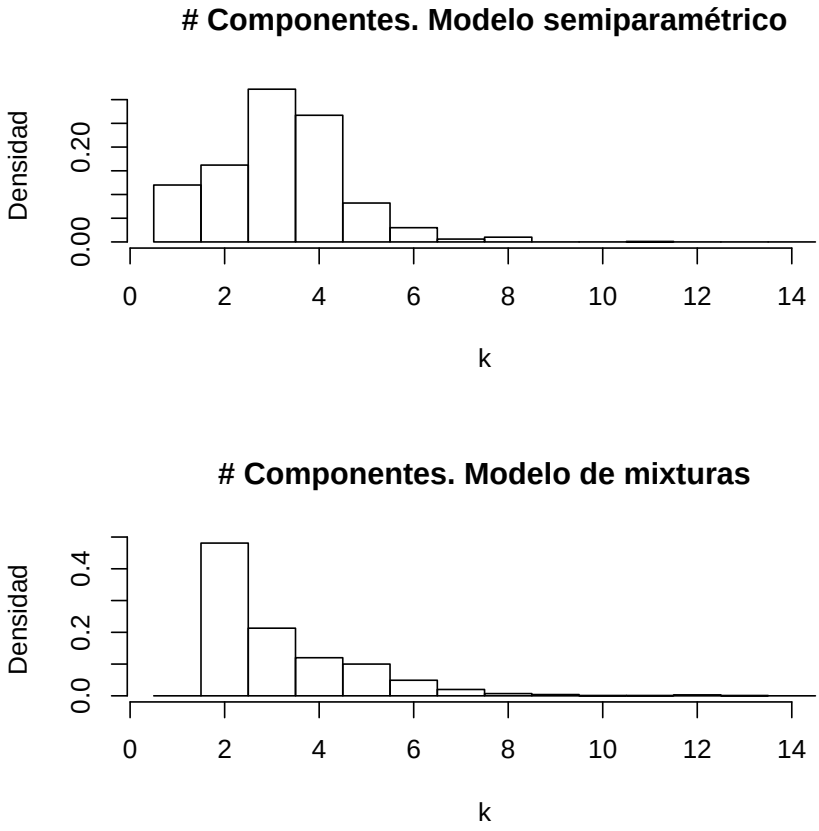

Figura 6.13: Análisis brote 3. Distribución posterior del número de componentes de la mixtura en el modelo semiparamétrico y en el de mixturas.

la hora de precisar dicho valor. La autocorrelación mostrada por las cadenas ha sido de 0.78 para el proceso semiparamétrico y de 0.77 para el modelo de mixturas.

En cuanto a β, la media de las matrices de varianza-covarianza de las componentes de la mixtura, en el proceso semiparamétrico la media a posteriori de esta matriz ha resultado

$$\begin{pmatrix} 0{,}745 & -0{,}136 \\ -0{,}136 & 0{,}073 \end{pmatrix}.$$

Así pues, el valor esperado de la desviación típica sobre el primer eje es de 0.496 kilómetros, de 0.155 kilómetros sobre el segundo y existe una correlación de -0.58 entre ambos ejes. Para el modelo de mixturas la media a posteriori de β resulta:

$$\begin{pmatrix} 0{,}481 & -0{,}120 \\ -0{,}120 & 0{,}053 \end{pmatrix},$$

es decir, el valor esperado de la desviación típica en la primera dimensión es 0.398 kilómetros, 0.133 sobre el segundo y la correlación entre ambos ejes es -0.75. Por tanto la forma de las matrices de varianza-covarianza en ambos modelos es similar; matrices con una amplia correlación negativa, aunque el volumen de las matrices en el modelo de mixturas es ligeramente inferior sobre el primer eje.

En la figura 6.14 hemos representado, para el modelo semiparamétrico, la intensidad suavizada de las componentes de su mixtura. Se aprecia que en esta ocasión la distribución de las medias de las componentes sí que se adapta bastante a la localización de los casos. Concretamente, se aprecian ciertas localizaciones con una probabilidad bastante alta de albergar una componente, en concreto las coordenadas (-0.8,-0.1), (0.6,0.2) y (1.2,-0.2).

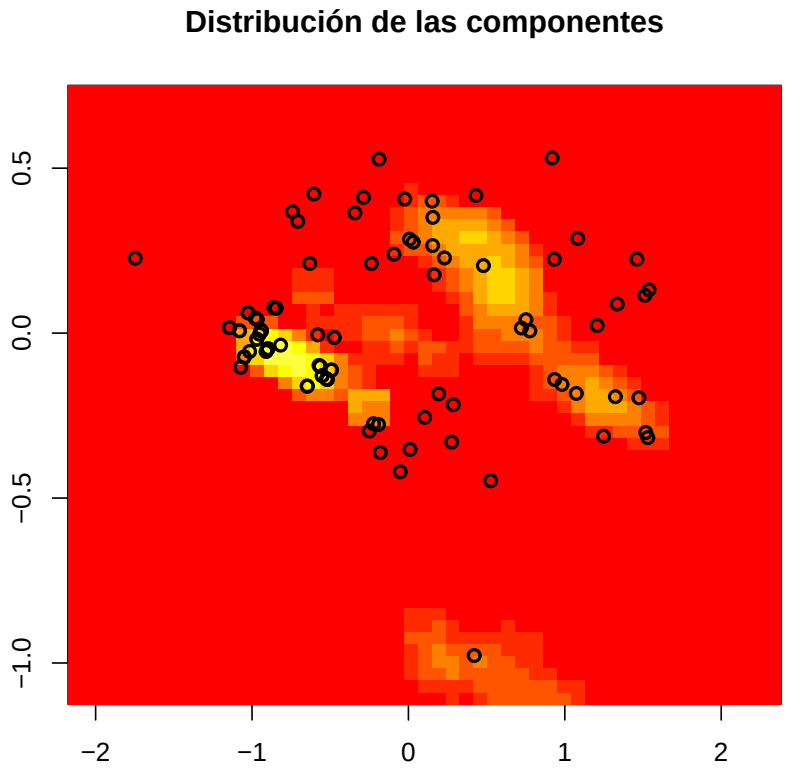

Figura 6.14: Análisis brote 3 mediante el modelo semiparamétrico. Distribución final de las medias de las componentes. El color de cada celda representa el número de simulaciones que hay en ella o en cualquiera de sus colindantes.

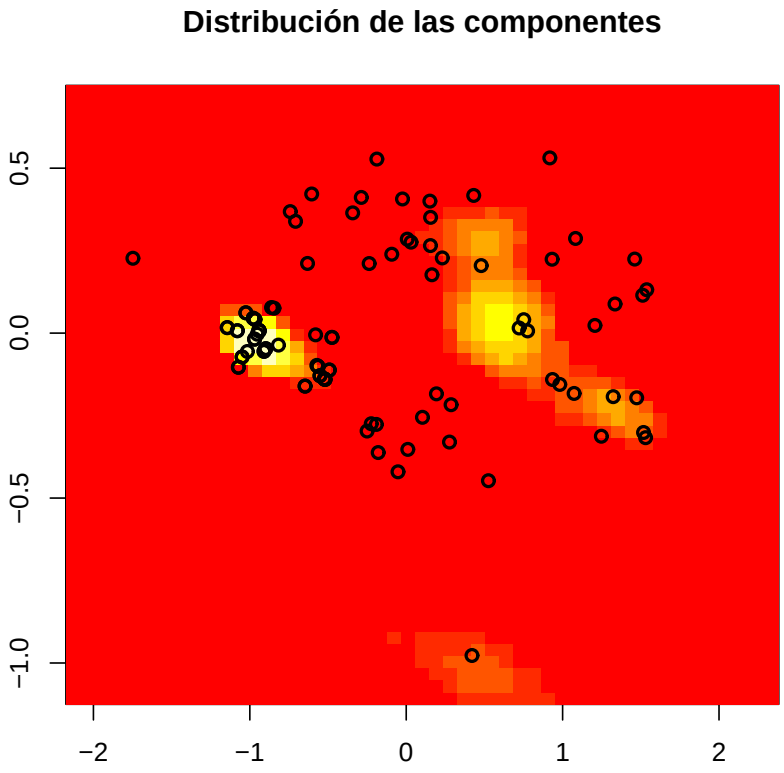

Figura 6.15: Análisis brote 3 mediante el modelo de mixturas. Distribución final de las medias de las componentes. El color de cada celda representa el número de simulaciones que hay en ella o en cualquiera de sus colindantes.

El modelo también parece contemplar la existencia de alguna componente alrededor de la coordenada (0.2,-1) para aumentar la verosimilitud del caso que reside en esa región. Además, la función de intensidad parece tomar un valor moderadamente elevado alrededor de la coordenada (0,0), cuando no se observa ningún caso alrededor de esta localización. Este resultado es consecuencia de que el modelo no descarta completamente el que haya una única instalación de riesgo que haya generado todos los casos, en cuyo caso la localización más probable para la instalación recaería alrededor del baricentro de los datos, la coordenada (0,0).

Para el modelo de mixturas hemos representado en la figura 6.15 la intensidad suavizada de sus componentes. En ella se aprecia que las zonas de riesgo señaladas por ambos modelos coinciden sustancialmente. La diferencia principal entre ambos resultados parece darse en la componente que el modelo semiparamétrico situaba alrededor de la coordenada (0.6,0.2), ya que en el modelo de mixturas, esta componente parece apuntar más claramente hacia la combinación de dos, la primera centrada alrededor de (0.5,0.3) y la segunda en (0.7,0.3). Además, el modelo de mixturas precisa en mayor medida la localización de las componentes, ya que nuestra propuesta parece determinar éstas de forma ligeramente más difusa. En los resultados del modelo de mixturas no se aprecia rastro de ninguna componente alrededor de la coordenada (0,0), ya que este modelo no considera probable la existencia de una única componente.

Respecto a los parámetros del proceso log-gaussiano que define la función de intensidad de los casos en el modelo semiparamétrico, en la figura 6.16 presentamos la distribución posterior de la desviación típica y la correlación espacial de dicho proceso. La media de la distribución posterior de la desviación típica es 1.67, y el intervalo de credibilidad al 95 % de este parámetro es [0.97,2.47]. La autocorrelación de orden 1 de las cadenas si-

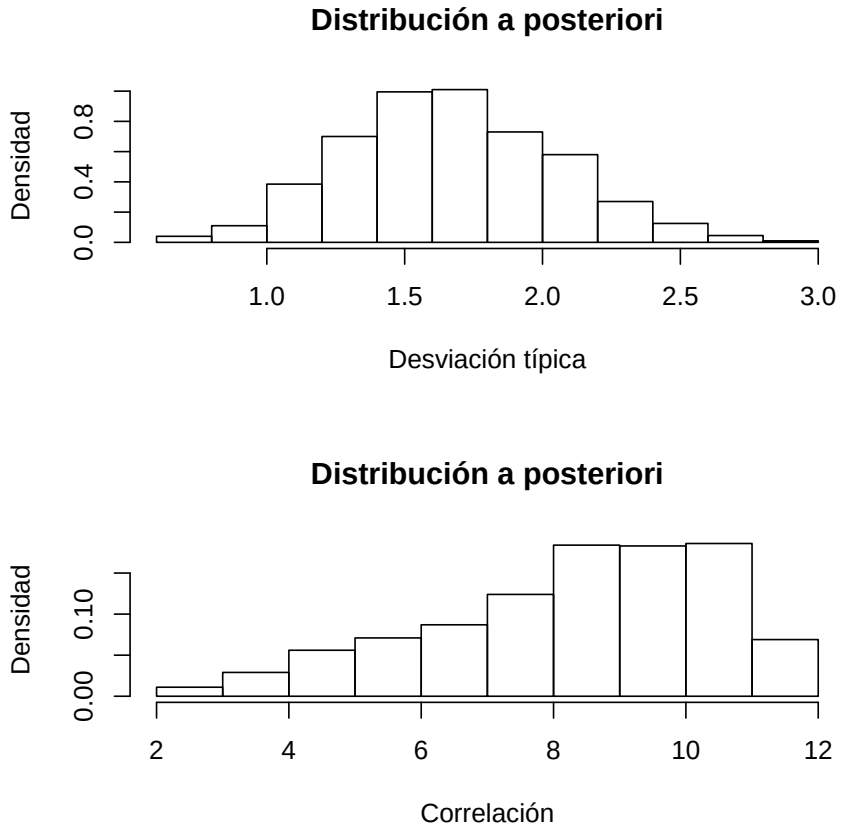

Figura 6.16: Análisis brote 3 mediante el modelo semiparamétrico. Inferencia sobre la desviación típica y la correlación del proceso log-gaussiano de los casos.

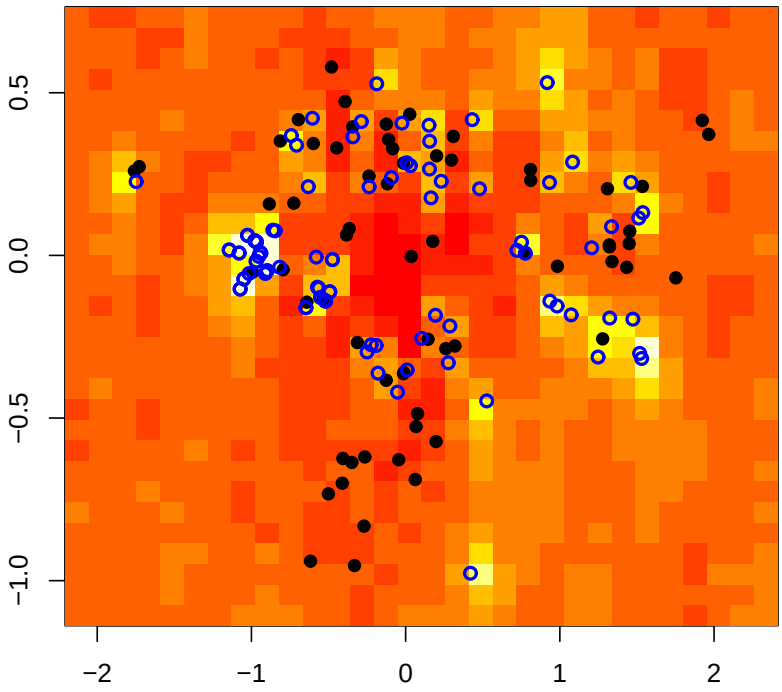

Figura 6.17: Análisis brote 3 mediante el modelo semiparamétrico. Media estimada en cada celda del proceso log-gaussiano de los casos. Los puntos negros representan la localización de los controles, los azules la localización de los casos

muladas para esta parámetro es de 0.73. Para el parámetro de correlación espacial, la media a posteriori es de 8.32, mientras que su intervalo de credibilidad al 95 % comprende el rango [3.59,11.27]. Para este parámetro la autocorrelación de orden 1 de las cadenas simuladas vale 0.44.

Por último, en la figura 6.17 hemos representado la estimación del proceso log-gaussiano en cada celda. Observamos que dicho proceso tiene nuevamente un comportamiento bastante heterogéneo, tal y como cabría esperar a la vista del parámetro de correlación espacial.

6.2.3. Análisis conjunto

Finalmente presentamos el análisis conjunto de todos los brotes que han tenido lugar en Alcoi desde septiembre de 1999 a noviembre de 2003. El número total de casos que se han incluido en este análisis es de 212 y contempla los casos también de los dos anteriores.

En la figura 6.18 se puede observar la distribución posterior del número de componentes de la mixtura para el modelo semiparamétrico y el de mixturas. Para nuestra propuesta la media de dicha distribución vale 4.59, y su intervalo de credibilidad al 95 % va desde 4 hasta 7. En el caso del modelo de mixturas, la media de la distribución posterior asciende a 7.48, mientras que su intervalo de credibilidad al 95 % varia de 4 a 13. Por tanto, los resultados proporcionados por ambos modelos son bastante diferentes en cuanto a este parámetro. Además, a diferencia del modelo semiparamétrico, el modelo de mixturas presenta mayor variabilidad en la estimación del número de componentes cuanto mayor es el tamaño muestral disponible en el estudio. Este comportamiento no parece observarse en nuestra propuesta donde el número de componentes presenta un comportamiento bastante

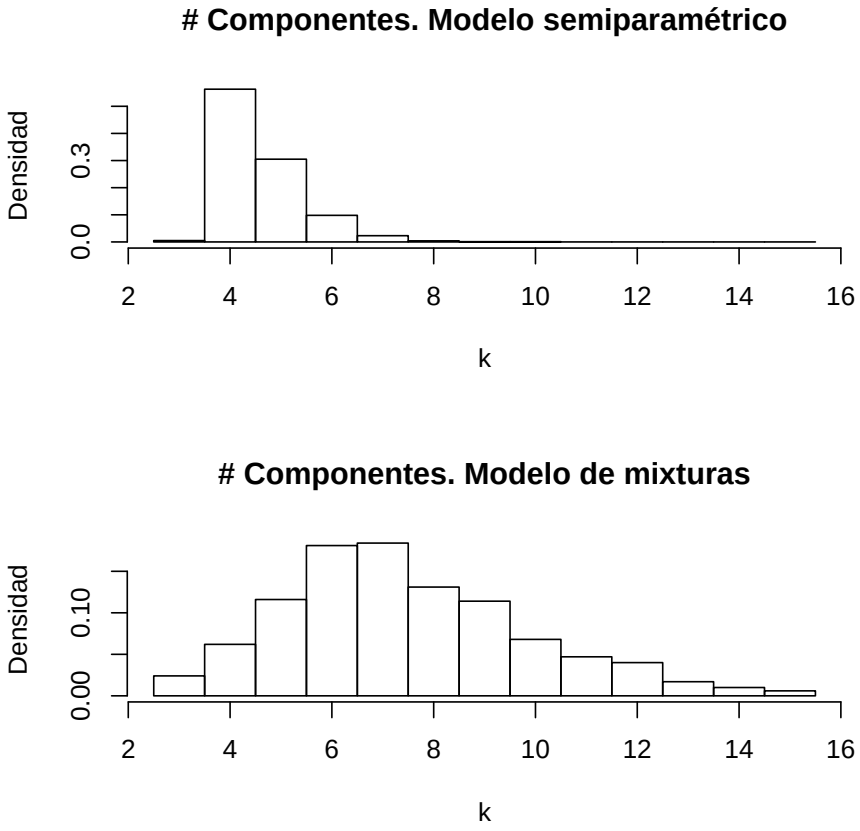

Figura 6.18: Análisis conjunto. Distribución posterior del número de componentes de la mixtura en el modelo semiparamétrico y en el de mixturas.

Modelo semiparamétrico							
Componentes	3	4	5	6	7	8	9
Frecuencia	0.005	0.563	0.305	0.098	0.023	0.004	0.001
Componentes	10	11	12	13	14	15	
Frecuencia	0.001	0	0	0	0	0	
Modelo de mixturas							
Componentes	3	4	5	6	7	8	9
Frecuencia	0.024	0.062	0.116	0.181	0.184	0.131	0.114
Componentes	10	11	12	13	14	15	
Frecuencia	0.068	0.047	0.040	0.017	0.010	0.006	

Cuadro 6.4: Análisis conjunto. Frecuencias de la distribución posterior del número de componentes del modelo de mixturas.

más estable.

En el cuadro 6.4 se presenta la distribución posterior del número de componentes para los dos modelos. La autocorrelación mostrada por las cadenas ha sido de 0.56 para el proceso semiparamétrico y de 0.80 para el modelo de mixturas.

Respecto a la inferencia sobre β, la media a posteriori de esta matriz de covarianzas en el proceso semiparamétrico ha resultado:

$$\begin{pmatrix} 0{,}402 & -0{,}081 \\ -0{,}081 & 0{,}048 \end{pmatrix},$$

por lo que la desviación típica sobre el primer eje es de 0.363 kilómetros, de 0.127 kilómetros sobre el segundo y existe una correlación de -0.58 entre

ambos ejes. Para el modelo de mixturas dicha media a posteriori resulta:

$$\begin{pmatrix} 0{,}074 & -0{,}013 \\ -0{,}013 & 0{,}020 \end{pmatrix},$$

así pues, la desviación típica en la primera dimensión es 0.155 kilómetros, 0.08 sobre el segundo y la correlación entre ambos ejes es -0.35. La forma de las matrices de varianza-covarianza en ambos modelos también es similar en esta ocasión; matrices con una amplia correlación negativa, aunque el volumen de las matrices en el modelo de mixturas es bastante inferior en el modelo de mixturas. En consecuencia, el modelo de mixturas proporcionará una superficie más apuntada y menos parsimoniosa que la mixtura del modelo semiparamétrico.

En la figura 6.19 se ha representado la intensidad suavizada de las componentes para nuestra propuesta. En ella se aprecian claramente los 4 clusters correspondientes a la moda de la distribución del número de componentes en este modelo. Los 3 clusters situados en la parte superior de la representación parecen estar bastante bien definidos, mientras que el cluster de la parte inferior parece tener una localización menos precisa que los otros. Este hecho se puede deber a que quizás esta componente responda a la existencia de casos en esta zona de la ciudad alejada del resto de focos y que podrían haberse contagiado en sus desplazamientos a otras partes de la ciudad. Como el modelo semiparamétrico sólo contempla la existencia de casos alrededor de los focos de riesgo, se ve obligado a introducir una nueva componente en esta zona, de esta forma da mayor verosimilitud a los casos que residen en ella. Este artefacto se podría solucionar con la presencia de una distribución uniforme además de las distribuciones normales en el modelo de mixturas.

La intensidad suavizada de las componentes para el modelo de mixturas se ha representado en la figura 6.20. Tal y como habíamos comentado, se

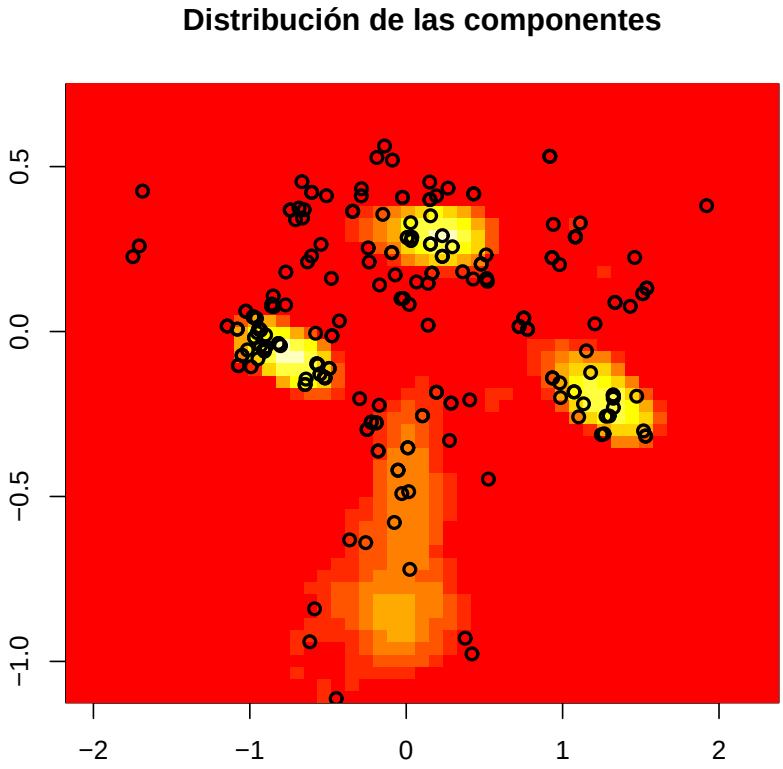

Figura 6.19: Análisis conjunto mediante el modelo semiparamétrico. Distribución final de las medias de las componentes. El color de cada celda representa el número de simulaciones que hay en ella o en cualquiera de sus colindantes.

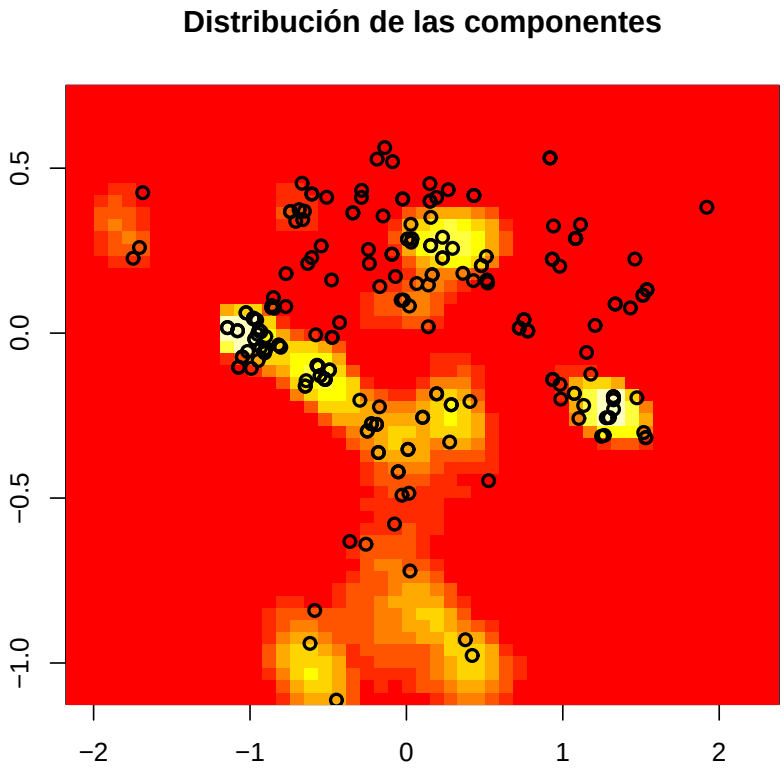

Figura 6.20: Análisis conjunto mediante el modelo de mixturas. Distribución final de las medias de las componentes. El color de cada celda representa el número de simulaciones que hay en ella o en cualquiera de sus colindantes.

aprecia un comportamiento bastante menos parsimonioso en la distribución de las componentes. En esta ocasión los 3 focos que observábamos en la parte superior del gráfico vuelven a aparecer, pero además se apuntan algunas localizaciones más como posibles fuentes de riesgo, aunque en general éstas tienen una intensidad menor que las señaladas en el modelo semiparamétrico. Respecto a la parte inferior del gráfico el modelo de mixturas apunta a la presencia de 2 clusters, ambos dos con un número de casos muy pequeño. Al igual que en nuestra propuesta, la inclusión de una distribución uniforme sobre la mixtura podría hacer innecesaria la presencia de estas componentes.

En la figura 6.21 hemos representado la distribución posterior de la desviación típica y la correlación espacial del proceso log-gaussiano del modelo semiparamétrico. La media a posteriori de la desviación típica es 1.63, y el intervalo de credibilidad al 95 % de este parámetro es [1.14,2.15]. La autocorrelación de orden 1 de las cadenas simuladas para esta parámetro es de 0.55. Para el parámetro de correlación espacial, la media de la distribución posterior es de 9.57, mientras que su intervalo de credibilidad al 95 % comprende el rango [6.08,11.27]. Para este parámetro la autocorrelación de orden 1 de las cadenas simuladas vale 0.31. Así, el parámetro de correlación espacial apunta en esta ocasión hacia una gran heterogeneidad espacial. Por su parte, la distribución posterior de la desviación típica, bastante alejada del valor 0, reafirma la necesidad de incluir este proceso en la modelización de la función de intensidad.

En la figura 6.22 hemos representado la estimación del proceso log-gaussiano en cada celda. Observamos que dicho proceso tiene un comportamiento bastante heterogéneo, tal y como apuntaba la distribución posterior del parámetro de correlación espacial.

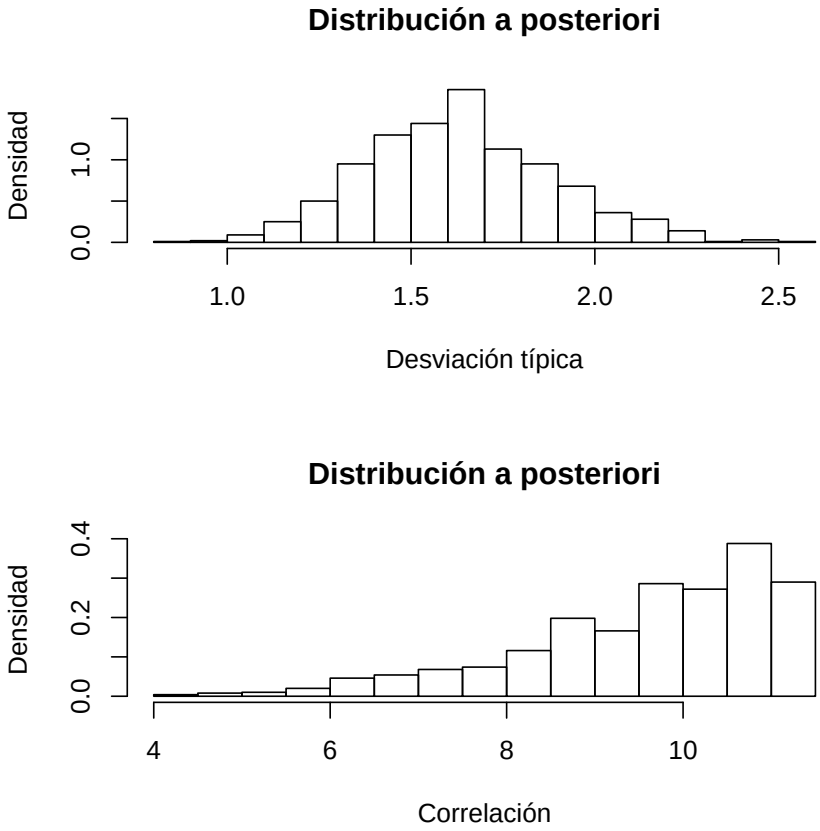

Figura 6.21: Análisis conjunto mediante el modelo semiparamétrico. Inferencia sobre la desviación típica y la correlación del proceso log-gaussiano de los casos.

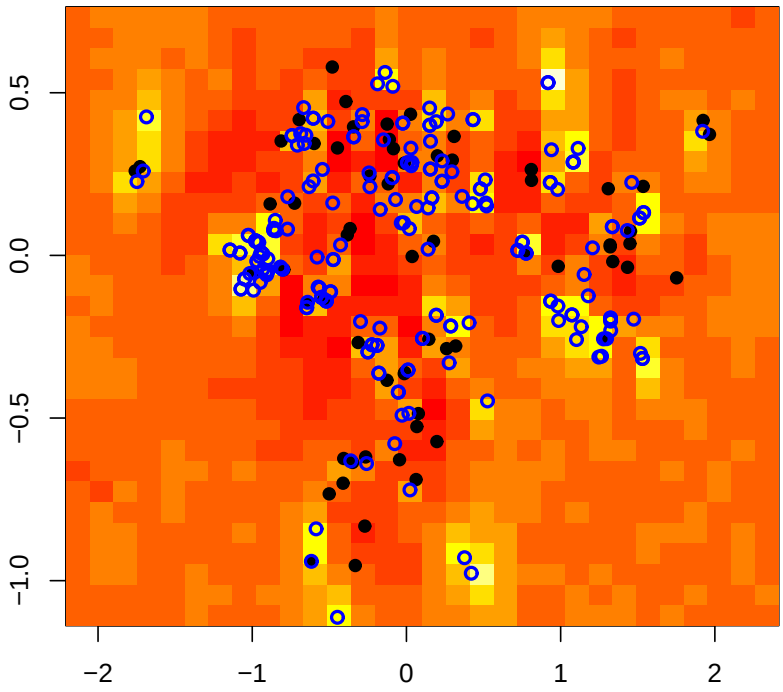

Figura 6.22: Análisis conjunto mediante el modelo semiparamétrico. Media estimada en cada celda del proceso log-gaussiano de los casos. Los puntos negros representan la localización de los controles, los azules la localización de los casos

Al observar conjuntamente los 3 análisis realizados, se pueden extraer varias conclusiones de interés. En primer lugar parece que, en todos los casos, los resultados del modelo semiparamétrico resultan bastante más parsimoniosos. Este hecho también se observaba en las pruebas numéricas del capítulo 5, donde este modelo además presentaba menor sesgo y precisión en la distribución posterior del número de componentes de la mixtura. Según podemos observar, el proceso log-gaussiano incluido para la descripción de los casos se ajusta en gran medida a la localización de éstos, haciendo menos necesaria la inclusión de una componente en la mixtura para contemplar la presencia de casos individuales más o menos aislados. Encontramos más adecuado contemplar la susceptibilidad de la población a contraer la enfermedad (que de hecho no suele ser un parámetro geográficamente constante), a la alternativa de añadir muchas componentes en la mixtura ante la presencia de cualquier agrupación mínima de casos.

Por otro lado, pensamos que el comportamiento de la distribución posterior del número de componentes en el modelo semiparamétrico es mucho más consistente. Este modelo ya apuntaba a 3 o 4 componentes en el estudio del tercer brote, y tras el estudio conjunto de todos los casos se ha reafirmado la presencia de 4 agregaciones de origen individual. Sin embargo, en el modelo de mixturas, el estudio del brote 3 apuntaba hacia 2 o 3 agregaciones, mientras que el estudio conjunto apunta a la existencia de entre 5 y 9 clusters. Resulta paradójico que al aumentar el tamaño muestral del patrón aumente tanto la variabilidad de la distribución final, este hecho se puede deber al comportamiento poco parsimonioso de este último modelo. Además, respecto a la distribución posterior del número de componentes del primer brote, podemos observar que cuando el tamaño muestral del patrón es bastante pequeño, ambos modelos se decantan por ajustar un número de componentes también pequeño. Este hecho es más acusado en el caso de nuestra propuesta donde la variabilidad del patrón se ajusta principal-

mente con la componente log-gaussiana. Así, creemos que la utilización de ambas modelizaciones se ha de realizar únicamente si el número de casos disponible es razonablemente grande. A la vista de los resultados del tercer brote sugeriríamos que al menos se dispusieran de 80 casos en el patrón si se quiere aplicar este tipo de modelización.

Desconocemos la importancia que puede estar teniendo el número de controles utilizados para realizar el estudio. Parece que la función de intensidad del proceso log-gaussiano de los controles proporciona un ajuste satisfactorio de ellos, ya que incluso se puede apreciar la orografía de la ciudad en la representación de la función de intensidad. Sin embargo, un aumento del tamaño muestral de los controles podría producir también una inferencia más precisa sobre los parámetros del modelo de los casos o incluso una representación menos abrupta de su heterogeneidad medioambiental.

Por último, nos gustaría reflejar la posibilidad de añadir información inicial en el modelo basada en la literatura o conocimiento previo del problema. La realización de estudios de meta-análisis sobre ciertos aspectos de interés para el brote nos puede proporcionar información que ayude a estimar las distribuciones posteriores de los parámetros del modelo. Por ejemplo, puede resultar muy interesante añadir información sobre la distancia a la que puede viajar la bacteria a partir de los resultados de la literatura o incluso información sobre la orografía de la ciudad que ayude a determinar la forma de las matrices de varianza-covarianza alrededor de cada foco. De todas formas querríamos dejar constancia de que los resultados obtenidos en el análisis del tercer brote según ambos modelos, y el análisis conjunto según el modelo semiparamétrico proporcionan unos resultados muy acordes con la literatura en cuanto a la estimación de la matriz de varianza-covarianza de cada componente. En dichos casos, la distancia máxima a la que cada foco representaría un riesgo sería próxima a una milla, tal y como se ha

comentado en el capítulo 1.

Capítulo 7

Conclusiones y futuras líneas de trabajo

Mediante este trabajo se ha pretendido realizar una modesta contribución dentro del campo de la detección de focos de riesgo en brotes epidémicos. Consideramos que el número de trabajos que se han realizado hasta la fecha en este campo es muy escaso, aunque creemos que el desarrollo de este tipo de aplicaciones se incrementará próximamente conforme se desarrollen nuevas técnicas de simulación que permitan plantearse nuevas modelizaciones. Si bien los tests estadísticos de agregación gozan de bastante más popularidad en el ámbito epidemiológico, en nuestra opinión, la modelización de procesos puntuales supone una alternativa mucho más flexible y potente.

El modelo que hemos planteado se adecúa al estudio de la distribución espacial en aquellos casos que existan dos fuentes de agregación que influyan en dicha distribución, la variabilidad de origen general y individual. Desde nuestro punto de vista, el proceso de agregación general habría de incluirse en cualquier modelización espacial de un patrón puntual procedente de

un proceso epidémico. Resulta muy difícil imaginar una enfermedad que se distribuya de forma completamente aleatoria en la población, que no distinga edades, sexos, clases sociales ni factores genéticos, por ejemplo. Como la distribución geográfica de estos factores en la población en general no es homogénea sino que suele variar de forma más o menos suave, consideramos necesario incluir un término aleatorio que tenga en cuenta esta variación geográfica del riesgo. Sin embargo, el proceso de agregación general no suele ser el objeto de principal interés desde un punto de vista epidemiológico, sino que resulta mucho más interesante la detección de las agregaciones de tipo individual. En consecuencia, cuando se tenga la intención de estudiar la existencia de agregaciones individuales en un patrón puntual se habrá de incluir específicamente dicho término en el modelo, aunque no se habrá de ignorar la variabilidad general para no incurrir en resultados sesgados.

En el modelo semiparamétrico que hemos descrito, a parte de controlarse el efecto de la agregación de tipo general, también se realiza de forma muy flexible la inclusión de la agregación individual. El modelo contemplado no impone la existencia de un número concreto de agregaciones. Puede parecer que nuestro modelo incorpora demasiada flexibilidad en su planteamiento y que resultará difícil distinguir entre las distintas fuentes de variabilidad existentes, cuando éstas tienen un comportamiento tan flexible. Consideramos que esta flexibilidad supone una gran ventaja ya que en caso de utilizarse modelos más restrictivos (por ejemplo con un número fijo de componentes) los resultados obtenidos estarían condicionados a dichas restricciones y serían válidos únicamente en el caso de que semejantes restricciones fuesen correctas. Así pues, como resulta muy difícil de determinar de antemano el número de focos de riesgo que intervienen en un brote de este tipo, consideramos que el modelo utilizado es la mejor opción posible, ya que sus resultados serán válidos independientemente del número de focos que hayan intervenido realmente en el brote.

La propuesta de modelo que hemos utilizado ha sido guiada por el problema concreto al que lo queríamos aplicar, los sucesivos brotes de legionelosis de Alcoi. Sin embargo este modelo es aplicable a muchas más situaciones. En concreto, cualquier enfermedad en la que no exista contagio directo entre casos y su incidencia se agregue en torno a ciertas localizaciones geográficas posiblemente desconocidas, es susceptible de ser estudiada con el modelo propuesto. Pero además la utilización de modelización estadística permite contemplar situaciones diferentes a la estudiada y adecuar el planteamiento estadístico a cada situación.

El modelo semiparamétrico utilizado, además de mejorar conceptualmente otras propuestas más simples que no incorporan agregaciones de tipo general, también parece obtener mejores resultados en la práctica. Dicha mejora ha quedado patente en las pruebas numéricas realizadas, ya que en ellas nuestra propuesta ha demostrado tener menos sesgo y ser más preciso a la hora de determinar el número de focos de riesgo que el modelo de mixturas. Sobre los datos reales de los brotes de Alcoi, el modelo semiparamétrico también ha demostrado tener un comportamiento más parsimonioso que el modelo de mixturas. Además, el modelo semiparamétrico parecía determinar con mayor precisión el número de focos de riesgo al aumentar el número de casos en el patrón, al contrario que el modelo de mixturas que mostraba mayor variabilidad al contemplar conjuntamente todos los brotes.

Por último, desde nuestro punto de vista resulta necesario un conocimiento más profundo de los modelos de mixturas con un número indeterminado de componentes. Tal y como se ha señalado en el capítulo 3, los resultados de los distintos análisis de estos modelos en la literatura difieren considerablemente entre sí. Además, en el desarrollo de nuestro trabajo hemos podido observar una gran sensibilidad de los resultados a distintos parámetros del modelo de mixturas. Consideramos que con vista a futuras

extensiones del modelo a nuevas aplicaciones se habría de hacer un análisis de sensibilidad más concienzudo que podría dar alguna clave sobre la discrepancia de resultados que hemos podido observar en la literatura.

Esperamos que el trabajo realizado sea de utilidad a la hora de plantear nuevas modelizaciones de procesos epidémicos en presencia de focos de riesgo. En ese caso nuestra propuesta puede servir de base para plantear nuevas modelizaciones. A modo de conclusión presentamos algunas de las líneas que pensamos podrían mejorar o complementar el trabajo realizado hasta este momento.

7.1. Líneas futuras de trabajo

Consideramos que la línea futura de trabajo más importante y necesaria consistiría en el análisis de sensibilidad de los resultados. Según hemos podido comprobar hay varios parámetros a los que el modelo resulta particularmente sensible. En concreto nos ha llamado la atención la sensibilidad respecto al parámetro δ de la distribución Dirichlet de los pesos de la mixtura. En general, al aumentar el valor de δ, disminuía considerablemente la variabilidad en la distribución posterior del número de componentes de la mixtura. Pensamos que este hecho se puede deber a la distribución marginal de los pesos como función del número de componentes del modelo. En la figura 7.1 se puede observar dichas distribuciones para $\delta = 1$ en la parte superior y para $\delta = 2$ en la parte inferior donde el número de componentes de la mixtura varía desde 2 hasta 15. En ambas representaciones las distribuciones más planas corresponden a los valores más bajos del número de componentes, mientras que las distribuciones más apuntadas corresponden a los valores mayores. Desde nuestro punto de vista la utilización del valor 1 para el parámetro δ tiene el inconveniente que la moda de las distribuciones

de los pesos se sitúa en 0, por tanto, este valor favorece la existencia de componentes vacías. Además, tal y como se puede observar cuanto mayor es el número de componentes de la mixtura más se favorecen las componentes de peso muy pequeño. Así, no resulta extraño el que la cola derecha de la distribución posterior del número de componentes sea tan alargada, ya que el modelo tenderá a aceptar componentes de bajo peso al tener éstas una gran verosimilitud a priori. Sin embargo, este hecho no sucede cuando tomamos valores superiores de δ, ya que en ese caso la moda de los pesos no se sitúa en 0, tal y como podemos observar en la parte inferior de la figura 7.1.

Así, a tenor de lo expuesto, nos parece razonable explorar la posibilidad de utilizar otros valores de δ distintos de 1. Incluso creemos razonable hacer depender δ del número de componentes de la mixtura. Así por ejemplo, para cada valor de k podríamos elegir δ de forma que la probabilidad de que un peso exceda o sea inferior a cierto umbral fuera constante. Hemos realizado algunas pruebas en este sentido y aunque los resultados obtenidos parecían bastante prometedores, hemos podido comprobar que seguía existiendo una gran sensibilidad de los resultados a los valores utilizados para δ. Aun así consideramos que se debería explorar este aspecto de la modelización con más calma ya que podría ayudar a comprender la variabilidad de la distribución posterior del número de componentes.

En cuanto a la sensibilidad del resto de parámetros del modelo, creemos que habría que explorar más detalladamente la utilización de expected posterior priors sobre la media y varianza de las componentes de la mixtura, tal y como proponen Pérez y Berger en sus distintos trabajos ([76], [77], [78]). En principio no parece demasiado satisfactorio el que la distribución posterior obtenida del número de parámetros dependa de la expected posterior prior utilizada, aun así parece que los resultados aplicando estas distribucio-

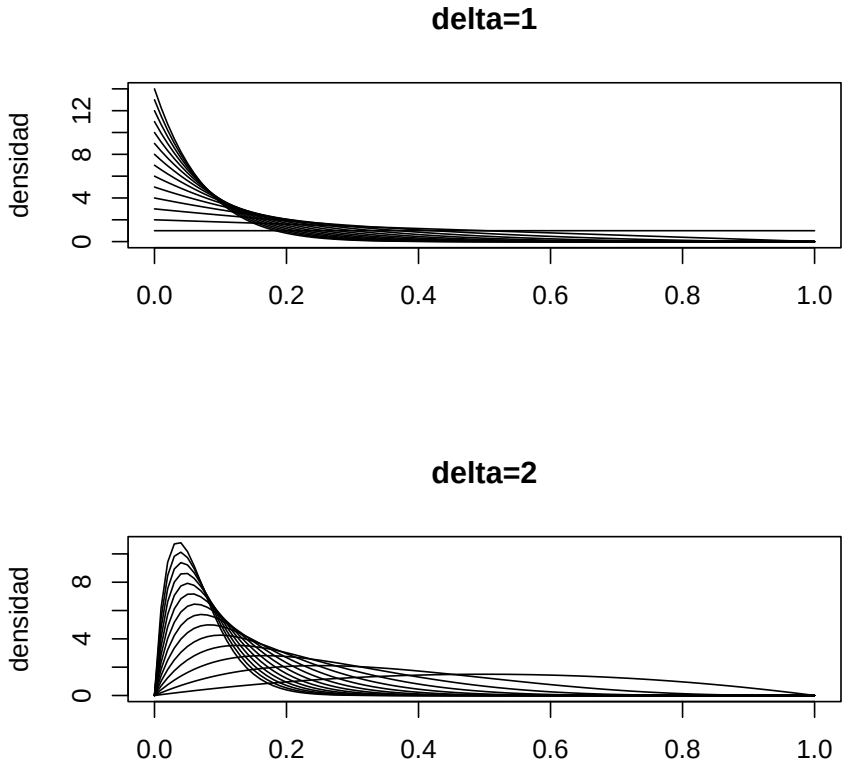

Figura 7.1: Distribución marginal de los pesos de las componentes de la mixtura para distintos valores de δ y k (el número de componentes de la mixtura). Las distribuciones más planas corresponden a los valores de k más bajos en ambas representaciones.

nes eran bastante prometedores. En concreto, en Pérez y Berger (1999) [76] se estudia el banco de datos *enzyme data*, también estudiado en Richardson y Green (1997) [80], obteniendo una distribución posterior del número de componentes con bastante menos variabilidad. Además, también se pueden observar en la literatura otros ejemplos de resultados bastante diferentes al cambiar la distribución inicial de la media de las componentes. Así vemos que en Dellaportas y Papageorgiou (2004) [35] analizan el *Old Faithful data*, también analizado por Stephens (1999) [91] con resultados bastante distintos en cuanto a la distribución posterior del número de componentes de la mixtura. Por tanto creemos que se debería realizar un estudio más pormenorizado de la sensibilidad de los resultados a las distribuciones iniciales de las medias y las varianzas de las componentes.

Otra mejora del modelo sería la inclusión de un término uniforme sobre el modelo de mixturas que contemple la existencia de casos que no se deban a un proceso de agregación sino a la presencia de la enfermedad en la población general. La inclusión de dicho término ya ha sido contemplada entre otros en Lawson y Clark (1999) [62] o Banfield y Raftery (1993) [9]. Este término ampliaría en gran medida el número de situaciones a la que puede aplicarse el modelo, ya que por ejemplo la propuesta realizada en esta tesis no es aplicable para valorar la agregación en la incidencia de cualquier tumor. En ese caso es previsible que existan un gran número de casos en la población que no respondan a un proceso de agregación espacial sino a la incidencia de la enfermedad en la población en general, por tanto se habrán de tener en cuenta dichos casos añadiendo un término uniforme a la mixtura. La inclusión de dicho término para el análisis de los brotes de legionelosis podría contrarestar el efecto de la movilidad urbana, ya que aquellos casos que hubieran contraído la enfermedad en sus desplazamientos se ubicarían en la componente uniforme y no necesitarían la inclusión de una nueva componente para ajustarlos.

También sería importante idear alguna forma de introducir covariables en la modelización. Este hecho no resulta demasiado sencillo ya que en general las covariables que dispongamos se referirán a los individuos y no a las localizaciones, en cuyo caso sería bastante más fácil de modelizar. En el caso de contemplar una componente uniforme sobre la mixtura, resulta particularmente interesante la inclusión de covariables sobre la probabilidad de pertenecer a esta componente o no. De esta forma se podría, por ejemplo, relacionar dicha probabilidad con información laboral, hábitos de vida,... y así las covariables podrían ayudar a identificar los casos atribuibles a clusters del resto. Esta información ayudaría a localizar los clusters individuales presentes en el patrón.

Otro tema en el que se podría profundizar es en la modelización espacio-temporal del problema. Hasta la fecha la única modelización de ese tipo de la que tenemos noticia en el ámbito epidemiológico es la desarrollada en Lawson y Clark (2002) [63]. Esta aproximación nos permitiría abordar el estudio de todos los brotes conjuntamente sin necesidad de agregarlos. Así, podríamos estudiar la información de todos los brotes por separado pero contemplando que los brotes próximos en el tiempo tienen mayor probabilidad de compartir focos de riesgo u otros factores.

Por último, como línea futura de trabajo se podría profundizar en la validación de los resultados obtenidos. Dicha validación podría realizarse mediante el uso de las técnicas habituales, utilizando los estadísticos K, F, G o J descritos en el capítulo 3. Además, recientemente se ha propuesto en Baddeley et al. (2005) [7] el uso de residuos para la valoración del ajuste de modelos en procesos puntuales. No obstante, nos encontramos con el problema de que la gran mayoría de las técnicas que conocemos de validación en procesos puntuales han sido ideadas dentro del marco de la estadística clásica, mientras que el planteamiento de nuestra modelización se ha hecho

desde una perspectiva bayesiana. Así pues, la mayoría de las técnicas existentes se basan en una estimación puntual de los parámetros del modelo, mientras que en nuestro caso disponemos de la distribución posterior de éstos. Es por ello que se requeriría de una adaptación de las técnicas existentes al marco bayesiano o, como alternativa, la aplicación de las técnicas de validación de modelos bayesiana al campo de los procesos puntuales, que hasta donde nosotros sabemos está poco extendida.

Bibliografía

[1] Abellán, J. J., Martínez-Beneito, M. A., Zurriaga, O., Jorques, G., Ferrándiz, J., y López-Quílez, A. (2002). Procesos puntuales como herramienta para el análisis de posibles fuentes de contaminación. *Gaceta Sanitaria*, **16**(5):445–449.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B.Ñ. y Csaki, F., editors, *Proc. 2nd International Symposium on Information Theory*, pages 267–281.

[3] Alexander, F. E. y Boyle, P., editors (1996). *Methods for Investigating localized clustering of disease*. Number 135 in IARC Scientific Publications. International Agency for Research on Cancer, Lyon.

[4] Alexander, F. E. y Boyle, P. (2000). Do cancers cluster? In *Spatial Epidemiology: Methods and applications*, chapter 17, pages 302–316. Oxford University Press.

[5] Arjas, E. y Gasbarra, D. (1994). Nonparametric bayesian inference from right censored survival data, using the Gibbs sampler. *Statist. Sinica*, **4**:505–524.

[6] Arjas, E. y Heikkinen, J. (1977). An algorithm for nonparametric bayesian estimation of a Poisson intensity. *Comput. Statist.*, **12**:384–402.

[7] Baddeley, A., Turner, R., Moller, J., y Hazelton, M. (2005). Residual analysis for spatial point processes. *Journal of the Royal Statistical Society: Series B*, **67**(5):1–35.

[8] Bailey, T. C. (2001). Spatial statistics analysis in health. *Cadernos de Saude Publica*, **17**(5):1083–1098.

[9] Banfield, J. D. y Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, **49**:803–821.

[10] Benes, V., Bodlak, K., Möller, J., y Waagepetersen, R. (2003). Application of log-Gaussian Cox Processes in disease mapping. In Mateu, J., Holland, D., y Gonzalez-Mantiega, W., editors, *Proceedings of the ISI Conference on Environmental Statistics and Health, Santiago de Compostela, 2003*, pages 95–105. International Statistical Institute.

[11] Berger, J., De Oliveira, V., y Sansó, B. (2001). Objective bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **96**:1361–1374.

[12] Bernardo, J. M. y Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, New York.

[13] Besag, J. (1994). Discussion to "representation of knowledge in complex systems" by grenander and miller. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **56**:591–592.

[14] Besag, J. y Newell, J. (1991). The detection of clusters in rare diseases. *Journal of the Royal Statistical Society: Series A*, **154**:143–155.

[15] Besag, J., York, J., y Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**:1–21.

[16] Best, N. G., Ickstadt, K., Wolpert, R. L., y Briggs, D. J. (2000). Spatial Poisson regression for health and exposure data measured at disparate resolutions. *Journal of the American Statistical Association*, **95**:1076–1088.

[17] Bhopal, R. S., Fallon, R. J., Buist, E. C., Black, R. J., y Urquhart, J. D. (1991). Proximity of the home to a cooling tower and risk of non-outbreak Legionnaire's Disease. *British Medical Journal*, **302**(6773):378–383.

[18] Birch, J. M., Alexander, F. E., Blair, V., Eden, O. B., Taylor, G. M., y McNally, R. J. Q. (2000). Space-time clustering patterns in childhood leukaemia support a role for infection. *British Journal of Cancer*, **82**:1571–1576.

[19] Black, R. J., Sharp, L., y Urquhart, J. D. (1991). An analysis of the geographical distribution of childhood Leukemia and Non-Hodgkin Lymphomas in Great Britain using areas of approximately equal population size. In Draper, G. J., editor, *The geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin Lymphoma in Great Britain 1966-1983*, pages 61–68. HMSO, London.

[20] Brix, A. (1999). Generalized Gamma Measures and Shot-Noise Cox Processes. *Advances in Applied Probability*, **31**:929–953.

[21] Brooks, S. P. y Gelman, A. (1998). Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**:434–455.

[22] Brooks, S. P. y Giudici, P. (2000). MCMC convergence assessment via two-way ANOVA. *Journal of Computational and Graphical Statistics*, **9**:266–285.

[23] Brown, C. M., Nuorty, P. J., Breiman, R. F., Hathcock, A. L., Fields, B. S., Lipman, H. B., Llewellyn, G. C., Hoffman, J., y Cetron, M. (1999). A community outbreak of Legionnaire's Disease linked to hospital cooling towers: an epidemiological method to calculate dose of exposure. *International Journal of Epidemiology*, **28**:353–359.

[24] Cappé, O., Robert, C. P., y Rydén, T. (2003). Reversible jump MCMC converging to birth-and-death MCMC and more general continuous time samplers. *Journal of the Royal Statistical Society: Series B*, **65**(3):679–700.

[25] Carlin, B. P. y Chib, S. (1995). Bayesian model choice via Markov Chain Monte Carlo. *Journal of the Royal Statistical Society: Series B*, **57**:473–484.

[26] Castelloe, J. M. (1998). *Issues in reversible jump Markov Chain Monte Carlo and composite EM analysis, applied to spatial poisson cluster processes.* PhD thesis, University of Iowa.

[27] Celeux, G., Forbes, F., Robert, C. P., y Titterington, D. M. (2003). Deviance information criteria for missing data models. Technical Report 30, Centre de Recherche en Economie et Statistique.

[28] Chardot, C., Carton, M., Spire-Bendelac, M., Le Pommelet, C., Golmard, J. L., y Auvert, B. (1999). Epidemiology of biliary atresia in france: a national study 1986-96. *Journal of Hepatology*, **1999**(6):1006–1013.

[29] Christensen, O. F., Möller, J., y Waagepetersen, R. P. (2001). Geometric ergodicity of metropolis-hastings algorithms for conditional simulation in generalized linear mixed models. *Methodology and Computing in Applied Probability*, **3**:309–327.

[30] Cox, D. R. (1955). Some statistical methods related with series of events (with discussion). *Journal of the Royal Statistical Society: Series B*, **17**:129–164.

[31] Cressie, N. A. (1993). *Statistics for spatial data. Revised edition.* John Wiley & Sons.

[32] Cuzick, J. y Edwards, R. (1990). Spatial clustering for inhomogeneous populations (with discussion). *JRSSB*, **52**:73–104.

[33] Dasgupta, A. y Raftery, A. E. (1998). Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association*, **93**:294–302.

[34] Dellaportas, P., Forster, J. J., y Ntzoufras, I. (2002). On bayesian model and variable selection using MCMC. *Statistics and Computing*, **12**:27–36.

[35] Dellaportas, P. y Papageorgiou, I. (2004). Multivariate mixtures of normals with unknown number of components. http://stat-athens.aueb.gr/ ptd/finmix.pdf.

[36] Dempster, A. P., Laird, N. M., y Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, **39**:1–38.

[37] Diggle, P. (1985). A kernel method for smoothing point process data. *Journal of the Royal Statistical Society: Series A*, **34**:138–147.

[38] Diggle, P. y Rowlingson, B. (1994). A conditional approach to point process modelling of elevated risk. *Journal of the Royal Statistical Society: Series A*, **157**:433–440.

[39] Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns*. Arnold, London, 2 edition.

[40] Diggle, P. J. y Chetwynd, A. G. (1991). Second-order analysis of spatial clustering for inhomogeneous populations. *Biometrics*, **47**:1155–1163.

[41] Diggle, P. J., Tawn, J. A., y Moyeed, R. A. (1998). Model based geostatistics (with discussion). *Journal of the Royal Statistical Society: Series A*, **43**(3):299–350.

[42] Escobar, M. y West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, **90**:577–588.

[43] Ferreira, J. T. A. S., Denison, D. G. T., y Holmes, C. C. (2002). Partition modelling. In Lawson, A. B. y Denison, D. G. T., editors, *Spatial Cluster Modelling*, pages 125–145. CRC.

[44] Fisher, R. A., Thornton, H. G., y Mackenzie, W. A. (1922). The accuracy of the plating method of estimating the density of bacterial populations. *Annals of Applied Biology*, **9**:325–359.

[45] Fraley, C. y Raftery, A. E. (1998). How many clusters?. Which clustering method?. Answers via model-based cluster analysis. *Computer Journal*, **41**:578–588.

[46] Fraley, C. y Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, **97**:611–631.

[47] Gelman, A. (2004). Prior distributions for variance parameters in hierarchical models. http://www.stat.columbia.edu/ gelman/research/unpublished/tau7.pdf.

[48] Gelman, A., Carlin, J. B., Stern, H. S., y Rubin, D. B. (2004). *Bayesian data analysis*. CRC, Boca Raton, 2 edition.

[49] Geman, S. y Geman, D. (1984). Stochastic relaxation, gibbs distribution and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**:721–741.

[50] Gilks, W. R., Richardson, S., y Spiegelhalter, D. J. (1995). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

[51] Godsill, S. J. (2000). On the relationship between MCMC model uncertainty methods. *Journal of Computational and Graphical Statistics*, **10**:230–248.

[52] Green, P. J. (1995). Reversible Jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika*, **82**(4):711–732.

[53] Green, P. (2003). Trans-dimensional Markov Chain Monte Carlo. In Green, P. J., Hjort, N. L., y Richardson, S., editors, *Highly structured stochastic systems*, chapter 8, pages 179–196. Oxford University Press.

[54] Grenander, U. y Miller, M. I. (1994). Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B*, **56**:549–603.

[55] Heikkinen, J. y Arjas, E. (1998). Non-parametric bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, **25**:435–450.

[56] Ickstadt, K. y Wolpert, R. L. (1999). Spatial regresion for Marked Point Processes. In Bernardo, J. M., Berger, J. O., Dawid, A. P., y Smith, A. F. M., editors, *Bayesian Statistics 6*, pages 323–341. Oxford University Press.

[57] Ishwaran, H. y Zarepour, M. (2000). Markov Chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika*, **87**:371–390.

[58] Kelsall, J. E. y Diggle, P. J. (1998). Spatial variation in risk of disease: A nonparametric binary regression approach. *Journal of the Royal Statistical Society: Series A*, **47**:559–573.

[59] Knorr-Held, L. y Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, **56**(13-21).

[60] Last, J., editor (2001). *A dictionary of epidemiology*. Oxford University Press, 4 edition.

[61] Lawson, A. B. (2001). *Statistical Methods in Spatial Epidemiology*. John Wiley & Sons, 1 edition.

[62] Lawson, A. B. y Clark, A. (1999). Markov Chain Monte Carlo methods for putative sources of hazard and general clustering. In Lawson, A., Biggeri, A., Bohning, D., Lesaffre, E., Viel, J. F., y Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, pages 119–142. Wiley.

[63] Lawson, A. B. y Clark, A. (2002). Spatio-temporal cluster modelling of small area health data. In *Spatial cluster modelling*, chapter 14, pages 235–256. John Wiley & Sons.

[64] Lawson, A. B. y Denison, D. G. T., editors (2002). *Spatial Cluster Modelling*. CRC.

[65] Lawson, A. B. y Denison, D. G. T. (2002). Spatial cluster modelling: An overview. In Lawson, A. B. y Denison, D. G. T., editors, *Spatial Cluster Modelling*, pages 1–19. CRC.

[66] Martínez-Beneito, M. A., Abellán, J., López-Quílez, A., Vanaclocha, H., Zurriaga, O., Jorques, G., y Fenollar, J. (2005). Source detection in an outbreak of Legionnaire's Disease. In Baddeley, A., Gregori, P., Mateu, J., Stoica, R., y Stoyan, D., editors, *Case studies in Spatial Point Process Models*, pages 169–181. Springer-Verlag.

[67] Ministerio de Sanidad y Consumo (1998). *Recomendaciones para la prevención y control de la legionelosis*.

[68] Möller, J. (2003). A comparison of spatial point process models in epidemiological applications. In Green, P. J., Hjort, N. L., y Richardson, S., editors, *Highly Structured Stochastic Systems*, chapter 8B, pages 264–270. Oxford University Press.

[69] Möller, J., Syversveen, A. R., y Waagepetersen (1998). Log-Gaussian Cox Process. *Scandinavian Journal of Statistics*, **25**:451–482.

[70] Möller, J. y Waagepetersen, R. (2004). *Statistical inference and simulation for spatial point processes*. CRC.

[71] Moreno, E. y Liseo, B. (2003). A default bayesian test for the number of components of a mixture. *Journal of Statistical Planning and Inference*, **111**:129–142.

[72] Mwalili, Z. M., Lesaffre, E., y Declerck, D. (2005). A bayesian ordinal logistic regression model to correct for interobserver measurement error in a geographical oral health study. *Applied Statistics*, **54**(1):77–93.

[73] Neyman, J. y Scott, E. L. (1958). Statistical approach to problems of cosmology (with discussion). *Journal of the Royal Statistical Society: Series B*, **20**:1–43.

[74] Nobile, A. (2004). On the posterior distribution of the number of components in a finite mixture. *Annals of Statistics*, **32**(5):2044–2073.

[75] Openshaw, S. y Craft, A. (1991). Using the geographical analysis machine to search for evidence of clusters and clustering in childhood Leukaemia and Non-Hodgkin Lymphomas in Britain. In Draper, G. J., editor, *The geographical Epidemiology of Childhood Leukaemia and Non-Hodgkin Lymphoma in Great Britain 1966-1983*, pages 109–122. HMSO, London.

[76] Pérez, J. M. y Berger, J. (1999). Default analysis of mixture models using Expected Posterior Priors. Technical Report 99-12, Centro de estadística y software matemático.

[77] Pérez, J. M. y Berger, J. (2001). Analysis of mixture models using Expected Posterior Priors, with application to classification of gamma ray bursts. In George, E. y Nanopoulos, P., editors, *Bayesian Methods, with applications to science, policy and official statistics*, pages 401–410. Official Publications of the European Communities, Luxembourg.

[78] Pérez, J. M. y Berger, J. (2002). Expected Posterior Prior distributions for model selection. *Biometrika*, **89**:491–512.

[79] Raftery, A. (1993). Bayesian model selection in structural equation models. In Bollen, K. A. y Long, J. S., editors, *Testing Stuctural Equation Models*, pages 163–180. Sage.

[80] Richardson, S. y Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society: Series B*, **59**(4):731–792.

[81] Ripley, B. D. (1976). The second-order analysis of stationary of stationary point processes. *JAP*, **13**:255–266.

[82] Ripley, B. D. (1977). Modelling spatial patterns (with discussion). *Journal of the Royal Statistical Society: Series B*, **39**:172–212.

[83] Roberts, G. O. y Rosenthal, J. S. (1998). Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society: Series B*, **60**:255–268.

[84] Roberts, G. O. y Tweedie, R. L. (1996). Exponential convergence of Langevin diffusions and their discrete approximations. *Bernoulli*, **2**:341–363.

[85] Roeder, K. y Wasserman, L. (1997). Practical bayesian estimation using mixtures of normals. *Journal of the American Statistical Association*, **92**:894–902.

[86] Schlather, M. (1999). Introduction to positive definite functions and to unconditional simulation of random fields. Technical Report ST-99-10, Department of Mathematics and Statitstics, Lancaster University.

[87] Schwarz, C. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**:461–464.

[88] Sisson, S. A. (2004). Trans-dimensional Markov-chains: A decade of progress and future perspectives. Technical report, Department of Statistics, University of New South Wales.

[89] Spiegelhalter, D. J., Abrams, K. R., y Miles, J. P. (2004). *Bayesian approaches to clinical trials and Health-Care evalutaion*. John Wiley & Sons, Chichester.

[90] Spiegelhalter, D. J., Best, N. G., Carlin, B. P., y VanDerLinde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society: Series B*, **64**:583–641.

[91] Stephens, M. (1999). *Bayesian analysis of mixtures*. PhD thesis, Oxford.

[92] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, **62**(4):795–809.

[93] Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B*, **62**(4):795–809.

[94] Stephenson, J., Gallagher, K., y Holmes, C. C. (2003). Beyond kriging - dealing with discontinuous spatial data fields using adaptative prior information and bayesian partition modelling.

[95] Stoyan, D., Kendall, W., y Mecke, J. (1996). *Stochastic Geometry and its applications*. John Wiley & Sons, 2 edition.

[96] Thomas, A., Best, N., Lunn, D., Arnold, R., y Spiegelhalter, D. (2004). *GeoBUGS user manual*, 1.2 edition.

[97] Waagepetersen, R. y Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward QTL-mapping. *International Statistical Review*, **69**:49–61.

[98] Wakefield, J. C., Kelsall, J. E., y Morris, S. E. (2000). Clustering, cluster detection, and spatial variation in risk. In Elliott, P., Wakefield, J. C., Best, N. G., y Briggs, D. J., editors, *Spatial Epidemiology*, pages 128–152. Oxford University Press.

[99] Wartenberg, D. (2001). Investigating disease clusters: Why, When and How? *Journal of the Royal Statistical Society: Series A*, **164**(1):13–22.

[100] Wolpert, R. L. y Ickstadt, K. (1998). Poisson/Gamma random field models for spatial statistics. *Biometrika*, **85**:251–267.