

UNIVERSITAT DE VALÈNCIA  
Facultat de Matemàtiques  
Departament d'Estadística i Investigació Operativa



**MODELO DE CONGLOMERADOS PARA EL ANÁLISIS  
BAYESIANO DE DATOS EPIDEMIOLÓGICOS EN ÁREAS  
PEQUEÑAS**

Tesis doctoral  
Realizada por: Karen Cecilia Flórez Lozano  
Dirigida por: José D. Bermúdez Edo y  
Ana Corberán Vallet



D. José Domingo Bermúdez Edo, profesor titular y Dña. Ana Corberán Vallet, ayudante Doctor del Departamento de Estadística e Investigación Operativa de la Universidad de Valencia

CERTIFICAN que la presente memoria de investigación

**“MODELO DE CONGLOMERADOS PARA EL ANÁLISIS  
BAYESIANO DE DATOS  
EPIDEMIOLÓGICOS EN ÁREAS PEQUEÑAS”**

ha sido realizada bajo su dirección por Karen Cecilia Flórez Lozano, y constituye su tesis para optar al grado de Doctor en Estadística y Optimización.

Y para que así conste, en cumplimiento con la normativa vigente, autorizan su presentación ante la Facultad de Matemáticas de la Universidad de Valencia para que pueda ser tramitada su lectura y defensa pública.

Valencia, 13 de Abril de 2016.

José D. Bermúdez Edo.

Ana Corberán Vallet.



**A mi esposo e hijos**



# Agradecimientos

*Estando en tierras lejanas mi corazón siempre alegre y agradecido estará, por haber encontrado bellas personas que me acogieron con amor y sinceridad.*

Doy infinitas gracias a Dios por darme la vida y permitirme vivir esta maravillosa experiencia de formación. A la Virgen María Auxiliadora por guiarme y ser mi compañera en los momentos más difíciles.

Y será poco dar un gúgolduplex de gracias a mis queridos tutores José D. Bermúdez E. y Ana Corberán Vallet, porque con mucha paciencia y dedicación me han acompañado y orientado para hacer realidad este proyecto. Al profe José gracias por enseñarme entre muchas otras cosas de tipo académico, a no enquistar el dolor, a ser fuerte y tener cada día un motivo para estar feliz y mostrar mi mejor sonrisa. Ana, gracias por confiar en mí y por siempre creer que sí lo lograría, gracias por tus consejos y por ofrecerme tu amistad. Gracias a los dos por hacer de cada reunión un encuentro ameno y lograr que de la mayoría de ellos saliera dando brincos por el pasillo.

A todos los profesores, profesoras, secretarias y personal de apoyo del departamento de Estadística e Investigación Operativa, por su acogida y por ser tan especiales conmigo. Cada palabra, cada saludo y cada gesto de cariño hicieron más fácil este trabajo fuera de mi país, gracias por hacerme sentir parte de este maravillosos equipo.

A Mariana, Mónica y Miguel, gracias por ser nuestra familia en España, gracias por acogernos con tanto cariño y ayudarnos desde nuestra llegada. A Silvia, Adina, Francesca, amigas del corazón, las quiero mucho y les agradezco por todo el cariño y por ser tan especiales conmigo y con mi familia. Estoy muy agradecida con mis compis colombianos y mis compañeros y compañeras de camino, a los del despacho y a los que están fuera de él. Me llevo en el corazón gratos recuerdos y muchos momentos compartidos. Los quiero mucho y gracias por los aportes recibidos.

Quiero agradecer a la Universidad del Norte, por todo el apoyo recibido y por confiar en mi futura labor. Gracias de todo corazón a los profesores Joachim Hahn, Guillermo Cervantes, Rafael Tuesca, Humberto Llinás y Jairo Hernández.

A toda mi familia, y muy especialmente a mis padres, Rosa y Julio, por creer en mí, por apoyarme desde la distancia y animarme con palabras de cariño. A mis hermanos, Katy y Fabio por su apoyo incondicional.

Gracias a mis suegros Cira y Pichocho por estar pendiente de nosotros y animarnos a seguir adelante. A mis cuñados Evelyn, Luis Guillermo y Hugo con sus respectivas familias, gracias muchas gracias por todos los detalles de cariño, por cuidarnos y mimarnos desde la distancia.

Y aunque ya se lo he dicho muchas veces aquí lo quiero plasmar, gracias a mi esposo amado Alfon, que con su ayuda y apoyo siempre estuvo allí para animar. Gracias por estar a mi lado, gracias por creer, gracias por superar juntos los momentos de agobio y gracias por formar nuestra familia. Hoy no solo dedico a ti esta tesis, también están presentes nuestros angelitos que desde el cielo nos acompañan Isa y Alfonsito, y aquellas dos chicuelas que nos animan a seguir adelante Kamila y la princesita Valentina que llegó a nuestras vidas aliviando el dolor y motivando nuestros sueños. A Kamy por compartir la felicidad de nuestros triunfos. Y a mi Valentina, nuestra Valentina, porque desde la barriguita me ayudaste con tu buen comportamiento y me permitiste terminar el capítulo que hacía falta. ¡A todos los amo!

# Prólogo

El análisis de datos relacionados con la incidencia y/o mortalidad de una enfermedad ha ido en constante crecimiento por parte de investigadores, además dicho análisis es una herramienta básica en el estudio de datos en salud pública. La aparición de una enfermedad en una zona geográfica, su localización y quizá otros factores de tipo medio-ambiental, por ejemplo, proporcionan herramientas para el estudio y la obtención de una visión general de la distribución espacial del riesgo. También es posible obtener información acerca de la etiología de la enfermedad y su evolución en el tiempo.

Es habitual estudiar la distribución geográfica de una enfermedad en áreas pequeñas, de tal manera que perdamos la menor información posible. Sin embargo, existen problemas de estimación del riesgo en este tipo de áreas y es entonces cuando se utilizan técnicas más específicas que están incluidas en el mapeo de enfermedades. A través del mapeo de enfermedades se pretende estudiar la variación geográfica y/o temporal del comportamiento de una o más enfermedades en una zona de estudio. Entre los objetivos que se abordan en esta línea de investigación encontramos: la formulación de hipótesis, estimación de la variabilidad de riesgos subyacentes en un área de estudio, identificación de zonas con alto riesgo que representan una amenaza para la salud de las poblaciones, entre otros.

El objetivo principal de esta tesis es el estudio de la variación geográfica del riesgo de una cierta enfermedad en una determinada zona de estudio. Dentro de esta temática se han propuesto trabajos que si bien logran estimar el riesgo relativo en áreas pequeñas, en ocasiones tienden a tener un exceso de suavizado, haciendo compleja la tarea de identificar zonas de alto riesgo. Nuestra propuesta intenta abordar los dos objetivos de mapeo de enfermedades simultáneamente, por un lado estimar el riesgo relativo en áreas pequeñas y al mismo tiempo tiempo detectar discontinuidades entre las mismas. Con el modelo de conglomerados que presentamos, hemos pretendido realizar una aportación al análisis bayesiano de datos epidemiológicos en áreas pequeñas. Dicho modelo permite obtener estimaciones del riesgo en cada una de las áreas que conforman la zona de estudio, como también estudiar el número de clases o conglomerados que pueden existir en una zona geográfica.

Existen factores de riesgos que en ocasiones no podemos observar y que involucran a los vecinos de las diferentes áreas que conforman la zona de estudio, logrando así que áreas próximas tengan riesgo similar. Esta es la razón por la cual una gran cantidad de modelos de suavización modelizan la correlación espacial entre observaciones bajo el supuesto de que áreas próximas se parecerán más que las que se encuentran más distantes. La novedad de nuestra propuesta es que rompemos con el esquema habitual de tener en cuenta algún criterio de vecindad para la estimación del riesgo. El modelo propuesto incluye ideas de modelos de

mixturas, de conglomerados y de modelos ocultos.

Todo lo mencionado hasta aquí lo exponemos detalladamente en los distintos capítulos de la tesis. En el primer capítulo hacemos una revisión de la literatura en mapeo de enfermedades y describimos las características, similitudes y discrepancias de algunos modelos que hacen parte de la base de nuestra propuesta. En el segundo capítulo presentamos el modelo de conglomerados y hacemos el análisis bayesiano respectivo. El tercer capítulo está dedicado a los resultados numéricos; en este capítulo se explica en detalle el diseño del estudio de simulación, la generación de los datos y el análisis de los escenarios que forman parte de toda la simulación. En el cuarto capítulo presentamos una extensión del modelo incluyendo covariables y estudiamos tres ejemplos con los datos de varicela en Valencia. Finalizando, el quinto capítulo está dedicado a conclusiones y futuras líneas de investigación.

# Índice general

<b>1. Introducción</b>	<b>1</b>
1.1. Mapeo de enfermedades . . . . .	1
1.2. Estimación del riesgo . . . . .	2
1.3. Detección de conglomerados . . . . .	12
1.3.1. Métodos de diagnóstico para la detección de conglomerados	13
1.3.2. Modelos de Conglomerados . . . . .	15
1.4. Modelos en mapeo de enfermedades que incluyen covariables. . . . .	19
1.5. Objetivos de la tesis . . . . .	21
<b>2. Un modelo de conglomerados con estructura de clases latentes para el análisis de mapas de enfermedades</b>	<b>23</b>
2.1. Un modelo de mixturas de Poisson . . . . .	24
2.1.1. Análisis bayesiano del modelo: distribuciones previas y a posteriori . . . . .	26
2.1.2. Distribuciones condicionales completas . . . . .	32
2.1.3. Muestreo de las distribuciones condicionales completas . . . . .	34
2.2. Situación general: número desconocido de clases . . . . .	38
2.3. Conclusiones sobre cada área pequeña . . . . .	42
2.4. Un ejemplo numérico . . . . .	43
<b>3. Resultados numéricos</b>	<b>53</b>
3.1. Estudio de simulación: presentación . . . . .	53
3.1.1. Generación de los datos y diseño del estudio . . . . .	53
3.1.2. Análisis estadístico y evaluación de los resultados . . . . .	55

3.2. Estudio de simulación: Análisis de algunos escenarios . . . . .	56
3.2.1. Análisis del escenario 1: $k = 1$ . . . . .	57
3.2.2. Análisis del escenario 2: $k = 3$ y riesgos lejanos . . . . .	58
3.2.3. Análisis del escenario 3: $k = 3$ y riesgos parecidos . . . . .	61
3.2.4. Análisis del escenario 4: $k = 5$ . . . . .	64
3.3. Estudio de simulación: comparación de escenarios . . . . .	67
3.4. Estudio de un caso real: varicela en la ciudad de Valencia en el período 2008-2013. . . . .	77
3.4.1. Resultados y comparación con estudios previos. . . . .	79
<b>4. Análisis del modelo de conglomerados incluyendo covariables</b>	<b>85</b>
4.1. Modelo de conglomerados con estructura de clases latentes inclu- yendo covariables y $k$ conocido. . . . .	85
4.1.1. Análisis bayesiano del modelo con covariables: Distribucio- nes previas y a posteriori . . . . .	87
4.1.2. Distribuciones condicionales completas . . . . .	88
4.1.3. Muestreo de las distribuciones condicionales completas . . .	92
4.2. Situación general: número desconocido de clases . . . . .	93
4.3. Conclusiones sobre cada área pequeña . . . . .	95
4.4. Ejemplos numéricos . . . . .	96
4.4.1. Análisis de la covariable ‘tamaño familiar’ . . . . .	96
4.4.2. Análisis de la covariable ‘distancia al centro’ . . . . .	98
4.4.3. Análisis de la covariable ‘vecinos’ . . . . .	99
<b>5. Conclusiones y futuras líneas de investigación</b>	<b>103</b>
5.1. Conclusiones . . . . .	103
5.2. Líneas futuras de investigación . . . . .	105
<b>6. Anexos</b>	<b>109</b>
<b>Referencias</b>	<b>169</b>

# Índice de figuras

2.1.	Trazas de las series para los logaritmos de los riesgos en cada clase.	27
2.2.	Trazas de las series para las probabilidades en cada clase. . . . .	28
2.3.	Riesgos relativos estimados bajo el modelo de conglomerados para los 87 barrios de la ciudad de Valencia, España. . . . .	44
2.4.	Traza del riesgo latente 1 para el banco de datos. . . . .	45
2.5.	Traza del riesgo latente 2 para el banco de datos. . . . .	46
2.6.	Gráfica del diagnóstico de Geweke para el área 73. . . . .	46
2.7.	Curva ROC para las probabilidades de pertenencia a cada una de las clases. . . . .	47
2.8.	Curva ROC para las probabilidades de pertenencia a cada una de las clases con $\eta = (0.8, 1.2)$ . . . . .	48
3.1.	Probabilidades finales de $k$ en escenario 2 . . . . .	59
3.2.	Errores de ajuste del modelo en el escenario 2 . . . . .	61
3.3.	Probabilidad final de $k$ en el escenario 3 . . . . .	63
3.4.	Errores de ajuste del modelo en el escenario 3 . . . . .	64
3.5.	Probabilidad final de $k$ en el escenario 4 . . . . .	66
3.6.	Mapa de la ciudad de Valencia por barrios. Fuente: Oficina de estadística del Ayuntamiento de Valencia ( <a href="http://www.valencia.es/ayuntamiento/estadistica">http://www.valencia.es/ayuntamiento/estadistica</a> ) . . . . .	78
3.7.	Mapa de riesgos estimados de varicela año 2008 con el modelo expuesto en Iftimi et al. (2015b). Fuente: Extraído del artículo de Iftimi et al. (2015b). . . . .	80
3.8.	Mapa de riesgos estimados de varicela año 2013 con el modelo expuesto en Iftimi et al. (2015b). Fuente: Extraído del artículo de Iftimi et al. (2015b). . . . .	80

---

3.9. Mapa de riesgos estimados para la varicela en el año 2008 con el modelo propuesto. . . . .	82
3.10. Mapa de riesgos estimados para la varicela en el año 2013 con el modelo propuesto. . . . .	82
4.1. Mapa de riesgos estimados para la varicela en el año 2008 con la covariable ‘tamaño familiar’ . . . . .	98
4.2. Mapa de riesgos estimados para la varicela en el año 2008 con la covariable ‘distancia al centro’. . . . .	100
4.3. Mapa de riesgos estimados para la varicela en el año 2008 con la covariable ‘vecinos’. . . . .	101

# Índice de tablas

2.1. Código WinBUGS del modelo . . . . .	27
2.2. Parámetros de la simulación con $k = 2$ . . . . .	45
2.3. Resumen de los valores $y_i$ generados en la simulación . . . . .	45
2.4. Resumen numérico e intervalo de probabilidad al 95 % de los parámetros. . . . .	46
2.5. Estimaciones de los parámetros con $k$ real= 2 y $k$ supuesto = 3 . . . . .	49
2.6. Probabilidades finales de $k$ . . . . .	49
2.7. Resumen de los riesgos relativos estimados de cada área pequeña . . . . .	49
2.8. Resumen de los riesgos relativos estimados de cada área pequeña con diferentes $k$ . . . . .	50
2.9. Resumen de los riesgos relativos de cada área pequeña. $k$ conocido vs $k$ desconocido. . . . .	50
2.10. Medidas de la bondad de los resultados . . . . .	50
3.1. Número de escenarios en el estudio de simulación . . . . .	54
3.2. Riesgos subyacentes supuestos . . . . .	54
3.3. Probabilidad de cada una de las clases supuestas . . . . .	55
3.4. Probabilidades finales de $k$ en escenario 1 . . . . .	57
3.5. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ en escenario 1 . . . . .	57
3.6. Errores de ajuste del modelo bajo diferentes incidencias en escenario 1 . . . . .	57
3.7. Mediana (rango intercuartílico) de los valores de $\eta$ y $p$ estimados en el escenario 2 . . . . .	58
3.8. Probabilidades finales de $k$ en escenario 2 . . . . .	58

3.9. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ en escenario 2 . . . . .	58
3.10. Errores de ajuste del modelo . . . . .	60
3.11. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ variando las incidencias. . . . .	60
3.12. Errores de ajuste bajo diferentes valores de la incidencia . . . . .	61
3.13. Mediana (rango intercuartílico) de los valores de $\eta$ y $p$ estimados en el escenario 3 . . . . .	62
3.14. Probabilidades finales de $k$ en el escenario 3 . . . . .	62
3.15. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ .	62
3.16. Errores de ajuste del modelo en el escenario 3 . . . . .	63
3.17. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con diferentes incidencias . . . . .	64
3.18. Errores de ajuste bajo diferentes valores de la incidencia . . . . .	65
3.19. Mediana (rango intercuartílico) de los valores de $\eta$ y $p$ estimados en el escenario 4 . . . . .	65
3.20. Probabilidades finales de $k$ con $k = 5$ en el escenario 4 . . . . .	65
3.21. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ en el escenario 4 . . . . .	66
3.22. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con varias incidencias. . . . .	66
3.23. Errores de ajuste del modelo en el escenario 4 . . . . .	66
3.24. Errores de ajuste bajo diferentes valores de la incidencia . . . . .	67
3.25. $RMSE_{\text{mod}}$ con $r = 3/10^4$ y $k = 2$ . . . . .	67
3.26. $RMSE_{\text{mod}}$ con $r = 30/10^4$ y $k = 2$ . . . . .	68
3.27. $RMSE_{\text{mod}}$ con $r = 300/10^4$ y $k = 2$ . . . . .	68
3.28. $RMSE_{\text{mod}}$ con $r = 3/10^4$ y $k = 3$ . . . . .	69
3.29. $RMSE_{\text{mod}}$ con $r = 30/10^4$ y $k = 3$ . . . . .	69
3.30. $RMSE_{\text{mod}}$ con $r = 300/10^4$ y $k = 3$ . . . . .	69
3.31. $RMSE_{\text{mod}}$ con $r = 3/10^4$ y $k = 5$ . . . . .	70
3.32. $RMSE_{\text{mod}}$ con $r = 30/10^4$ y $k = 5$ . . . . .	70
3.33. $RMSE_{\text{mod}}$ con $r = 300/10^4$ y $k = 5$ . . . . .	70
3.34. $RMSE_{\text{mod}}$ con $r = 3/10^4$ y $k = 7$ . . . . .	71

3.35. $RMSE_{\text{mod}}$ con $r = 30/10^4$ y $k = 7$ . . . . .	71
3.36. $RMSE_{\text{mod}}$ con $r = 300/10^4$ y $k = 7$ . . . . .	71
3.37. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 3/10^4$ y $k = 2$ , . . . . .	72
3.38. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 30/10^4$ y $k = 2$ . . . . .	72
3.39. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 300/10^4$ y $k = 2$ . . . . .	73
3.40. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 3/10^4$ y $k = 3$ . . . . .	73
3.41. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 30/10^4$ y $k = 3$ . . . . .	74
3.42. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 300/10^4$ y $k = 3$ . . . . .	74
3.43. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 3/10^4$ y $k = 5$ . . . . .	74
3.44. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 30/10^4$ y $k = 5$ . . . . .	75
3.45. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 300/10^4$ y $k = 5$ . . . . .	75
3.46. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 3/10^4$ y $k = 7$ . . . . .	75
3.47. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 30/10^4$ y $k = 7$ . . . . .	76
3.48. Porcentaje en el que la mayor probabilidad corresponde a cada $k$ con $r = 300/10^4$ y $k = 7$ . . . . .	76
3.49. Casos de varicela en Valencia en el período 2008 a 2013. . . . .	78
3.50. Resumen de medidas para el riesgo estimado en cada barrio . . . . .	81
3.51. Probabilidades finales de $k$ . Años 2008 y 2013 . . . . .	81
3.52. Resumen de los riesgos relativos de cada área pequeña . . . . .	81
4.1. Resumen de la covariable ‘tamaño familiar’ . . . . .	97
4.2. Estimaciones, por intervalo al 95 %, de $\beta$ y $\sigma$ (‘tamaño familiar’) . . . . .	97
4.3. Resumen de la estimación de los riesgos en áreas pequeñas (‘tamaño familiar’) . . . . .	97

4.4. Resumen de la covariable ‘distancia al centro’ . . . . .	98
4.5. Estimaciones, por intervalo al 95 %, de $\beta$ y $\sigma$ (‘distancia al centro’) . . . . .	99
4.6. Resumen de la estimación de los riesgos en áreas pequeñas (‘distancia al centro’) . . . . .	99
4.7. Resumen de la covariable ‘vecinos’ . . . . .	99
4.8. Estimaciones, por intervalo al 95 %, de $\beta$ y $\sigma$ (‘vecinos’) . . . . .	100
4.9. Resumen de la estimación de los riesgos en áreas pequeñas (‘vecinos’) . . . . .	101
6.1. Estimaciones de los riesgos latentes y probabilidades $k = 2$ . Incidencia $3/10^4$ hab. . . . .	119
6.2. Estimaciones de los riesgos latentes y probabilidades $k = 2$ . Incidencia $3/10^4$ hab. Continuación. . . . .	120
6.3. Estimaciones de los riesgos latentes y probabilidades $k = 2$ . Incidencia $30/10^4$ hab. . . . .	121
6.4. Estimaciones de los riesgos latentes y probabilidades $k = 2$ . Incidencia $30/10^4$ hab. Continuación. . . . .	122
6.5. Estimaciones de los riesgos latentes y probabilidades $k = 2$ . Incidencia $300/10^4$ hab. . . . .	123
6.6. Estimaciones de los riesgos latentes y probabilidades $k = 2$ . Incidencia $300/10^4$ hab. Continuación. . . . .	124
6.7. Estimaciones de los riesgos latentes y probabilidades $k = 3$ . Incidencia $3/10^4$ hab. . . . .	125
6.8. Estimaciones de los riesgos latentes y probabilidades $k = 3$ . Incidencia $3/10^4$ hab. Continuación. . . . .	126
6.9. Estimaciones de los riesgos latentes y probabilidades $k = 3$ . Incidencia $30/10^4$ hab. . . . .	127
6.10. Estimaciones de los riesgos latentes y probabilidades $k = 3$ . Incidencia $30/10^4$ hab. Continuación. . . . .	128
6.11. Estimaciones de los riesgos latentes y probabilidades $k = 3$ . Incidencia $300/10^4$ hab. . . . .	129
6.12. Estimaciones de los riesgos latentes y probabilidades $k = 3$ . Incidencia $300/10^4$ hab. Continuación. . . . .	130
6.13. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $3/10^4$ hab. . . . .	131
6.14. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $3/10^4$ hab. Continuación. . . . .	132

6.15. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $3/10^4$ hab. Continuación. . . . .	133
6.16. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $30/10^4$ hab. . . . .	134
6.17. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $30/10^4$ hab.Continuación . . . . .	135
6.18. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $30/10^4$ hab.Continuación . . . . .	136
6.19. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $300/10^4$ hab. . . . .	137
6.20. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $300/10^4$ hab.Continuación. . . . .	138
6.21. Estimaciones de los riesgos latentes y probabilidades $k = 5$ . Incidencia $300/10^4$ hab.Continuación . . . . .	139
6.22. Estimaciones de los riesgos latentes $k = 7$ . Incidencia $3/10^4$ hab. . .	140
6.23. Estimaciones de los riesgos latentes $k = 7$ . Incidencia $3/10^4$ hab. Continuación. . . . .	141
6.24. Estimaciones de las probabilidades $k = 7$ . Incidencia $3/10^4$ hab. . .	142
6.25. Estimaciones de las probabilidades $k = 7$ . Incidencia $3/10^4$ hab. Continuación. . . . .	143
6.26. Probabilidades finales con $k$ real $k = 1$ ( $\eta = 1, p = 1$ ) . . . . .	145
6.27. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 3.5)$ , Inc. = $3/10^4$ hab . . .	145
6.28. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 3.5)$ , Inc.= $30/10^4$ hab . .	145
6.29. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 3.5)$ , Inc.= $300/10^4$ hab . .	146
6.30. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 2.5)$ , Inc. = $3/10^4$ hab . . .	146
6.31. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 2.5)$ , Inc. = $30/10^4$ hab . .	146
6.32. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 2.5)$ , Inc. = $300/10^4$ hab .	146
6.33. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 1.5)$ , Inc. = $3/10^4$ hab . . .	146
6.34. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 1.5)$ , Inc. = $30/10^4$ hab . .	147
6.35. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.5, 1.5)$ , Inc. = $300/10^4$ hab .	147
6.36. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.8, 1.2)$ , Inc. = $3/10^4$ hab . . .	147
6.37. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.8, 1.2)$ , Inc. = $30/10^4$ hab . .	147
6.38. Prob. finales con $k = 2, (\eta_1, \eta_2) = (0.8, 1.2)$ , Inc. = $300/10^4$ hab .	147
6.39. Prob. finales con $k = 3, (\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc.= $3/10^4$ hab.	148

6.40. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc. = $30/10^4$ hab. . . . .	148
6.41. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc. = $300/10^4$ hab. . . . .	148
6.42. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc. = $3/10^4$ hab.	148
6.43. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc. = $30/10^4$ hab. . . . .	148
6.44. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc. = $300/10^4$ hab. . . . .	149
6.45. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc. = $3/10^4$ hab.	149
6.46. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc. = $30/10^4$ hab. . . . .	149
6.47. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc. = $300/10^4$ hab. . . . .	149
6.48. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc. = $3/10^4$ hab.	149
6.49. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc. = $30/10^4$ hab. . . . .	150
6.50. Prob. finales con $k = 3$ , $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc. = $300/10^4$ hab. . . . .	150
6.51. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc. = $3/10^4$ hab. . . . .	150
6.52. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc. = $30/10^4$ hab. . . . .	150
6.53. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc. = $300/10^4$ hab. . . . .	150
6.54. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc. = $3/10^4$ hab. . . . .	151
6.55. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc. = $30/10^4$ hab. . . . .	151
6.56. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc. = $300/10^4$ hab. . . . .	151
6.57. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc. = $3/10^4$ hab. . . . .	151
6.58. Prob. finales con $k = 5$ , $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc. = $30/10^4$ hab. . . . .	151

6.59. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.= $300/10^4$ hab. . . . .	152
6.60. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.= $3/10^4$ hab. . . . .	152
6.61. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.= $30/10^4$ hab. . . . .	152
6.62. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.= $300/10^4$ hab. . . . .	152
6.63. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.= $3/10^4$ hab. . . . .	152
6.64. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.= $30/10^4$ hab. . . . .	153
6.65. Prob. finales con $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.= $300/10^4$ hab. . . . .	153
6.66. Prob. finales con $k = 7, \eta' s = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.= $3/10^4$ hab. . . . .	153
6.67. Prob. finales con $k = 7, \eta' s = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.= $30/10^4$ hab. . . . .	153
6.68. Prob. finales con $k = 7, \eta' s = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.= $300/10^4$ hab. . . . .	154
6.69. Prob. finales con $k = 7, \eta' s = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.= $3/10^4$ hab. . . . .	154
6.70. Prob. finales con $k = 7, \eta' s = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.= $30/10^4$ hab. . . . .	154
6.71. Prob. finales con $k = 7, \eta' s = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.= $300/10^4$ hab. . . . .	154
6.72. Prob. finales con $k = 7, \eta' s = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.= $3/10^4$ hab. . . . .	155
6.73. Prob. finales con $k = 7, \eta' s = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.= $30/10^4$ hab. . . . .	155
6.74. Prob. finales con $k = 7, \eta' s = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.= $300/10^4$ hab. . . . .	155
6.75. Prob. finales con $k = 7, \eta' s = (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5)$ , Inc.= $3/10^4$ hab. . . . .	155
6.76. Prob. finales con $k = 7, \eta' s = (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5)$ , Inc.= $30/10^4$ hab. . . . .	156

6.77. Prob. finales con $k = 7$ , $\eta' s = (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	156
6.78. Prob. finales con $k = 7$ , $\eta' s = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	156
6.79. Prob. finales con $k = 7$ , $\eta' s = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	156
6.80. Prob. finales con $k = 7$ , $\eta' s = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	157
6.81. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=1. $\eta = 1$ , $p = 1$ . . . . .	157
6.82. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2, $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	157
6.83. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2, $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	158
6.84. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2, $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	158
6.85. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2, $(\eta_1, \eta_2) = (0.5, 2.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	158
6.86. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2, $(\eta_1, \eta_2) = (0.5, 2.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	158
6.87. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2, $(\eta_1, \eta_2) = (0.5, 2.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	158
6.88. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2 $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	159
6.89. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2 $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	159
6.90. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2 $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	159
6.91. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2 $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	159
6.92. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2 $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	159
6.93. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=2 $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	160
6.94. Porcentaje de correspondencia entre el máximo de la probabilidad final con k real=3, $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc.= 3/10 <sup>4</sup> hab. . .	160



6.111	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	164
6.112	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	164
6.113	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	164
6.114	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	164
6.115	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	165
6.116	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	165
6.117	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	165
6.118	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	165
6.119	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	165
6.120	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=5, $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	166
6.121	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.= 3/10 <sup>4</sup> hab. . . . .	166
6.122	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.= 30/10 <sup>4</sup> hab. . . . .	166
6.123	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.= 300/10 <sup>4</sup> hab. . . . .	166

---

6.124	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.= $3/10^4$ hab. . . . .	167
6.125	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.= $30/10^4$ hab. . . . .	167
6.126	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.= $300/10^4$ hab. . . . .	167
6.127	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.= $3/10^4$ hab. . . . .	167
6.128	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.= $30/10^4$ hab. . . . .	168
6.129	Porcentaje de correspondencia entre el máximo de la probabilidad final con $k$ real=7, $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.= $300/10^4$ hab. . . . .	168



# Capítulo 1

## Introducción

### 1.1. Mapeo de enfermedades

Desde hace un poco más de dos décadas ha ido creciendo en el campo de la salud, y más específicamente en Epidemiología, un gran interés por estudiar la evolución de ciertas enfermedades que de algún modo representan una amenaza para la salud de las personas. Para hablar de dichas amenazas tenemos que referirnos en la literatura epidemiológica al término riesgo o factor de riesgo, definido por la Organización Mundial de la salud (OMS) como ‘cualquier rasgo, característica o exposición de un individuo que aumente su probabilidad de sufrir una enfermedad o lesión.’ Bajo esta definición, muchos organismos han expresado su preocupación acerca de cómo los factores de riesgo medioambientales y en general de contaminación pueden estar afectando la salud de las poblaciones. Esto ha generado que investigadores de diferentes áreas estén trabajando en el desarrollo de metodologías estadísticas para el análisis del riesgo de enfermedades. Uno de los trabajos pioneros en esta materia de investigación es la monografía de Doll (1980), donde se describen algunas de las primeras hipótesis concernientes a la influencia de factores medioambientales y del estilo de vida en la mortalidad por cáncer y discute cómo dichas influencias pueden ser observadas al estudiar la distribución geográfica de la enfermedad.

El objetivo principal de esta tesis es el estudio de la variación geográfica del riesgo de una cierta enfermedad en una determinada zona de estudio. Así pues, nuestro tema de estudio se enmarca en el ámbito de la epidemiología espacial cuyo foco es el análisis de datos de salud en áreas pequeñas. Cuando hablamos de datos espaciales nos referimos a aquellas mediciones u observaciones realizadas en localizaciones o en áreas específicas. Uno de los principales objetivos de la epidemiología espacial es el de mostrar qué parte de la variación espacial de la ocurrencia de una enfermedad no está explicada ni por la distribución espacial de factores explicativos conocidos ni por una variación aleatoria. De hecho, a

menudo interesa encontrar pistas sobre algún factor de riesgo desconocido de esta enfermedad (Barceló et al., 2008).

Abordaremos nuestro estudio desde el mapeo de enfermedades (disease mapping, en inglés), el cual constituye una de las herramientas más usadas en la epidemiología espacial para el estudio del comportamiento de una enfermedad en una zona geográfica determinada. A través de dicha herramienta se desea conocer la distribución espacial y/o espacio-temporal de una enfermedad y trabajar en la identificación de factores de riesgo que podrían explicar dicho comportamiento. El mapeo de enfermedades intenta resumir la variación espacial del riesgo para evaluar y cuantificar la magnitud de la heterogeneidad espacial y patrones asociados, destacando así zonas de alto y bajo riesgo (Best et al., 2005).

Mediante el mapeo de enfermedades se pueden abordar los siguientes objetivos (Lawson, 2006; Botella et al., 2013; Etxeberria, 2012): identificar patrones espaciales de la enfermedad; mostrar la evolución de la enfermedad en el espacio y tiempo; y formular hipótesis asociadas a la etiología de la enfermedad, que permiten determinar con mayor precisión las causas subyacentes en la aparición (ya sea en forma de incidencia, prevalencia o mortalidad) de enfermedades. La exploración geográfica de distribuciones diferentes de enfermedades supone una herramienta de gran utilidad para la toma de decisiones. De esta forma, los gestores de sistemas de salud pueden tomar decisiones basadas en información real y objetiva que permitirán desarrollar intervenciones acertadas desde un punto de vista costo-beneficio.

## 1.2. Estimación del riesgo

Por lo general, los datos de partida en la estimación del riesgo de una enfermedad son de naturaleza discreta y corresponden a conteos de la enfermedad de interés dentro de un área geográfica dividida en áreas pequeñas (Elliott et al., 2000). Si en las áreas pequeñas contamos ocurrencias locales de la enfermedad en un corto período de tiempo hablamos de incidencia. Cuando estudiamos la acumulación de casos de la enfermedad a largo plazo nos referimos a prevalencia. Si la información es de datos relacionados con el número de muertes debido a cierta enfermedad, estamos estudiando mortalidad. Y si, en general, tenemos información acerca de número de personas que enferman en un lugar y tiempo determinado hablamos de morbilidad.

Para analizar en detalle lo anterior vamos a asumir que la región de estudio está dividida en  $m$  áreas pequeñas contiguas. En cada una de esas áreas revisaremos los casos observados  $y_i$ , los cuales corresponden al número de personas enfermas o fallecidas debido a la enfermedad en estudio. En la literatura encontramos comúnmente que los casos observados  $y_i$  siguen una distribución Poisson con media  $e_i\lambda_i$ :

$$y_i | e_i, \lambda_i \sim \text{Po}(e_i \lambda_i), \quad (1.1)$$

con  $i = 1, \dots, m$ , donde  $e_i$  es el número de casos esperados en la  $i$ -ésima área y  $\lambda_i$  es el riesgo relativo. Dados estos riesgos, los datos son independientes. Los valores esperados son calculados bajo un modelo que supone igual riesgo en todas las áreas, usualmente mediante estandarización indirecta por edad o sexo sobre la población de interés (Fleiss, 1981; Elliott et al., 2000):

$$e_i = \sum_{j=1}^J N_{ij} \frac{y_{.j}}{N_j},$$

donde  $N_{ij}$  representa el número de personas en riesgo en el área  $i$  y grupo de edad o sexo  $j$ .  $N_j$  es el número de personas en riesgo en el grupo con edad o sexo  $j$ , es decir  $N_j = \sum_{i=1}^m N_{ij}$ . Además  $y_{.j}$  es el número de casos observados en el grupo

(edad o sexo)  $j$ , es decir  $y_{.j} = \sum_{i=1}^m y_{ij}$ .

En estos estudios el interés recae sobre la estimación de los riesgos relativos  $\{\lambda_i\}_{i=1}^m$ . El estimador de máxima verosimilitud del riesgo relativo en cada área es llamado tasa de mortalidad o morbilidad estandarizada (SMR):

$$\text{SMR}_i = \hat{\lambda}_i = \frac{y_i}{e_i}.$$

La varianza del estimador máximo verosímil viene dada por:

$$\text{Var}(\hat{\lambda}_i) = \text{Var}\left(\frac{y_i}{e_i}\right) = \frac{\text{Var}(y_i)}{e_i^2} = \frac{e_i \lambda_i}{e_i^2} = \frac{\lambda_i}{e_i},$$

y entonces:

$$\widehat{\text{Var}}(\hat{\lambda}_i) = \frac{\hat{\lambda}_i}{e_i} = \frac{y_i}{e_i^2},$$

por lo que estas tasas trabajan bien si  $e_i$  es grande, pero son extremadamente variables cuando  $e_i$  es pequeño, produciendo lo que se denomina extra-variabilidad. Además, su representación a través de un mapa exhibe una variación amplia y poco informativa de los patrones existentes (Silva et al., 2003). Todo lo anterior trae como consecuencia la dificultad o incluso el impedimento de hacer una correcta estimación del riesgo subyacente en los datos. De ahí la importancia del

desarrollo de modelos de estimación y suavización del riesgo en áreas pequeñas que ofrezcan una estimación más estable del riesgo.

El modelo con tendencia simple, explicado en Rotejanaprasert (2014) y en el capítulo 3 de Lawson (2013), modeliza el logaritmo del riesgo utilizando las coordenadas geográficas de los centroides de cada una de las áreas:

$$\log(\lambda_i) = \alpha_0 + \alpha_x S_{x_i} + \alpha_y S_{y_i},$$

donde  $\alpha_0$  es el riesgo general,  $\alpha_x$  y  $\alpha_y$  son los coeficientes de las coordenadas geográficas y  $S_{x_i}$  y  $S_{y_i}$  son las coordenadas geográficas del centroide de la  $i$ -ésima área. Las distribuciones previas asumidas para los coeficientes son  $\alpha_0 \sim N(0, \sigma_{\alpha_0}^2)$ ,  $\alpha_x \sim N(0, \sigma_{\alpha_x}^2)$ ,  $\alpha_y \sim N(0, \sigma_{\alpha_y}^2)$ , donde  $\sigma_* \sim U(0, 100)$ .

Este modelo en algunos casos es más parsimonioso y ajusta mejor que modelos más sofisticados de efectos aleatorios en ciertas situaciones.

Al estimar el riesgo de una enfermedad en un área pequeña deseamos obtener una estimación fiable del mismo al tiempo que nos gustaría recoger la mayor variabilidad posible. Una primera propuesta consiste en asumir que no hay dependencia espacial entre las observaciones. Entonces, el logaritmo del riesgo de la enfermedad se puede modelizar como (Lawson, 2013):

$$\log(\lambda_i) = \alpha_0 + v_i,$$

donde  $\alpha_0$  es el intercepto y describe un riesgo común o general en el área de estudio, y  $v_i$  es un efecto aleatorio que está asociado al ruido observacional, generalmente conocido como heterogeneidad no correlacionada.

En un contexto jerárquico bayesiano, el modelo anterior puede especificarse como:

$$\begin{aligned} y_i | e_i, \lambda_i &\sim \text{Po}(e_i \lambda_i) \\ \log(\lambda_i) &= \alpha_0 + v_i \\ \alpha_0 &\sim N(0, \sigma_\alpha^2) \\ v_i &\sim N(0, \sigma_v^2). \end{aligned}$$

Para los efectos aleatorios generalmente se recomienda que las precisiones  $\sigma_\alpha^{-2}$  y  $\sigma_v^{-2}$  sean estimables y hay que asignarles una distribución previa. Usualmente se emplean distribuciones Gamma para estas precisiones:

$$\sigma_*^{-2} \sim \text{Ga}(a, b),$$

donde  $a$  y  $b$  suelen suponerse que siguen una distribución poco informativa o con varianza grande. Alternativamente, Gelman (2006) sugiere una distribución Uniforme  $U(0, 10)$  sobre  $\sigma_*$ .

El trabajo realizado por Besag et al. (1991), más conocido como el modelo de convolución BYM, es uno de los trabajos más importantes realizados en la modelización espacial del riesgo y es frecuentemente utilizado en la literatura epidemiológica (Mollié, 1996; Ferrándiz et al., 2002, 2004; Barceló et al., 2008). La idea principal del modelo es suponer que existen factores de riesgo que abarcan más de un área de estudio y, en consecuencia, los riesgos relativos son espacialmente dependientes. Así pues, en el segundo nivel de jerarquía, el logaritmo del riesgo relativo es modelizado en función de efectos aleatorios espacialmente correlacionados y no correlacionados:

$$\log(\lambda_i) = \alpha_0 + v_i + u_i,$$

donde  $\alpha_0$  identifica el riesgo común en la zona de estudio;  $u_i$  es la componente aleatoria con dependencia espacial y  $v_i$  es la heterogeneidad no correlacionada.

Como distribución previa para la componente espacial suele utilizarse la distribución auto-regresiva condicional intrínseca (ICAR):

$$f(u) \propto \frac{1}{\sigma_u^m} \exp \left\{ \frac{-1}{2\sigma_u^2} \sum_i \sum_{j \in \delta_i} (u_i - u_j)^2 \right\}, \quad (1.2)$$

donde  $\delta_i$  es la vecindad de la  $i$ -ésima área. Generalmente se trabaja con una estructura de vecindad de primer orden; es decir,  $\delta_i$  está compuesto por el conjunto de áreas pequeñas que comparten límites geográficos con el área  $i$ . En este caso, la distribución sería un ejemplo de campo aleatorio de Markov (Rue and Held, 2005). Notar que la distribución ICAR supone que el valor esperado de cada elemento de  $u$  coincide con la media de los elementos correspondientes a las regiones vecinas. Así pues, la distribución ICAR puede formularse alternativamente como:

$$u_i | u_{-i} \sim N(\bar{u}_i, \frac{\sigma_u^2}{n_{\delta_i}}),$$

donde  $n_{\delta_i}$  es el número de vecinos de la región  $i$  y  $\sigma_u^{-2}$  es el parámetro de precisión que controla el grado de suavizado. Además,  $\bar{u}_i = \sum_{j \in \delta_i} u_j / n_{\delta_i}$ .

Una ventaja de esta formulación intrínseca Gaussiana es que los momentos condicionales son definidos como simples funciones de los valores vecinos y número de vecinos:

$$\begin{aligned} E(u_i | u_{-i}) &= \bar{u}_i \\ \text{Var}(u_i | u_{-i}) &= \sigma_u^2 / n_{\delta_i}. \end{aligned}$$

Otra posible distribución para la previa de la componente correlacionada  $u_i$  es la distribución CAR propia:

$$u_i \sim N(\mu_i, \sigma_u^2/n_{\delta_i}),$$

siendo

$$\mu_i = t_i + \phi \sum_{j \in \delta_i} (u_j - t_j)/n_{\delta_i},$$

donde  $t_i = x_i' \beta$ , siendo  $x_i'$  la  $i$ -ésima fila de la matriz de diseño que incluye covariables y  $\beta$  el vector de parámetros en el modelo. El término  $\phi$  controla la estructura de correlación de forma que para  $\phi > 0$  las regiones próximas tomarán valores similares (dependencia espacial), para  $\phi = 0$  las observaciones serán independientes y para  $\phi < 0$  las regiones próximas tenderán a tomar valores opuestos (dependencia espacial negativa). En un simple caso donde no hay tendencia entonces la expresión anterior se reduce a:

$$\mu_i = \phi \bar{u}_i.$$

La heterogeneidad no correlacionada  $v_i$  fue definida por Besag et al. (1991) para tener como previa una distribución Gaussiana con media cero. Por lo tanto la previa para el vector  $v$  es:

$$f(v) \propto \frac{1}{\sigma_v^m} \exp \left\{ \frac{-1}{2\sigma_v^2} \sum_{i=1}^m v_i^2 \right\}. \quad (1.3)$$

La previa para el intercepto es:

$$\alpha_0 \sim N(0, \sigma_\alpha^2).$$

Finalmente, la distribución posterior conjunta dado el vector de datos  $y = (y_1, \dots, y_m)$  se puede expresar como:

$$\begin{aligned}
f(\alpha_0, u, v, \sigma_\alpha, \sigma_u, \sigma_v | y) &\propto \prod_{i=1}^m \exp\{-e_i \lambda_i\} (e_i \lambda_i)^{y_i} / y_i! \\
&\times \frac{1}{\sigma_{\alpha_0}} \exp\{-\alpha_0^2 / 2\sigma_{\alpha_0}^2\} \\
&\times \frac{1}{\sigma_u^m} \exp\left\{\frac{-1}{2\sigma_u^2} \sum_i \sum_{j \in \delta_i} (u_i - u_j)^2\right\} \\
&\times \frac{1}{\sigma_v^m} \exp\left\{\frac{-1}{2\sigma_v^2} \sum_{i=1}^m v_i^2\right\} \\
&\times \text{previa}(\sigma_\alpha, \sigma_u, \sigma_v).
\end{aligned}$$

Esta distribución posterior puede ser muestreada usando algoritmos MCMC tales como el Gibbs Sampling (Geman and Geman, 1984) o Metropolis-Hastings (Metropolis and Ulam, 1949; Hastings, 1970). En la actualidad, algunos autores están implementando el enfoque numérico basado en la aproximación de Laplace denominado INLA (Integrated Nested Laplace Approximation) (Rue et al., 2009). Es habitual encontrar que para hacer inferencia bayesiana en modelos Gaussianos con estructura latente se implemente INLA por su alta velocidad computacional. Sin embargo, para el estudio de modelos de mixturas no hemos encontrado implementaciones en INLA, quizá esta puede ser una interesante línea de trabajo.

Como se mencionó anteriormente, en este modelo se ha incorporado dependencia espacial atendiendo a la idea de que observaciones de áreas geográficas próximas se parecerán más entre sí que las observaciones de áreas geográficas más distantes, logrando así una estimación suavizada del riesgo. La introducción de esta estructura de correlación espacial en el modelo da información adicional y permite la obtención de estimaciones más fiables de los riesgos relativos que la estimación máximo verosímil.

Algunos autores han evaluado este modelo. Por ejemplo, Lawson et al. (2000) realizan varios ejercicios de simulación para estudiar diferentes aspectos de los rendimientos de los modelos de mapeo de enfermedades. En dicho estudio se concluyó que el modelo más robusto entre aquellos con estructura espacial fue el modelo de BYM. Otro trabajo más reciente en este ámbito es el de Ugarte et al. (2009c), donde se revisa el comportamiento de modelos condicionales autoregresivos espacio-temporales en la detección de áreas con riesgo extremo. Los autores comparan seis modelos de mapeo de enfermedades espacio-temporales, la mayoría de ellos con evolución temporal paramétrica. Nuevamente se evidencia la robustez del modelo de BYM. Otros trabajos que incluyen análisis y evaluación de este modelo son Best et al. (2005); Lawson et al. (2000); Botella et al. (2013) y Rotejanaprasert (2014).

Botella (2010) presenta una propuesta de suavización espacial de los riesgos que incorpora ideas de los procesos de medias móviles del análisis de series temporales. Bajo dicho modelo, el logaritmo del riesgo relativo de cada área pequeña se modeliza como:

$$\log(\lambda_i) = \mu + \phi_i^{-1}(\omega_0\theta_i + \omega_1(\sum_{j \sim 1^i} \theta_j) + \dots + \omega_m(\sum_{j \sim m^i} \theta_j)), \quad (1.4)$$

donde  $\theta_i$  representa el riesgo latente asociado a la  $i$ -ésima área. Así pues, el logaritmo del riesgo relativo de cada área se expresa como suma de un término de interceptación  $\mu$ , que modeliza el valor medio de los riesgos de toda la región de estudio, más un segundo término proporcional a la suma de los efectos de los riesgos latentes, tanto de la propia región como de las vecinas. En concreto, suponen que el riesgo de un área depende de su propio efecto latente, más el efecto latente de sus regiones vecinas de orden 1, más el efecto latente de sus regiones vecinas de orden 2, y así sucesivamente hasta llegar a las regiones vecinas de orden  $m$ . La notación  $j \sim v^i$  indica que el área  $j$  es vecina de orden  $v$  del área  $i$ . Los riesgos latentes de los vecinos de diferente orden son ponderados mediante unos pesos, que controlan la importancia relativa de las regiones más alejadas respecto a las más próximas en el proceso de suavización; es decir, cuanto mayor sea el valor de  $\omega_1$  en relación a  $\omega_0$  el efecto espacial será más marcado.

En un análisis bayesiano del modelo, las previas sugeridas por Botella (2010) son:

$$\begin{aligned} \mu &\sim U(-\infty, \infty) \\ \theta|\sigma &\sim N_m(0, \sigma^2\mathbf{I}) \\ \sigma &\sim U(0, \infty) \\ \omega|m &\sim \text{Dir}(1_{m+1}) \\ p(m) &\propto \frac{1}{m!}. \end{aligned}$$

Debido a que la varianza del logaritmo del riesgo relativo en cada área está en función de la configuración de vecinos de dicha área, es posible que aparezcan efectos frontera (y quizás otros efectos menos evidentes) sobre la estimación de la variación espacial del riesgo. Para corregir este efecto, el parámetro  $\phi_i$  es definido como:

$$\phi_i = (\omega_0^2 + \omega_1^2 n_1(i) + \dots + \omega_m^2 n_m(i))^{1/2}.$$

Alternativamente, es posible abordar el problema de estimación del riesgo desde una perspectiva en la cual se busque algún tipo de agrupamiento entre las

áreas pequeñas. Existen dos clases de agrupamiento: el centrado y el no centrado. El primer tipo de agrupamiento asume que existe una fuente de riesgo (por ejemplo una antena telefónica, un vertedero de residuos, etc.) de manera que áreas cercanas a la fuente de riesgo son más probables para tener la enfermedad que aquellas que están más lejos; es decir, se asume que el riesgo decrece conforme aumenta la distancia a la fuente.

Supongamos que los datos  $y_i$  siguen un comportamiento Poisson. Entonces, bajo el supuesto de agrupamiento centrado, el logaritmo del riesgo es modelizado como sigue:

$$\log(\lambda_i) = \log[1 + \exp(-\alpha d_i)] + X_i' \beta + Z_i' \gamma, \quad (1.5)$$

donde  $d_i$  es la distancia al área pequeña  $i$  desde el punto focal y  $\alpha$  define la relación de distancia (Wakefield and Morris, 2001; Lawson, 2010).  $X$  es la matriz de covariables;  $Z$  es una matriz de efectos aleatorios y  $\gamma$  es el vector unitario.

En un análisis bayesiano del modelo, las previas sugeridas son:

$$\begin{aligned} \alpha &\sim N(0, \sigma_\alpha^2) \\ \beta &\sim N(0, \sigma_\beta^2). \end{aligned}$$

Del agrupamiento no centrado se hablará en la siguiente sección.

Una alternativa convencional de modelización del riesgo es considerar que el riesgo está compuesto por una combinación de niveles de riesgo no observados o latentes. En esta línea, encontramos los denominados modelos de mixturas (ver Lawson, 2009, cap. 2). Por ejemplo, se puede modelizar el logaritmo del riesgo de la siguiente forma:

$$\log(\lambda_i) = v_i + w_i u_{1,i} + (1 - w_i) u_{2,i} \quad (1.6)$$

$$w_i \sim Be(\alpha, \alpha) \quad (1.7)$$

donde  $u_{1,i}$  y  $u_{2,i}$  son términos asociados a la heterogeneidad correlacionada y a los cuales se les asigna como previa una distribución CAR. Para la heterogeneidad no correlacionada se asume  $v_i \sim N(0, \sigma_v)$ . El parámetro de mixtura  $\{w_i\}$  varía espacialmente.

Uno de los primeros modelos bayesianos de mixtura fue propuesto por Schlattman and Böhning (1993). Su enfoque asume que la distribución de los datos observados es una mixtura de distribuciones Poisson:

$$f(y_i | p, e_i, \lambda) = \sum_{k=1}^K p_k \text{Po}(y_i | e_i \lambda_k),$$

con probabilidades de la mixtura  $\{p_k\}$  tales que  $\sum_k p_k = 1$ . Ambos parámetros  $p$  y  $\lambda$  son desconocidos. Una vez estimados  $p$  y  $\lambda$ , se puede estimar el riesgo en cada área como:

$$\hat{\lambda}_i = \frac{1}{G} \sum_{g=1}^G \sum_{k=1}^K \lambda_k^g p_k^g \text{Po}(y_i | e_i \lambda_k^g) / \sum_{k=1}^K p_k^g \text{Po}(y_i | e_i \lambda_k^g).$$

La previa para  $p \sim \text{Dir}(\alpha)$ , donde  $\{\alpha_k\}$ ,  $k = 1, \dots, K$ , mientras que para  $\{\alpha_k\}$  una Gamma puede ser especificada. La previa para cada  $\lambda_k$  puede estar basada en una distribución Gamma, por ejemplo:

$$\lambda_k \sim \text{Ga}(c_k a, a).$$

El parámetro  $K$  puede estar fijado o, alternativamente, puede ser asignada una distribución previa como por ejemplo, una Poisson con tasa fija o una Uniforme  $K \sim U(1, K_{max})$ . Fernández and Green (2002), sugieren distintas extensiones de este modelo.

Continuando con los enfoques basados en modelos de mixturas, encontramos el trabajo de Militino et al. (2001), donde se comparan dos enfoques para introducir efectos aleatorios, con el fin de revisar cómo cada uno de esos modelos identifica zonas de alto riesgo.

El primer modelo es un modelo de mixtura no paramétrico donde se asume que la población viene de una mixtura discreta de distribuciones Poisson:

$$y_i \sim \sum_{k=1}^K p_k \text{Po}(e_i R_k),$$

donde asociados con cada región se encuentran los efectos aleatorios que modelizan heterogeneidad. El modelo de mixtura no paramétrico asume que, en una región específica, los efectos aleatorios tienen una distribución de probabilidad discreta tomando  $k$  valores  $R_1, \dots, R_k$ , con probabilidades  $p_1, \dots, p_k$ , respectivamente. Cada una de esas  $K$  componentes de la mixtura representa un conglomerado que contiene una proporción de la población  $p_k$  con tasa de mortalidad  $R_k$  y con la restricción de que  $\sum_{k=1}^K p_k = 1$ . Si se conoce el conglomerado al cual pertenece cada área pequeña, la verosimilitud tiene la siguiente forma:

$$\prod_{i=1}^N \left[ \frac{(e_i R_1)^{y_i} \exp(-e_i R_1)}{y_i!} p_1 \right]^{g_{i1}} \left[ \frac{(e_i R_2)^{y_i} \exp(-e_i R_2)}{y_i!} p_2 \right]^{g_{i2}} \dots \\ \times \left[ \frac{(e_i R_k)^{y_i} \exp(-e_i R_k)}{y_i!} p_k \right]^{g_{ik}},$$

donde  $g_{ik} = 1$  si la observación  $i$  pertenece a la componente  $k$  y 0 de otra forma y  $\sum_{k=1}^K g_{ik} = 1$ .

El segundo enfoque que estudian es el modelo de mixtura Poisson-Normal, el cual permite autocorrelación espacial a través del término  $u_i$  que incluye variación espacial en el riesgo relativo a través de un modelo de regresión Gaussiano.

Concluyen que los modelos de mixtura discretos son capaces de localizar regiones con alto riesgo, los modelos de mixtura Normal también trabajan bien y mejoran notablemente la estimación cuando la autocorrelación espacial está presente.

Otra propuesta dentro de los modelos de mixtura para la estimación del riesgo es la de Lawson and Clark (2002), en la cual trabajan un modelo de mixtura espacial en mapeo de enfermedades. En esta propuesta modelizan el parámetro de interés a través del logaritmo del riesgo, así:

$$\log(\lambda_i) = p_i u_i + (1 - p_i) w_i + v_i, \quad (1.8)$$

donde los términos  $p_i$  y  $(1 - p_i)$  controlan la mixtura de las dos componentes espaciales ( $u_i$  y  $w_i$ ) que tienen como previa una distribución CAR. En el caso de que  $p_i = 1$ , entonces el modelo quedará reducido al BYM. Para  $p_i$  se sugiere una previa poco informativa, por ejemplo:

$$p_i \sim \text{Be}(a, b),$$

donde  $a$  y  $b$  pueden tener como previas una Uniforme (0,10).

Para finalizar esta sección incluimos en nuestra revisión los modelos basados en splines utilizados en suavización espacial. La idea de estos modelos se centra en el supuesto de un operador de suavización para representar la estructura media del proceso (Ugarte et al., 2009a, 2010). Como es habitual los datos observados se distribuyen con una Poisson. Además definen la geo-referencia para la observación  $i$  como  $s_i : (x_{i1}, x_{i2})$ , pudiendo ser este el centroide de la  $i$ -ésima área u otro punto de referencia asociado. Los modelos basados en splines son atractivos en aplicaciones y tienen fuertes enlaces con modelos que incluyen procesos Gaussianos (por ejemplo, French y Wand, 2004). Entonces,

$$\begin{aligned} \log(\mu_i) &= \alpha_0 + \sum_{j=1}^2 \alpha_j x_{ij} + \sum_{j=1}^{n_k} \psi_j C\{\|s_i - k_j\|\} \\ &= x_i' \alpha + z_i' \psi, \end{aligned}$$

donde  $\{k_j\}$ ,  $j = 1, \dots, n_k$ , es un conjunto de localizaciones fijadas en el espacio,  $\{\psi_j\}$  es un efecto aleatorio Gaussiano,  $z_i' = (z_1, \dots, z_{n_k})$  y  $[C\{\|s_i - k_j\|/\rho\}]_{1 \leq i \leq m, 1 \leq j \leq n_k}$ , es la función de covarianza definida como:

$$C\{a\} = (1 + |a|) \exp\{-|a|\}.$$

Definen la matriz cuadrada

$$\omega = [C\{\|k_i - k_j\|/\rho\}]_{1 \leq i, j \leq n_k},$$

donde  $\rho$  es un término que controla el grado de suavizado y la distribución previa para el efecto aleatorio conjunto es:

$$\psi \sim N(0, \tau\omega^{-1}).$$

Además de los modelos estudiados en esta sección para la estimación del riesgo, es necesario que revisemos una segunda herramienta principal en el área de mapeo de enfermedades. Los modelos que estudiaremos a continuación se centran en detectar discontinuidades y detectar grupos de riesgo inusual.

### 1.3. Detección de conglomerados

En ocasiones, el objetivo del estudio se centra en dar respuesta a preguntas del tipo: ¿Dónde están localizadas las áreas con alto riesgo? ¿Están esas localizaciones estructuradas en el espacio? Si es así ¿cómo? ¿Están enlazadas a factores medioambientales? Intentando responder a los interrogantes anteriores, en las últimas décadas se ha incrementado el interés en el análisis y detección de conglomerados en áreas pequeñas. Como un ejemplo de aplicación podemos mencionar el trabajo de Wheeler (2007), donde la detección de conglomerados es una importante herramienta en la vigilancia de cáncer para identificar áreas de elevado riesgo y generar hipótesis acerca de la etiología de dicha enfermedad.

Antes de adentrarnos en la modelización de conglomerados por riesgo de enfermedades debemos revisar algunas definiciones que nos ayudarán a comprender mejor la terminología empleada en esta sección. Cuando se habla de conglomerado (cluster en inglés) o agrupamiento global, se asume que la superficie de riesgo está agrupada o tiene áreas de igual riesgo. Por otro lado, una superficie no agrupada muestra cambios aleatorios de riesgo; es decir, el nivel de riesgo en cada una de las áreas es mucho más variable y tiene pocas áreas contiguas de igual riesgo. Un conglomerado espacial es cualquier zona espacialmente delimitada con riesgos significativamente elevados o reducidos. Cuando se habla de que una zona está espacialmente delimitada significa que la región debe tener integridad espacial por ejemplo a través de un criterio de vecindad (aquí las áreas son contiguas). La forma más intuitiva de agrupamiento es el conglomerado por punto caliente (hot-spot clustering) donde cualquier área pequeña o región puede ser considerada por sí sola como un conglomerado, esto debido al criterio de cero vecinos. Richardson

et al. (2004) definen reglas de decisión para clasificar las regiones como zonas de alto riesgo. Definen un valor o límite de referencia en la distribución posterior del riesgo relativo, y un valor de corte en la probabilidad posterior de que el riesgo relativo es superior al valor de referencia. Luego, una región es clasificada como una zona de alto riesgo si la probabilidad posterior de que el riesgo relativo sea superior al umbral de referencia es mayor que el valor de corte. Además, se refieren a la idoneidad de la regla con el número de casos esperados, de manera que una zona de alto riesgo se detecta si el exceso de riesgo es 3 o más, o si el riesgo es 2 y el número de casos esperados es alta (cincuenta o más).

Dentro de la detección de conglomerados podemos hablar de agrupamiento no centrado cuando tenemos localizaciones no conocidas de agrupamiento. Aquí, las localizaciones de los conglomerados son desconocidas, así como el número y el tamaño de los mismos.

### 1.3.1. Métodos de diagnóstico para la detección de conglomerados

A continuación presentamos algunos métodos que han sido utilizados para identificar grupos con riesgo inusual (elevado/bajo). Trabajos como los de Richardson et al. (2004), Hossain and Lawson (2006) y Ugarte et al. (2009b) entre otros, han explorado los residuales de un modelo de estimación del riesgo con el objetivo de encontrar áreas con riesgos excesivos que no han sido explicados por el modelo. Los residuales pueden ser calculados fácilmente como:

$$r_i = (y_i - e_i \hat{\lambda}_i) / \sqrt{e_i \hat{\lambda}_i}, \quad (1.9)$$

donde  $\hat{\lambda}_i$  es el valor medio de una muestra  $\{\lambda_i^{(j)}\}_{j=1}^J$  obtenida de la distribución a posteriori. Los conglomerados pueden detectarse a partir de la probabilidad de que el valor residual supere un determinado valor, por ejemplo  $Pr(r_i > 2)$  o  $Pr(r_i > 3)$ . Sin embargo, hay que tener en cuenta que los residuales siempre incluirán alguna forma de ruido no relacionado con el conglomerado. Incluso cualquier modelo Poisson siempre tendrá ruido alrededor de la media  $e_i \lambda_i$  (Hossain and Lawson, 2006).

También es posible abordar la detección de conglomerados estudiando la significancia del riesgo  $\hat{\lambda}_i$  mediante las probabilidades de excedencia, las cuales bajo un análisis bayesiano son estimadas a partir de los valores de la muestra posterior:

$$q_i^c = Pr(\lambda_i > c).$$

Esta probabilidad es una estimación de qué tan frecuente el riesgo relativo excede el valor nulo del riesgo ( $\lambda_i = 1$ ) y puede ser considerada como un indicador

de cómo es de inusual el riesgo en esa área de estudio. Grandes valores de  $\hat{q}_i^c$  pueden sugerir áreas con riesgo inusual.

Uno de los trabajos más relevantes en este tema es el de Richardson et al. (2004), quienes estudian cómo diferentes modelos recobran la verdadera superficie de riesgo. En concreto, los autores enfatizan que la información obtenida de la posterior puede ser mejor estudiada a través de los riesgos relativos estimados y, usando algunas de las reglas de decisión expuestas en este trabajo, discriminan entre áreas mostrando altos riesgos y áreas con riesgos relativos cercanos al nivel de referencia (un nivel de referencia puede ser la tasa de incidencia de la enfermedad en cierta área de estudio). Sin embargo, expresan que existe una gran preocupación relacionada con el uso de probabilidades de excedencia para una única región y es el hecho de que está diseñada solamente para detectar hot-spot clusters y no considera cualquier otra información concerniente a formas posibles de conglomerados o incluso información de vecindad. Además, dichas probabilidades son estimadas desde las colas superiores de la distribución posterior, las cuales son muy sensibles a los supuestos del modelo. Otro problema con las probabilidades de excedencia es que la utilidad de la medida depende del modelo que ha sido fijado para los datos. Así, un modelo pobre puede dejar al conglomerado de interés en el ruido residual.

Algunos trabajos recientes han intentado mejorar estas medidas incluyendo estructuras de vecindad. Por ejemplo, Hossain and Lawson (2006, 2010) estudian diferentes diagnósticos de detección de conglomerados para datos de salud en áreas pequeñas con el fin de evaluar el grado en el que un método de conglomerado recupera el verdadero comportamiento de agrupamiento. En su trabajo proponen además un modelo de riesgo relativo local basado en la relación de un dato a otro dato dentro de una distancia de enlace (distancia lazo) sin especificar centros de conglomerados. Este modelo se basa en estimar una distancia de enlace para cada región dentro de la cual son consideradas regiones para ser agrupadas. El modelo es como sigue:

Sea  $y_i$  el conteo observado en el área  $i$  y  $y_{\delta_i}$  el conteo total observado dentro de la distancia de enlace o distancia lazo. Entonces,  $y_{\delta_i}$  se modeliza a través de una Poisson condicionada en  $e_{\delta_i}$  y el parámetro del modelo  $\theta_{\delta_i}$ :

$$y_{\delta_i} | e_{\delta_i}, \theta_{\delta_i} \sim \text{Po}(e_{\delta_i} \theta_{\delta_i}).$$

En el segundo nivel de jerarquía el logaritmo del riesgo es modelizado como:

$$\log(\theta_{\delta_i}) = \delta_i + \epsilon_i, \tag{1.10}$$

donde  $\delta_i$  es la componente espacial y  $\epsilon_i$  es la componente no estructurada. En este modelo el suavizado de  $\delta_i$  es controlado por una distribución CAR y la fuerza de la correlación espacial es controlada por  $\beta_{\delta}$ :

$$\delta_i | \beta_\delta \propto \beta_{\delta_i}^{-n/2} \exp\left\{\frac{-1}{2\beta_\delta} \sum_{i < j} \sum_{j \in \delta_i} (\delta_i - \delta_j)^2\right\}.$$

Una previa alternativa para  $\delta_i$  es:

$$\delta_i | \beta_\delta \propto \beta_\delta^{-n/2} \exp\left\{\frac{-1}{\beta_\delta} \sum_{i < j} \sum_{j \in \delta_i} \{|\delta_i - \delta_j|\}\right\}.$$

La previa sugerida para la componente no estructurada  $\epsilon_i$  es:

$$\epsilon_i \sim N(0, \sigma_\epsilon^2).$$

Para el análisis y detección de conglomerados usan los residuales y/o estadísticos de dicho modelo de suavizado, que resumen la distribución posterior (usan los residuales porque ellos contienen la información removida de los riesgos extremos después del suavizado) y finalmente chequean la capacidad de cada modelo para recubrir el verdadero componente de conglomerado. Para formar conglomerados asumen qué valores bajos extremos y riesgos altos formarán los conglomerados.

Finalmente, comentar que dentro de los primeros trabajos que intentan detectar grupos de riesgo se encuentran los que pertenecen al enfoque Scan Statistics (Kulldorff, 1997). Cabe resaltar que estos no son modelos sino métodos que han sido ampliamente usados en procesos puntuales y en los cuales se identifican grupos de unidades de áreas que exhiben un elevado riesgo de la enfermedad. En este tipo de métodos se realiza una exploración estadística para probar si un proceso puntual unidimensional es aleatorio o si es posible detectar algún patrón o conglomerados en dicho proceso. Otros trabajos de gran relevancia en esta línea son el de Gangnon and Clayton (2000), Kulldorff et al. (2005) y el trabajo de Wakefield and Kim (2013) quien presenta una versión bayesiana del Scan Statistics de Kulldorff.

### 1.3.2. Modelos de Conglomerados

Notar que los enfoques mencionados anteriormente necesitan de un modelo para la estimación del riesgo relativo. Sin embargo, existen trabajos que intentan modelizar el riesgo de una enfermedad y al tiempo tratan de abordar el objetivo de detectar grupos de áreas pequeñas con riesgo similar.

Dentro de los modelos para la detección de conglomerados podemos hablar de los modelos con procesos ocultos. En ellos, el logaritmo del riesgo es modelizado como sigue:

$$\log(\lambda_i) = \alpha_0 + \alpha_1 \sum_{j=1}^k \phi_j h(C_i - c_j; \tau_h), \quad (1.11)$$

donde  $\alpha_0$  es el intercepto,  $\alpha_1$  es el parámetro de enlace,  $\{\phi_j\}$  el efecto aleatorio,  $C_i$  es el centroide del área  $i$ -ésima,  $\{c_j\}_{j=1}^k$  el conjunto de centros ocultos para ser estimados y  $h(C_i - c_j; \tau_h)$  es una función de distribución de conglomerados; dicha función asocia áreas pequeñas a centros ocultos,  $\tau_h$  es el parámetro asociado a la precisión. Algunas variaciones de este modelo las encontramos en Lawson (2006), donde:

$$\lambda_i = \exp(\phi_j) \left[ 1 + \sum_{j=1}^k h(C_i - c_j; \tau_h) \right],$$

con  $\phi_j \sim N(0, \sigma_\phi^2)$ .

Otra variante aparece en el trabajo de Gangnon (2006). En este caso, el logaritmo del riesgo relativo es estimado como:

$$\log(\lambda_i) = \alpha_0 + \Gamma_i + \epsilon_i,$$

con  $\Gamma_i = \sum_{j=1}^k \phi_j h(C_i - c_j; \tau_h)$ ;  $\epsilon_i$  es el efecto aleatorio al cual se le asigna una previa Normal  $\epsilon \sim N(0, \sigma_\epsilon^2)$ ,  $\phi_j$  es la componente del riesgo relativo asociado con el  $j$ -ésimo conglomerado y  $h(\cdot)$  es la función de pertenencia a un conglomerado asumida para ser uniforme en un área, además asocia áreas pequeñas a centros de conglomerados (que se corresponden con los centroides de las áreas pequeñas).

Otro enfoque distinto lo podemos encontrar en los modelos de partición. Dentro de estos modelos, cabe destacar el trabajo de Knorr-Held and Rasser (2000), donde se asignan regiones a conglomerados atendiendo a criterios claramente geográficos. A través de este modelo podemos divisar grupos de elevado o bajo riesgo, identificar factores de riesgo y detectar discontinuidades. En el primer nivel de jerarquía, los casos observados  $y_i$  en cada una de las áreas pequeñas siguen una distribución Poisson independiente cuando condicionamos al riesgo, con media  $e_i h_j$ , donde  $h_j$  es el riesgo relativo asociado al conglomerado  $j$ , con  $i \in C_j$ . La función de verosimilitud puede ser escrita como:

$$L(y|H_k) = \prod_{j=1}^k \prod_{i \in C_j} \frac{(e_i h_j)^{y_i}}{y_i!} \exp\{-e_i h_j\}, \quad (1.12)$$

donde  $k$  es el número de conglomerados, supuesto desconocido, y el vector de riesgos es  $H_k = (h_1, h_2, \dots, h_k)$ . Para estimar los parámetros del modelo, los autores utilizan un algoritmo MCMC de salto reversible (Lawson and Clark, 1999; Lawson and Denisson, 2002; Gangnon, 2006). El número de conglomerados es restringido para ser  $k \leq m$ , donde  $m$  es el número de áreas pequeñas. También

puede ser restringido de manera que  $k \leq k_{max} \leq m$ , donde  $k_{max}$  es una constante que expresa el número de conglomerados permitidos.

Para asignar las áreas a los conglomerados algunas regiones son marcadas y llamadas centros de conglomerados, de modo que cada una define a un conglomerado. A continuación, cada una de las áreas pequeñas es asignada al conglomerado más próximo. La distancia entre dos áreas es definida como el mínimo número de límites que tienen que ser cruzados para moverse de un área a otra.

Algunas extensiones del modelo de particiones las podemos encontrar en los trabajos de Denison and Holmes (2001), Denison et al. (2002), Ferreira et al. (2002), entre otros.

Lee and Lawson (2014) proponen un novedoso modelo bayesiano espacio-temporal que puede detectar conglomerados de áreas que exhiben riesgo inusual. El modelo es capaz de detectar conglomerados dinámicamente; es decir, la relación de pertenencia al conglomerado y su nivel de riesgo promedio puede evolucionar en el tiempo.

El modelo de conglomerados es representado por  $(\theta_{1t}, \dots, \theta_{mt})$ , que es el vector de riesgos para las  $m$  unidades de área durante el período de tiempo  $t$ . Existen al menos  $G$  distintos niveles de riesgo o clases  $(\lambda_{t1}, \dots, \lambda_{tG})$ , ordenadas para mitigar el problema de identificabilidad común en modelos de mixturas. La asignación del área  $i$  a una clase de riesgo durante un período de tiempo  $t$  es controlada por el vector de variables indicadoras para el área  $i$  y todos los períodos de tiempo  $T: (Z_{i1}, \dots, Z_{iT})$  donde  $Z_{it} \in 1, \dots, G$ . De esta manera, el logaritmo del riesgo es modelizado como sigue:

$$\log(\theta_{it}) = \lambda_{t, Z_{it}}. \quad (1.13)$$

El término  $G$  es el número máximo de clases, permitiendo que algunas de esas clases pueda estar vacía.  $G$  es fijado en el modelo y los autores implementan una restricción para penalizar modelos con muchas clases de riesgo. Sin embargo,  $G$  podrá ser tratado como un parámetro más del modelo y ser estimado a partir de un algoritmo MCMC de salto reversible.

Las previas sugeridas son las siguientes:

$$\begin{aligned} \lambda_{tj} | \lambda_{t-1, j} &\sim N(\lambda_{t-1, j}, \sigma^2) I_{[\lambda_{t, j-1}, \lambda_{t, j+1}]}(\lambda_{tj}) \\ \lambda_{1j} &\sim U(\lambda_{1, j-1}, \lambda_{1, j+1}) \\ \sigma^2 &\sim \text{Ga}^{-1}(a, b) \end{aligned}$$

con  $j = 1, \dots, G$  y  $t = 2, \dots, T$ .

Las  $Z_{it}$  son modelizadas como temporalmente autocorreladas. Modelos Markov independientes son especificados para cada unidad de área  $i$ . La distribución previa conjunta para  $(Z_1, \dots, Z_m)$  es descompuesta como:

$$f(Z_1, \dots, Z_m) = \prod_{i=1}^m f(Z_i) = \prod_{i=1}^m [f(Z_{i1}) \prod_{t=2}^T f(Z_{it}|Z_{i,t-1})].$$

Cada  $Z_{it} \in \{1, \dots, G\}$ , y en este punto los autores proponen distribuciones discretas  $f(Z_{i1})$  y  $f(Z_{it}|Z_{i,t-1})$  dadas por:

$$\begin{aligned} P(Z_{it}|Z_{i,t-1}) &= \frac{\exp(-\alpha(Z_{it} - Z_{i,t-1})^2 - \beta(Z_{it} - G^*)^2)}{\sum_{r=1}^G \exp(-\alpha(r - Z_{i,t-1})^2 - \beta(r - G^*)^2)} \\ P(Z_{i1}) &= \frac{\exp(-\beta(Z_{i1} - G^*)^2)}{\sum_{r=1}^G \exp(-\beta(r - G^*)^2)} \\ \alpha, \beta &\sim U(0, M) \end{aligned}$$

La autocorrelación temporal en las clases indicadoras es controlada por  $\alpha$  y  $\beta$  controla la fuerza de una restricción que favorece modelos cuyas clases indicadoras están cerca a la clase media  $G^* = \frac{G+1}{2}$ . El término  $\beta(Z_{it} - G^*)^2$  del modelo de Markov está incluido como una restricción para mejorar la convergencia del algoritmo MCMC.

Seguidamente los autores realizan una extensión del modelo de conglomerados para incorporar autocorrelación espacial en los riesgos  $\{\theta_{it}\}_{i=1}^m$  a través de la previa CAR Leroux et al. (1999). Entonces el modelo es como sigue:

$$\begin{aligned} \log(\theta_{it}) &= \lambda_{t,Z_{it}} + \phi_{it} \\ \phi_{it}|\phi_{-it} &\sim N\left(\frac{\rho \sum_{j=1}^N \omega_{ij} \phi_{jt}}{\sum_{j=1}^N \omega_{ij} + 1 - \rho}, \frac{\tau^2}{\rho \sum_{j=1}^N \omega_{ij} + 1 - \rho}\right) \\ \tau^2 &\sim \text{Ga}^{-1}(a, b) \\ \rho &\sim U(0, 1) \end{aligned}$$

Aquí  $\rho$  es un parámetro de autocorrelación espacial,  $\rho = 1$  corresponde a la previa ICAR propuesta por Besag et al. (1991). Si  $\rho = 0$  los efectos aleatorios son independientes con media cero y varianza constante  $\tau^2$ .

Algunas propuestas que también tratan de estimar el riesgo de la enfermedad y al mismo tiempo identificar conglomerados en el ámbito puramente espacial son: Gangnon and Clayton (2000), Knorr-Held and Rasser (2000), Green and Richardson (2002), Forbes et al. (2013), Wakefield and Kim (2013) y Anderson et al. (2014).

## 1.4. Modelos en mapeo de enfermedades que incluyen covariables.

En ocasiones resulta de gran interés incluir información relacionada con una o más covariables en el modelo estadístico. Es posible que, haciendo un ajuste para los efectos de la covariable, se obtengan indicios de algún factor influyente en la estimación del riesgo y que tenga que ver, por ejemplo, con la etiología de la enfermedad. Algunos de los trabajos que estudian la estimación del riesgo incluyendo covariables se explican a continuación. Green and Richardson (2002) presentan un modelo jerárquico bayesiano donde las covariables son incluidas de manera multiplicativa en la media de la distribución de Poisson:

$$y_i \sim \text{Po}(e_i \lambda_i \exp \left\{ \sum_j x_{ij} \gamma_j \right\}), \quad (1.14)$$

donde  $\gamma_j$  es el parámetro de regresión. Para el siguiente nivel de la jerarquía, los autores introducen un modelo de mixtura finita para la modelización del riesgo. Este tipo de modelización fue explicada en la página 9, sección 1.2 de este capítulo.

Hossain and Lawson (2005) ilustran dos formas para incluir covariables en el modelo: incluirlas en el nivel principal de la jerarquía o incorporar la dependencia de las covariables en la escala de conglomerados en el modelo.

Cuando se incorpora el efecto de las covariables en el nivel principal, el conteo observado es modelizado como:

$$y_{\delta_i} | e_{\delta_i}, \theta_{\delta_i} \sim \text{Po}(e_{\delta_i} \theta_{\delta_i}),$$

con

$$\log(\theta_{\delta_i}) = x_i \alpha + \delta_i + \xi_i, \quad (1.15)$$

donde  $x_i$  es el vector de covariables para la  $i$ -ésima área y  $\alpha$  es un vector de parámetros. Las componentes aleatorias son dadas por una componente espacial de conglomerados  $\delta_i$  y una componente incorrelada de ruido  $\xi_i$ .

Cuando la escala de la agrupación de los conglomerados depende directamente de las covariables, los datos son modelizados como:

$$y_{\delta_i^*} | e_{\delta_i^*}, \theta_{\delta_i^*} \sim \text{Po}(e_{\delta_i^*} \theta_{\delta_i^*}),$$

donde

$$\log(\theta_{\delta_i^*}) = \delta_i^* + \xi_i, \quad (1.16)$$

con  $\delta_i^* = f_i\zeta + \delta_i$ . El vector  $f_i$  es el de covariables en escala de conglomerados y  $\zeta$  vector de parámetros. Esos dos escenarios alternativos para las covariables llevan a muchas formas diferentes. El segundo caso provee un mecanismo para dependencia entre la escala de conglomerados y covariables.

También presentan una extensión para este modelo donde ambos tipos de ajuste de covariables pueden ser incluidos:

$$\begin{aligned} \log(\theta_{\delta_i^*}) &= x_i\alpha + \delta_i^* + \xi_i \\ \delta_i^* &= f_i\zeta + \delta_i \end{aligned}$$

Aquí el vector  $x_i$  y el vector de parámetros  $\alpha$  contiene la información principal de los niveles de las covariables y el vector  $f_i$  y el vector de parámetros  $\zeta$  representa las covariables a nivel de escala.

Hossain and Lawson (2010) hacen la siguiente propuesta para incluir covariables:

$$f(y_{it}|\lambda_t, \beta, e_{it}, X_{it}) = \text{Po}(e_{it} \exp\{\sum_j \beta_j X_{it}\} \lambda_t), \quad (1.17)$$

donde  $e_{it}$  son los valores esperados en el área  $i$  y período de tiempo  $t$ ,  $\beta_j$  vector de coeficientes de las covariables,  $X_{it}$  corresponde a la información de la  $j$ -ésima covariable en el área  $i$  y tiempo  $t$ , y  $\lambda_t$  el riesgo en cada una de las  $L$  componentes.

Lawson (2010) modeliza el riesgo relativo mediante tres componentes: un predictor el cual es una función de covariables fijas, una función de efectos aleatorios y un término de mixtura que describe cómo regiones se relacionan a diferentes componentes temporales. Esto puede ser definido de manera más general como:

$$\log(\lambda_{it}) = x'_{it}\beta_t + z'_{it}\gamma + \Gamma_{it}, \quad (1.18)$$

donde  $x'_{it}$  es un vector de valores de la covariable para la  $i$ -ésima área en el  $t$ -ésimo período de tiempo,  $\beta_t$  es el vector de parámetros indexado a un período de tiempo,  $z'_{it}$  es un vector de valores de efectos aleatorios para la  $i$ -ésima área y el  $t$ -ésimo período de tiempo,  $\gamma$  es un vector unitario y  $\Gamma_{it}$  es la componente de

mixtura indexada a la  $i$ -ésima área y al  $t$ -ésimo período de tiempo. En general, modelos convencionales de efectos aleatorios espacio-temporales asumen las dos primeras de esas componentes para permitir efectos de las covariables y efectos aleatorios.

Lee and Lawson (2014) comentan que en común con la literatura de mapeo de enfermedades, su modelo de conglomerados no incluye covariables, así que el riesgo de las clases relaciona al riesgo en un área y no al riesgo residual después del ajuste por covariables. Esto les permite clasificar grupos de áreas en elevado riesgo y así las intervenciones en salud pública pueden ser dirigidas apropiadamente. Sin embargo el clasificar la superficie de riesgo residual después del ajuste por covariables permite lograr una identificación de covariables desconocidas de tipo etiológico. Además, la estructura residual de las clases puede ser examinada para revisar similaridad con covariables que varían espacialmente. La información de las covariables puede también ser usada como predictores para el modelo de asignación de conglomerados y comentan que ambas propuestas serán investigadas en futuros trabajos.

Anderson et al. (2014) proponen un modelo similar a los anteriores para incluir el efecto de las covariables:

$$y_i \sim \text{Po}(e_i \lambda_i) \quad (1.19)$$

$$\log(\lambda_i) = x_i' \beta + \phi_i \quad (1.20)$$

con  $i = 1, \dots, m$ . Aquí  $\lambda_i$  representa el riesgo de la enfermedad en el área  $i$  y es modelizado mediante un vector de covariables  $x_i' = (1, x_{i1}, \dots, x_{ip})$  con coeficientes  $\beta = (\beta_0, \dots, \beta_p)$ , y un efecto aleatorio  $\phi_i$ . Dicho efecto aleatorio  $\phi = (\phi_1, \dots, \phi_m)$  cuantifica la autocorrelación espacial no explicada e inducida dentro de los datos de la enfermedad. En concreto,  $\phi$  es modelizado mediante una previa condicional autoregresiva.

## 1.5. Objetivos de la tesis

El objetivo principal de esta tesis es el estudio de la variación geográfica del riesgo de una cierta enfermedad en una determinada zona de estudio. Para ello, proponemos una nueva modelización del riesgo basada en conglomerados con estructuras de riesgos latentes a través de la cual podemos asignar áreas a clases sin necesidad de cumplir restricciones de vecindad.

La estructura de los siguientes capítulos es como sigue: en el Capítulo 2 presentamos el modelo de conglomerados con estructura de clases latentes que ha sido desarrollado y mostramos el análisis bayesiano del mismo. Para el estudio de la distribución conjunta a posteriori hemos implementado métodos MCMC debido a que el estudio analítico de dicha distribución es intratable.

En el Capítulo 3 presentamos el estudio de simulación elaborado para estudiar el rendimiento del modelo propuesto. Mostramos los diferentes escenarios estudiados así como los principales resultados obtenidos. Algunas medidas de ajuste del modelo son mencionadas y finalizamos este capítulo con la aplicación de la propuesta en datos reales correspondientes a la incidencia de la varicela en la ciudad de Valencia en los años 2008 y 2013.

En el Capítulo 4 proponemos una extensión del modelo presentado en el Capítulo 2 incluyendo covariables. En este capítulo mostramos el desarrollo teórico del modelo, su análisis bayesiano y algunas aplicaciones a datos reales.

Por último, en el Capítulo 5 exponemos a modo de resumen las conclusiones generales a las que hemos llegado con los modelos propuestos y mencionamos algunas líneas futuras de investigación.

## Capítulo 2

# Un modelo de conglomerados con estructura de clases latentes para el análisis de mapas de enfermedades

En este capítulo presentamos un modelo de conglomerados con estructura de clases latentes que nos permitirá estimar el riesgo de una cierta enfermedad en una región de estudio dividida en áreas pequeñas, así como identificar áreas que tienen un riesgo similar.

Recientemente, varios autores han mostrado interés en el desarrollo de modelos espaciales para estimar el riesgo de la enfermedad y simultáneamente identificar grupos de riesgo que forman conglomerados, entre estos encontramos el de Lee and Lawson (2014); quienes proponen un nuevo enfoque de modelización jerárquica bayesiana para estimar simultáneamente riesgos de enfermedad e identificar conglomerados de alto riesgo en el ámbito espacio-temporal. Este modelo es capaz de detectar conglomerados dinámicamente, estudiando la relación de pertenencia a un conglomerado y cómo su nivel de riesgo promedio evoluciona en el tiempo.

El modelo que presentamos en esta tesis no requiere definir desde el inicio la dependencia o distancia entre vecinos, sino que expone una formulación donde variables de asignación de los riesgos permiten capturar diferentes estructuras de riesgo. Así pues, es un enfoque alternativo donde los riesgos relativos de las áreas pequeñas son asignados a riesgos subyacentes. Nuestra propuesta aplica ideas de modelos de mixturas, detección de conglomerados y de modelos con estructura

latente (Knorr-Held and Rasser, 2000; Green and Richardson, 2002; Lee and Lawson, 2014).

Como detallamos en la siguiente sección, el vector de datos observados son los conteos de la enfermedad en cada una de las áreas. Los parámetros del modelo son los riesgos de las clases o componentes, el número de clases, componentes o conglomerados en los cuales se agruparán las áreas pequeñas según su riesgo y las probabilidades de pertenecer a cada componente. Un aspecto clave en este tipo de modelos es el número de clases o riesgos latentes a incluir. En principio hacemos el análisis del modelo bajo el supuesto de que el número de clases o conglomerados es conocido, y más adelante consideramos el número de conglomerados como otro parámetro desconocido.

## 2.1. Un modelo de mixturas de Poisson

La asignación aleatoria de objetos dentro de clases o componentes es de fundamental interés en estadística y otras áreas que interactúan con esta. Una forma común para asignar una distribución capaz de identificar y separar las diferentes características dentro de grupos homogéneos es a través de los modelos de mixturas.

Supongamos que la zona de estudio tiene  $m$  áreas pequeñas. El vector de datos observados es  $y = (y_1, \dots, y_m)'$ , donde  $y_i$  representa el número de casos observados en el área  $i$ . Asumimos que los conteos observados en cada área,  $y_i$ , provienen de una mixtura de distribuciones Poisson en la que cada una de sus  $k$  componentes tiene media dada por  $e_i \eta_j$  con  $j = 1, \dots, k$ . Donde  $e_i$  es el número de casos esperados en la  $i$ -ésima área bajo el modelo en el que todas las áreas tienen igual riesgo de contraer cierta enfermedad. El vector  $e = (e_1, \dots, e_m)'$  siempre se considera conocido. Además  $\eta_j$  representa el riesgo desconocido de la clase  $j$ -ésima.

Proponemos un modelo de mixtura de Poissons en el cual se tiene un número de clases  $k$  conocido y cada una de ellas tiene asociado un riesgo subyacente, además la probabilidad de pertenecer a cada una de las diferentes clases existentes en el modelo viene dada por el vector desconocido  $p = (p_1, \dots, p_k)'$  con  $\sum_{j=1}^k p_j = 1$  y  $p_j \geq 0$ .

Como hemos mencionado anteriormente, en principio consideraremos que  $k$  es una constante conocida, en un apartado posterior ya consideraremos a  $k$  como otro parámetro desconocido.

De este modo el modelo es como sigue:

$$f(y_i|k, \eta, p, e_i) = \sum_{j=1}^k \text{Po}(y_i|e_i\eta_j)p_j,$$

con  $i = 1, \dots, m$ . También supondremos que, dados los parámetros desconocidos, las observaciones son independientes por lo que:

$$f(y|k, \eta, p, e) = \prod_{i=1}^m \sum_{j=1}^k \text{Po}(y_i|e_i\eta_j)p_j. \quad (2.1)$$

El análisis estadístico de este modelo resulta complejo, debido a la explosión del número de sumandos en la función de verosimilitud.

Como alternativa proponemos un modelo con estructura latente a través de vectores de variables latentes (o variables de asignación no observadas)  $z_i = (z_{i1}, \dots, z_{ik})'$  con  $i = 1, \dots, m$ , donde  $z_{ij} = 1$  identifica la clase  $j$  a la cual pertenece cada área  $i$ , es decir:

$$z_{ij} = \begin{cases} 1, & \text{si el área } i \text{ pertenece a la componente } j \\ 0, & \text{de otro modo} \end{cases}$$

pues con esa definición se cumple que  $\sum_{j=1}^k z_{ij} = 1$ , para cada  $i = 1, \dots, m$ .

Ahora nuestro modelo incluye una estructura latente y lo podemos expresar así:

$$y_i|k, z_i, \eta, e_i \sim \text{Po}(y_i|e_i z_i' \eta) \quad (2.2)$$

$$z_i|p \sim \text{Mult}_k(z_i|n = 1, p) \quad (2.3)$$

Bajo completa incertidumbre sobre la estructura de las clases, la elección estándar para la distribución de  $z_i$  es la Multinomial ( $\text{Mult}_k$ ). Seguimos considerando que los vectores  $(y_i', z_i')'$ , con  $i = 1, \dots, m$ , son independientes dados los parámetros. Así, la distribución conjunta de los datos observados y las variables latentes es expresada ahora como una productoria de Poissons y Multinomiales, lo cual elimina el problema de la explosión del número de sumandos en la función de verosimilitud:

$$f(y, z|k, \eta, p, e) = \prod_{i=1}^m \text{Po}(y_i|e_i z_i' \eta) z_i' p.$$

Esta propuesta muestra una forma alternativa para escribir el modelo (2.1), pues aquí estudiamos a qué clase pertenece cada una de las áreas pequeñas. La ventaja es que esta nueva estructura que presentamos en las expresiones (2.2) y (2.3) muestra una explícita caracterización de la distribución de la relación de pertenencia de un área pequeña a una clase o componente. Obsérvese que si marginalizamos, integrando respecto a  $z$ , obtenemos la distribución dada en la ecuación (2.1), que es la que deseamos estudiar. Tener en cuenta que nos encontramos frente a un modelo de mixturas en los cuales es habitual encontrar problemas de identificabilidad de las clases. Es un problema relacionado con la habilidad de distinguir parámetros dentro de un modelo paramétrico (Lawson, 2009; Bernardo and Smith, 1994). En nuestro caso concreto, está relacionado con la habilidad para distinguir clases o componentes a las cuales serán asignadas cada una de las áreas pequeñas de la zona de estudio en cuestión.

A continuación mostramos el análisis bayesiano del modelo propuesto anteriormente.

### 2.1.1. Análisis bayesiano del modelo: distribuciones previas y a posteriori

Los riesgos  $\eta$  son valores que se encuentran en la semirecta real positiva, es por eso que como distribución previa para  $\eta$  podemos pensar en una Normal para el logaritmo de dichos riesgos. Sin embargo, al realizar exploraciones con esta previa, encontramos problemas de identificabilidad de las clases. Para ilustrar dicho problema presentamos un ejemplo obtenido como resultado durante el desarrollo de esta tesis.

En el modelo estudiado hemos denotado los logaritmos de los riesgos de las clases por  $\eta^* = \log(\eta)$ , como distribución previa asignamos una Normal  $(0, \tau)$  cuya desviación la asumimos con una distribución inicial Uniforme  $(0, 10)$ . Dicho modelo fue lanzado con la ayuda del software WinBUGS, el cual es utilizado para muestrear de la distribución posterior de modelos estadísticos mediante MCMC. La Tabla 2.1 muestra el código implementado en WinBUGS.

A través de las Figuras 2.1 y 2.2 de la cadena de Markov podemos observar el problema de identificabilidad que se presenta. En la Figura 2.1 para  $k = 3$  observamos que el primer valor, correspondiente al logaritmo del riesgo de la primera clase, está en torno a 1.3; dicho valor se mantiene durante todo el proceso. Sin embargo, para el segundo y tercer riesgo podemos observar que las etiquetas se cambian durante el proceso: hay una componente que va a -2 y otra que va a 0 pero a partir de la iteración 3000 se cambia. Existe un problema de identificabilidad. Dicho comportamiento ocurre a la vez, es decir en la misma iteración, tanto en riesgos como en probabilidades (ver Figura 2.2).

Teniendo en cuenta lo anterior, proponemos una distribución previa sobre

```

model
{
  for(i in 1:m){
    y[i] ~ dpois(meany[i])
    meany[i] ← e[i]*theta[i]
    theta[i] ← exp(lambda[i])
    lambda[i] ← inprod(z[i,1:k],etha[1:k])
    z[i,1:k] ~ dmulti(p[1:k],1)
    for(j in 1:k){
      etha[j] ~ dnorm(0,tauetha)
      tauetha ← pow(sdetha,-2)
      sdetha ~ dunif(0,10)
      p[1:k] ~ ddirch(alpha[1:k])
      for(j in 1:k){
        alpha[j] ← 1/k
      }
    }
  }
}

```

Tabla 2.1: Código WinBUGS del modelo

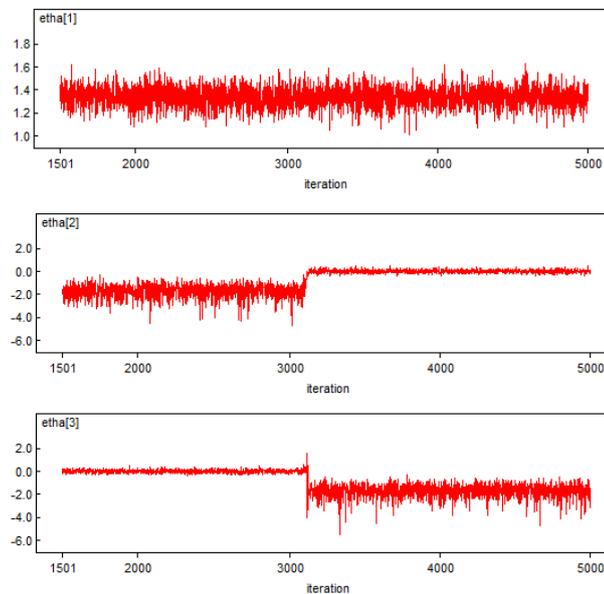


Figura 2.1: Trazas de las series para los logaritmos de los riesgos en cada clase.

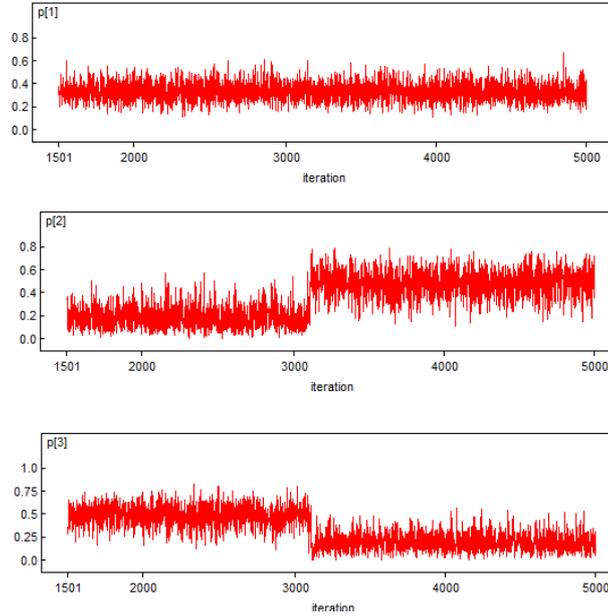


Figura 2.2: Trazas de las series para las probabilidades en cada clase.

el parámetro  $\eta$  que ordene los riesgos. Para ello estableceremos unos puntos de corte, de manera que cada uno de los riesgos  $\eta_j$  oscilará entre un punto de corte y otro.

Nótese que los riesgos se encuentran entre 0 y  $+\infty$ , intervalo en el cual resultaría complicado establecer puntos de corte y por ende definir una distribución de probabilidad para ellos. Nuestra propuesta consiste en trabajar con la siguiente transformación:

$$\eta_j^* = \frac{\eta_j}{1 + \eta_j}, \text{ con } j = 1, \dots, k,$$

que transforma el intervalo  $[0, +\infty)$  en el intervalo  $[0, 1]$ .

El vector  $v$  es el vector de todos los puntos de corte  $v = (v_0, v_1, \dots, v_{k-1}, v_k)$ , con  $v_0 = 0$  y  $v_k = 1$ . Teniendo en cuenta que las componentes del vector  $v$  representan los puntos de corte entre cada dos valores consecutivos de los riesgos transformados, tenemos una ordenación creciente en los componentes del vector  $\eta^*$  y de esta manera conseguiremos identificabilidad de las clases. Una vez fijados los puntos de corte, proponemos como previa para  $\eta^*$ :

$$f(\eta^*|v) = \prod_{j=1}^k f(\eta_j^*|v) = \prod_{j=1}^k \text{Un}(v_{j-1}, v_j).$$

Para obtener la previa  $f(\eta|v)$ , utilizaremos la siguiente proposición:

**Proposición 1.** Sea  $x \sim U(a, b)$ , con  $0 \leq a < b \leq 1$ . Y sea  $y$  tal que  $x = \frac{y}{1+y}$  entonces,

$$f(y|a, b) = \frac{1}{b-a} \left( \frac{1}{1+y} \right)^2, \text{ si } \frac{a}{1-a} < y < \frac{b}{1-b}$$

y 0 de otro modo.

*Demostración.* Si expresamos a  $x$  como:

$$x = \frac{y}{1+y}, \text{ entonces } y = \frac{x}{1-x} = g(x)$$

y,

$$\frac{dg^{-1}(y)}{dy} = \frac{1}{(1+y)^2}$$

Entonces,

$$f(y|a, b) = \frac{1}{b-a} \frac{1}{(1+y)^2} \text{ si } \frac{a}{1-a} < y < \frac{b}{1-b}$$

■

De donde se deduce el siguiente corolario.

**Corolario 1.** *Distribución inicial de los riesgos  $\eta$*

$$f(\eta|v) = \prod_{j=1}^k \frac{1}{(v_j - v_{j-1})} \frac{1}{(1 + \eta_j)^2} \mathbf{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j)$$

Donde  $\mathbf{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j)$  es la función indicatriz o indicadora, que indica la pertenencia o no de cada  $\eta_j$  al intervalo  $\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)$ , de modo que asignará el valor de 1 a todos los elementos que pertenezcan a dicho intervalo y 0 en caso contrario.

Para asignarle una previa a  $v$  nos basamos en la distancia que existe entre un punto de corte y el siguiente, es decir  $d_j = v_j - v_{j-1}$ . El vector de distancias  $d$  en términos de la transformación hecha para el análisis de la previa del riesgo de las clases son no negativas y suman 1, así la distribución previa natural para  $d$  es una Dirichlet.

$$d \sim \text{Dir}(d|\gamma)$$

con  $\gamma = (\gamma_1, \dots, \gamma_k)$  conocido y tal que  $\gamma_j > 0, \forall j$ . Nuestra primera propuesta es  $\gamma_1 = \dots = \gamma_k = 1$ .

Teniendo en cuenta lo anterior, la distribución inicial de  $v$  viene dada por:

**Proposición 2.** *Distribución inicial de los puntos de corte  $v$*

$$f(v) \propto \prod_{j=1}^k (v_j - v_{j-1})^{\gamma_j - 1}$$

si  $v_0 \leq v_1 \leq \dots \leq v_{k-1} \leq v_k$ , con  $v_0 = 0$  y  $v_k = 1$ .

*Demostración.* Realizando un adecuado cambio de variable de los puntos de corte en términos de la distancia.

$$\text{Sea } v_i = \sum_{j=1}^i d_j, i = 1, \dots, k-1.$$

Luego,  $v = Md$ , con

$$M = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 0 \\ & \dots & & \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Así,

$$v = Md = g(d)$$

El Jacobiano de la transformación es  $|M^{-1}| = |M|^{-1} = 1$ , luego:

$$f(v) = f_d(g^{-1}(v))|M| = f_d(M^{-1}v)$$

siendo  $f_d$  la función de densidad sobre  $d$ . Así:

$$f(v) \propto \prod_{j=1}^k (v_j - v_{j-1})^{\gamma_j - 1}$$

si  $v_0 \leq v_1 \leq \dots \leq v_{k-1} \leq v_k$ , con  $v_0 = 0$  y  $v_k = 1$ . ■

Como previa para el parámetro  $p$  proponemos una distribución Dirichlet, que es la previa estándar para vectores no negativos que suman 1, además, por ser la familia conjugada de la Multinomial:

$$p \sim Dir(p|\alpha),$$

con  $\alpha = (\alpha_1, \dots, \alpha_k)$  conocido y tal que  $\alpha_j > 0, \forall j$ . Entonces,

$$f(p) \propto \prod_{j=1}^k p_j^{\alpha_j - 1}.$$

Nuestra primera propuesta para este hiperparámetro fue  $\alpha_1 = \dots = \alpha_k = \frac{1}{k}$ , que coincide con la previa de referencia (Bernardo and Smith, 1994). Pero debido a los resultados empíricos obtenidos durante pruebas en el estudio de simulación, decidimos trabajar con  $\alpha_1 = \dots = \alpha_k = 2$ .

Consideramos que  $\eta$  y  $p$  son independientes a priori por lo que la distribución a posteriori conjunta viene dada por:

$$\begin{aligned} f(z, \eta, p, v, k|y, e) &\propto f(y, z|k, \eta, p, v, e) f(p, \eta, v) \\ &\propto f(y|k, z, \eta, e) f(z|p) f(p) f(\eta|v) f(v) \\ &\propto \left[ \prod_{i=1}^m f(y_i|z_i, \eta, e_i) \right] \left[ \prod_{i=1}^m f(z_i|p) \right] f(p) f(\eta|v) f(v) \\ &\propto \left[ \prod_{i=1}^m (z'_i \eta)^{y_i} \exp\{-e_i z'_i \eta\} \right] \left[ \prod_{i=1}^m z'_i p \right] \\ &\quad \prod_{j=1}^k p_j^{\alpha_j - 1} \\ &\quad \prod_{j=1}^k \frac{1}{(v_j - v_{j-1})} \frac{1}{(1 + \eta_j)^2} \mathbf{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j) \\ &\quad \prod_{j=1}^k (v_j - v_{j-1})^{\gamma_j - 1} \mathbf{I}_{(v_{j-1}, v_{j+1})}(v_j) \end{aligned}$$

Resulta complejo realizar un estudio analítico de dicha distribución posterior por lo que emplearemos el método de Gibbs sampler (Casella, 1992) para muestrear de la distribución conjunta. Para ello necesitamos conocer la distribución condicional completa para cada parámetro. Las distribuciones condicionales completas son las distribuciones de cada parámetro condicionando a la información conocida y todos los otros parámetros. Al tratarse de un modelo latente, las variables latentes (no observables) las trataremos como si fueran parámetros.

También debemos tener en cuenta que si marginalizamos respecto a  $z$  y  $v$  se obtiene la distribución final deseada para el modelo de mixtura:

$$f(\eta, p|y, e, k)$$

Además usando el Gibbs sampler es inmediato obtener una muestra de la distribución final del modelo de mixtura presentado en las ecuaciones (2.2) y (2.3) y, en consecuencia del modelo de mixturas presentado en la expresión (2.1).

### 2.1.2. Distribuciones condicionales completas

La distribución condicional completa del vector de variables latentes de asignación  $z_i$  dado los datos y los demás parámetros se describe a continuación.

**Proposición 3.** *La distribución condicional completa de  $z_i$  es*

$$f(z_i = E_l|k, \eta, p, y_i, e_i) = C \eta_l^{y_i} \exp\{-e_i \eta_l\} p_l, \forall l = 1, \dots, k,$$

siendo  $E_l$  la columna  $l$ -ésima de la matriz identidad  $k$ -dimensional y donde:

$$C = \frac{1}{\sum_{l=1}^k (\eta_l^{y_i} \exp\{-e_i \eta_l\} p_l)}.$$

*Demostración.* El vector aleatorio  $z_i$  sólo puede tomar  $k$  valores posibles,  $z_i \in \{E_1, \dots, E_k\}$ . Las probabilidades de cada uno de esos valores por el teorema de Bayes, son:

$$\begin{aligned} f(z_i = E_l|k, \eta, p, y_i, e_i) &\propto f(y_i|z_i = E_l, p, \eta, e_i) f(z_i = E_l|p) \\ &\propto \eta_l^{y_i} \exp\{-e_i \eta_l\} p_l \end{aligned}$$

con  $l = 1, \dots, k$ . Como se trata de una distribución discreta sobre  $z_i$ , con sólo  $k$  valores posibles ( $E_1, \dots, E_k$ ), su constante de proporcionalidad debe ser tal que las  $k$  probabilidades sumen 1, de donde se obtiene el valor de  $C$ . ■

A continuación se muestra la condicional completa sobre  $p$ , cuya demostración se omite por ser muy simple: la familia Dirichlet es la conjugada de la familia multinomial.

**Proposición 4.** *La distribución condicional completa sobre  $p$  es como sigue:*

$$f(p|k, z, y, e, \eta, v, \alpha) \propto \prod_{j=1}^k p_j^{(\sum_{i=1}^m z_{ij}) + \alpha_j - 1},$$

lo que significa que sigue una distribución Dirichlet  $(p|\alpha + \sum_{i=1}^m z_i)$ .

Ahora revisaremos la condicional completa para los riesgos de cada clase  $\eta_j$ . Para este análisis debemos tener en cuenta que en nuestro modelo existirán tantos riesgos como clases.

**Proposición 5.** *La distribución condicional completa del riesgo en la clase  $j$  es como sigue:*

$$f(\eta_j|k, z, v, y, e) \propto \text{Ga}[\eta_j|y_{(j)} + 1, e_{(j)}] \frac{1}{(1 + \eta_j)^2}$$

si  $\frac{v_{j-1}}{1-v_{j-1}} \leq \eta_j \leq \frac{v_j}{1-v_j}$  y 0 en caso contrario. Siendo  $y_{(j)} = \sum_{i=1}^m y_i z'_i E_j$  y

$$e_{(j)} = \sum_{i=1}^m e_i z'_i E_j.$$

*Demostración.*

$$\begin{aligned} f(\eta_j|k, z, v, y, e, \eta_{-j}) &\propto \prod_{i: z_i = E_j} f(y_i|\eta_j, e_i, z, v) f(\eta_j|v) \\ &\propto (\eta_j)^{y_{(j)}} \exp\{-e_{(j)}\eta_j\} \frac{1}{(1 + \eta_j)^2} I\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)(\eta_j) \\ &\propto \text{Ga}[\eta_j|y_{(j)} + 1, e_{(j)}] \frac{1}{(1 + \eta_j)^2} \end{aligned}$$

si  $\frac{v_{j-1}}{1-v_{j-1}} \leq \eta_j \leq \frac{v_j}{1-v_j}$  y 0 en otro caso. Siendo  $y_{(j)}$  la suma de los datos de las áreas que pertenecen a la clase  $j$  y similarmente  $e_{(j)}$ . ■

En la proposición anterior se supone que la clase  $j$  no está vacía; es decir, algún área pertenece a esa clase, con lo que la distribución Gamma que aparece en la proposición es propia. En otro caso, la distribución final coincidiría con la inicial.

Ahora trabajamos en la estructura de la distribución condicional completa de los puntos de corte  $v = (v_0, v_1, \dots, v_{k-1}, v_k)$  con  $v_0 = 0$  y  $v_k = 1$ .

**Proposición 6.** *La distribución condicional completa sobre cada punto de corte  $v_j$ , con  $j = 1, \dots, k - 1$ , es como sigue:*

$$f(v_j|k, \eta, p, z, v_{-j}, y, e) \propto (v_j - v_{j-1})^{\gamma_j - 2} (v_{j+1} - v_j)^{\gamma_{j+1} - 2}$$

$$\text{si } \frac{\eta_j}{1+\eta_j} < v_j < \frac{\eta_{j+1}}{1+\eta_{j+1}}.$$

*Demostración.*

$$\begin{aligned} f(v_j|k, \eta, p, z, v_{-j}, y, e) &\propto \frac{1}{v_j - v_{j-1}} \mathbf{I}\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)(\eta_j) \\ &\quad \frac{1}{v_{j+1} - v_j} \mathbf{I}\left(\frac{v_j}{1-v_j}, \frac{v_{j+1}}{1-v_{j+1}}\right)(\eta_{j+1}) \\ &\quad (v_j - v_{j-1})^{\gamma_j - 1} (v_{j+1} - v_j)^{\gamma_{j+1} - 1} \mathbf{I}_{(v_{j-1}, v_{j+1})}(v_j) \\ &\propto (v_j - v_{j-1})^{\gamma_j - 2} (v_{j+1} - v_j)^{\gamma_{j+1} - 2} \mathbf{I}\left(\frac{\eta_j}{1+\eta_j}, \frac{\eta_{j+1}}{1+\eta_{j+1}}\right)(v_j) \end{aligned}$$

Nótese que  $v_{j-1} < \frac{\eta_j}{1+\eta_j}$  y  $v_{j+1} > \frac{\eta_{j+1}}{1+\eta_{j+1}}$  por lo que las restricciones  $v_{j-1} \leq v_j \leq v_{j+1}$  ya están incluidas en la función indicatriz anterior. ■

La distribución obtenida en la proposición anterior es una transformación lineal de una Beta truncada. La distribución de  $Y = (v_j - v_{j-1}) / (v_{j+1} - v_{j-1})$  es una Beta de parámetros  $\gamma_j - 1$  y  $\gamma_{j+1} - 1$  pero truncada puesto que  $v_{j-1} < \frac{\eta_j}{1+\eta_j}$  y  $v_{j+1} > \frac{\eta_{j+1}}{1+\eta_{j+1}}$ . Al ser truncada, esta distribución es propia aunque los parámetros de la Beta sean negativos.

### 2.1.3. Muestreo de las distribuciones condicionales completas

El muestreo de las distribuciones condicionales completas de  $p$  y  $z$  es inmediato, pues conocemos la forma analítica de sus distribuciones. Sin embargo, es necesario tener en cuenta que para muestrear de la distribución condicional completa de los riesgos de las clases y de los puntos de corte, no existe una rutina de generación de variables aleatorias en forma automática. Aún así, podemos simular de ella estudiando diferentes casos.

Para muestrear de la distribución condicional completa de cada riesgo  $\eta_j$  hay que distinguir dos situaciones. Si  $e_{(j)} = 0$ , es decir la clase  $j$  está vacía, la distribución final coincide con la inicial. De esa distribución inicial se puede muestrear utilizando la transformada inversa de la función de distribución. La siguiente proposición muestra dicha transformada.

**Proposición 7.** *Sea  $x$  una variable aleatoria con densidad  $f(x|a, b) \propto (1+x)^{-2}$  si  $a \leq x \leq b$ . La transformada inversa de su función de distribución es:*

$$F^{-1}(y|a, b) = \left( \frac{1}{1+a} - \left( \frac{1}{1+a} - \frac{1}{1+b} \right) y \right)^{-1} - 1.$$

*Demostración.* Sea

$$f(x) \propto \frac{1}{(1+x)^2} \text{ si } a \leq x \leq b;$$

pero como

$$\int_a^b \frac{1}{(1+x)^2} dx = -(1+x)^{-1} \Big|_a^b = \frac{1}{1+a} - \frac{1}{1+b},$$

entonces

$$f(x) = \left( \frac{1}{1+a} - \frac{1}{1+b} \right)^{-1} (1+x)^{-2},$$

si  $a \leq x \leq b$ .

Además:

$$F(x) = \int_a^x f(t) dt = \left( \frac{1}{1+a} - \frac{1}{1+b} \right)^{-1} \left( \frac{1}{1+a} - \frac{1}{1+x} \right)$$

si  $a \leq x \leq b$ .

Si ahora hacemos  $y = F(x)$  y despejamos  $x$  se obtiene el resultado de la proposición. ■

La situación habitual, sin embargo, es que  $e_{(j)} > 0$ . En ese caso la condicional completa sobre  $\eta$  viene dada por la Proposición 5:

$$f(\eta_j | k, z, y, e, v, p) \propto \text{Ga}[(\eta_j | y_{(j)} + 1, e_{(j)})] \frac{1}{(1 + \eta_j)^2}$$

si  $\frac{v_{j-1}}{1-v_{j-1}} \leq \eta_j \leq \frac{v_j}{1-v_j}$  y 0 en otro caso.

Además  $y_{(j)} = \sum_{i=1}^m y_i(z_i == E_j)$  y  $e_{(j)} = \sum_{i=1}^n e_i(z_i == E_j)$ . El término  $(z_i == E_j)$  actúa como un operador lógico, donde el sumando sólo aparecerá cuando el área  $i$  vaya a la clase  $j$  adecuada. Con  $E_j$  la columna  $j$ -ésima de la matriz identidad  $k$ -dimensional.

Para muestrear de esta condicional completa haremos uso del método de aceptación y rechazo (von Neumann, 1951; Robert and Casella, 2004). En general el método de aceptación-rechazo se describe así: Deseamos simular de una variable

aleatoria  $X$  cuya densidad es  $f(x)$ ; no sabemos simular directamente de esa distribución pero sí de otra distribución con función de densidad  $g(x)$  que verifica que:

1. Si  $f(x) \neq 0$  entonces  $g(x) \neq 0$ ,
2. Existe una constante conocida  $c$  tal que  $f(x) \leq cg(x)$  para todo  $x$  tal que  $f(x) > 0$ .

Entonces podemos simular  $x$  de  $g(x)$  y  $u$  de una Uniforme en  $(0, 1)$ . Aceptaremos  $x$  si y sólo si  $u < \frac{f(x)}{cg(x)}$ . Repetiremos el proceso hasta que aceptemos el número de valores  $x$  requerido.

Como la condicional completa de cada  $\eta_j$  es de la forma

$$f(x) \propto \text{Ga}(x|a, b) \frac{1}{(1+x)^2} \text{ si } A \leq x \leq B,$$

utilizaremos como función  $g(x)$  la Gamma truncada si el valor de  $B$  es pequeño, tomando como constante  $c = (1+A)^{-2}$ . Pero si  $B$  es grande, el porcentaje de aceptación con ese método es muy bajo: el método no es eficiente. En ese caso, y si el valor de  $a$  es mayor a 2, nótese que:

$$\begin{aligned} f(x) &\propto x^{a-1} \exp\{-bx\}(1+x)^{-2} \\ &\propto x^{(a-2)-1} \exp\{-bx\}x^2(1+x)^{-2} \\ &\propto \text{Ga}(x|a-2, b) \frac{x^2}{(1+x)^2} \end{aligned}$$

si  $A \leq x \leq B$  por lo que podemos utilizar como  $g(x)$  esa nueva Gamma truncada y como constante  $c = B^2/(1+B)^{-2}$ , obteniendo así un porcentaje de aceptación razonable.

Por último, hemos de simular de la distribución condicionada completa de cada punto de corte, que está detallada en la Proposición 6. Para ello simularemos de la Beta truncada  $\text{Be}(\gamma_j^{-1}, \gamma_{j+1}^{-1})$ , comentada tras la demostración de la Proposición 6 y desharemos la transformación lineal. Esto, sin embargo, sólo es automático para las Betas propias; esto es, si  $\gamma_j^{-1} > 0$  y  $\gamma_{j+1}^{-1} > 0$ . Sin embargo, en nuestros resultados numéricos queremos utilizar  $\gamma_j = \gamma_{j+1} = 1$ . En ese caso podemos recurrir al siguiente resultado:

**Proposición 8.** *Sea  $X$  una variable aleatoria con función de densidad*

$$f(x) \propto \frac{1}{(x-a)(b-x)} \text{ si } a < A \leq x \leq B < b.$$

Entonces, su transformada inversa de la función de distribución es:

$$F^{-1}(y) = \frac{a + bk(y)}{1 + k(y)} \text{ siendo } k(y) = \frac{A - a}{b - A} \left[ \frac{(B - a)(b - A)}{(b - B)(A - a)} \right]^y$$

*Demostración.*

$$\begin{aligned} C^{-1} &= \int_A^B \frac{1}{(x - a)(b - x)} dx = \int_A^B \frac{1}{b - a} \left[ \frac{1}{x - a} + \frac{1}{b - x} \right] dx \\ &= \frac{1}{b - a} (\log(x - a) - \log(b - x)) \Big|_A^B \\ &= \frac{1}{b - a} \left( \log \frac{B - a}{b - B} - \log \frac{A - a}{b - A} \right), \end{aligned}$$

y

$$F(x) = \int_A^x C \frac{1}{(x - a)(b - x)} = C \frac{1}{b - a} \left( \log \frac{x - a}{b - x} - \log \frac{A - a}{b - A} \right),$$

con lo que

$$F(x) = y \Rightarrow \log \frac{x - a}{b - x} = (b - a)C^{-1}y + \log \frac{A - a}{b - A} = \log(k(y)).$$

Despejando  $x$

$$(x - a) = (b - x)k(y) \Rightarrow x = \frac{a + bk(y)}{1 + k(y)};$$

luego

$$F^{-1}(y) = \frac{a + bk(y)}{1 + k(y)}.$$

Por lo que:

$$\begin{aligned} k(y) &= \exp \left\{ y \left( \log \frac{B - a}{b - B} - \log \frac{A - a}{b - A} \right) + \log \frac{A - a}{b - A} \right\} \\ &= \left( \frac{(B - a)(b - A)}{(b - B)(A - a)} \right)^y \frac{A - a}{b - A} \end{aligned}$$

■

## 2.2. Situación general: número desconocido de clases

Para trabajar el caso donde el número de clases o componentes  $k$  es desconocido, vamos a suponer un número máximo de clases  $K$  conocido.

Para  $k$  asignamos una previa discreta sobre los enteros:

$$f(k) = q_k,$$

si  $k = 1, \dots, K$ .

La distribución final sobre  $k$  proponemos obtenerla a partir de la verosimilitud marginal de  $k$  conociendo las distribuciones iniciales de los riesgos  $\eta$  y de las probabilidades  $p$ ; es necesario que dichas iniciales sean propias para garantizar la existencia de la integral. Así, partiendo del supuesto de que  $k$  es conocido, la verosimilitud de los datos viene dada por:

$$f(y|k, \eta, p, e) \propto \prod_{i=1}^m \sum_{j=1}^K \eta_j^{y_i} \exp\{-e_i \eta_j\} p_j, \quad (2.4)$$

y, si la distribución inicial de los parámetros de la mixtura es  $f(\eta, p|k)$ , podemos definir la verosimilitud marginal de la siguiente forma.

**Definición 1.** *La verosimilitud marginal de los datos dado  $k$  es:*

$$f(y|k, e) = \int_{\Theta} f(y|k, \eta, p, e) f(\eta, p|k) d\theta, \quad (2.5)$$

siendo  $\theta = (\eta', p)'$ .

El cálculo de la verosimilitud marginal puede aproximarse utilizando integración Monte-Carlo, a partir de una muestra aleatoria  $(\eta^{(i)}, p^{(i)})$  obtenida de la distribución inicial de los parámetros:

$$\hat{f}(y|k, e) = \frac{1}{N} \sum_{i=1}^N f(y|k, \eta^{(i)}, p^{(i)}, e). \quad (2.6)$$

Pero este método de integración puede ser muy poco eficiente si las distribuciones iniciales son poco informativas.

Alternativamente, pueden buscarse métodos Monte-Carlo más eficientes; esto es, métodos que reduzcan el error de la aproximación Monte-Carlo utilizando el

mismo número de puntos  $N$ . El error Monte-Carlo puede estimarse fácilmente sin añadir complicación computacional; así, si  $s_f$  es la desviación típica muestral de los  $N$  sumandos en la ecuación (2.6), el error estándar de la estimación Monte-Carlo es:

$$SE(\hat{f}(y|k, e)) = \frac{s_f}{\sqrt{N}}.$$

Las proposiciones siguientes muestran alternativas a la ecuación (2.6) más eficientes.

**Proposición 9.** *La verosimilitud marginal cuando  $k = 1$  puede expresarse como:*

$$f(y|k = 1, e) = \Gamma(1 + \sum_{i=1}^m y_i) \left( \sum_{i=1}^m e_i \right)^{-(1 + \sum_{i=1}^m y_i)} E((1 + \eta)^{-2}),$$

donde la esperanza se calcula con respecto a  $\eta \sim \text{Ga}(\eta|1 + \sum_{i=1}^m y_i, \sum_{i=1}^m e_i)$ .

*Demostración.* Utilizando la expresión (2.4) y la inicial  $f(\eta)$  comentada en el apartado 2.1.1, la integral en la ecuación (2.5) es:

$$f(y|k = 1, e) = \int_0^\infty \eta^{\sum_{i=1}^m y_i} \exp\{-\eta \sum_{i=1}^m e_i\} \frac{1}{(1 + \eta)^2} d\eta.$$

Nótese que si  $k = 1$  el parámetro  $p$  carece de sentido por lo que el único parámetro del modelo es  $\eta$ .

En la integral anterior, los dos primeros factores son el núcleo de una densidad Gamma. Completando esa densidad:

$$\begin{aligned} f(y|k = 1, e) &= \Gamma(1 + \sum_{i=1}^m y_i) \left( \sum_{i=1}^m e_i \right)^{-(1 + \sum_{i=1}^m y_i)} \\ &\int_0^\infty \frac{1}{(1 + \eta)^2} \text{Ga}(\eta|1 + \sum_{i=1}^m y_i, \sum_{i=1}^m e_i) d\eta. \end{aligned}$$

■

La proposición anterior permite aproximar por Monte-Carlo la verosimilitud marginal para  $k = 1$  mediante:

$$\hat{f}(y|k = 1, e) = \Gamma(1 + \sum_{i=1}^m y_i) \left( \sum_{i=1}^m e_i \right)^{-(1 + \sum_{i=1}^m y_i)} \frac{1}{N} \sum_{i=1}^N \frac{1}{(1 + \eta^{(i)})^2},$$

a partir de una muestra aleatoria  $\{\eta^{(i)}, i = 1, \dots, N\}$  de una distribución  $\text{Ga}(\eta|1 + \sum_{i=1}^m y_i, \sum_{i=1}^m e_i)$ . Esa distribución Gamma, utilizada como función importante en la integración Monte-Carlo de la ecuación (2.5) es una mejor aproximación a la función que se desea integrar que la distribución inicial, por ello el error Monte-Carlo debe ser menor.

En el caso general,  $k > 1$ , proponemos utilizar como distribución importante una aproximación a  $f(\eta, p|k, y, e)$  en la que también vamos a tener en cuenta el hiperparámetro  $d$  asociado a la distribución inicial de  $\eta$ . Esa función importante,  $g(\eta, p, d|k, y, e)$ , la detallamos a continuación para el caso  $k = 2$ . Su generalización a  $k > 2$  es automática. Con ella pretendemos construir una aproximación *naive* a la distribución a posteriori.

Como densidad marginal del hiperparámetro  $d$  vamos a utilizar la inicial propuesta en el apartado 2.1.1:

$$g(d|k = 2, e) = \text{Dir}(d|\gamma_1, \gamma_2) = \text{Be}(d|\gamma_1, \gamma_2).$$

Pero las distribuciones de  $\eta$  y  $p$  vamos a hacerlas depender de  $d$  de la siguiente manera:

$$g(\eta, p, d|k, y, e) = g(\eta|d, k, y, e)g(p|d, k, y, e),$$

suponiendo independencia de los parámetros  $\eta$  y  $p$  dado el hiperparámetro  $d$ .

Una vez observado  $d$ , cada área pequeña se asigna a la primera clase si y sólo si el estimador (tasa de mortalidad estandarizada) de su riesgo,  $\hat{\eta}_i = y_i/e_i$ , es menor que  $d/(1 - d)$ ; esto es, si y sólo si  $y_i/(y_i + e_i) < d$ . Sea  $n_1$  el número de áreas pequeñas que cumplen esa restricción,  $\{\hat{\eta}_{(j)}, j = 1, \dots, n_1\}$  las estimaciones de los riesgos de esas áreas,  $n_2 = m - n_1$  el número de áreas pequeñas asignadas a la segunda clase y  $\{\hat{\eta}_{(j)}, j = n_1 + 1, \dots, m\}$  las estimaciones de los riesgos de esas áreas. Entonces,

$$g(p|d, k, y, e) = \text{Be}(p|\alpha_1 + n_1, \alpha_2 + n_2),$$

siendo  $\alpha_1$  y  $\alpha_2$  los parámetros de la distribución inicial sobre  $p$ .

La distribución  $g(\eta_1|d)$  se construye a partir de la distribución empírica de los  $\{\hat{\eta}_{(j)}, j = 1, \dots, n_1\}$ , de manera que coincida con dicha distribución en cada punto  $\eta_{(j)}$  y tenga una función de densidad constante entre cada dos valores consecutivos de los  $\{\hat{\eta}_{(j)}\}$ .

Por último, la distribución de  $g(\eta_2|d)$  se construye de manera similar.

La introducción de esta función importante  $g(\eta, p, d|k, y, e)$  permite demostrar la siguiente proposición.

**Proposición 10.** *La verosimilitud marginal cuando  $k > 1$  puede expresarse como:*

$$f(y|k, e) = E\left(\frac{f(y|k, \eta, p, e)f(p, \eta|k, d)}{g(\eta, p|d, k, y, e)}\right),$$

tomando esperanzas según la distribución  $g(\eta, p, d|k, y, e)$ .

*Demostración.*

$$\begin{aligned} f(y|k, e) &= \int_{\Theta} f(y|k, \eta, p, e)f(\eta, p|k)d\theta \\ &= \int_{\Theta} f(y|k, \eta, p, e)f(p|k) \int_D f(\eta|k, d)f(d|k)ddd\theta \\ &= \int_{\Theta^*} f(y|k, \eta, p, e)f(p|k)f(\eta|k, d)\frac{g(\eta, p, d|k, y, e)}{g(\eta, p|d, k, y, e)}d\theta^* \end{aligned}$$

siendo  $\theta^* = (\eta', p', d)'$  y las funciones  $g(\cdot)$  definidas más arriba. Nótese que  $f(d|k)$  coincide con  $g(d|k)$ .

Luego:

$$f(y|k, e) = E\left(\frac{f(y|k, \eta, p, e)f(p, \eta|k, d)}{g(\eta, p|d, k, y, e)}\right),$$

donde la esperanza es con respecto a la distribución importante  $g(\eta, p, d|k, y, e)$ . ■

Así pues, proponemos resolver la integral en la ecuación (2.5) por Monte-Carlo utilizando como distribución importante  $g(\eta, p, d|k, y, e)$ , de la que es fácil simular y calcular la densidad en los puntos simulados.

Una vez obtenidos los valores  $\hat{f}(y|k, e)$  para  $k = 1, \dots, K$ , la distribución final de  $k$  puede aproximarse mediante:

$$f(k|y, e) \propto \hat{f}(y|k, e)f(k), \quad (2.7)$$

con  $k = 1, \dots, K$ . En todos los casos utilizaremos el error Monte-Carlo para calcular el tamaño muestral  $N$  de cada integración de manera que dicho error no influya decisivamente en el cálculo de la distribución final de  $k$ .

Por último, la distribución final completa deseada es:

$$f(\eta, p, k|y, e) = f(\eta, p|y, e, k)f(k|y, e).$$

Obsérvese que con el primer factor de la expresión anterior obtenemos la distribución final cuando  $k$  es conocido. Con la expresión completa obtenemos la distribución final cuando  $k$  es desconocido.

Dentro de esta situación general en la que  $k$  es desconocido, uno de los casos a considerar es  $k = 1$ , caso no estudiado en el apartado 2.1 pero necesario cuando  $k$  es desconocido. Bajo este escenario existe una única clase, trabajamos con un único riesgo latente  $\eta$ . El vector de variables de asignación  $z_i$  y el vector de probabilidades  $p$  no hacen parte de este caso porque existe una única componente. Entonces, la distribución condicional completa del riesgo  $\eta$  dado los datos  $y$  y  $e$  (la cual se deduce de la Proposición 5) es:

$$\begin{aligned} f(\eta|k, y, e) &\propto f(y|k, \eta, e)f(\eta) \\ &\propto \left[ \prod_{i=1}^m \exp\{-e_i\eta\} \frac{(e_i\eta)^{y_i}}{y_i!} \right] \frac{1}{(1 + \eta)^2}, \end{aligned}$$

siendo  $0 < \eta < +\infty$ .

Nótese que aunque nos enfrentamos a un habitual modelo de Poisson, la inicial asumida para  $\eta$  es  $\frac{1}{(1 + \eta)^2}$ , de tal manera que dicha expresión sea compatible con la previa asumida para cuando  $k > 1$ .

### 2.3. Conclusiones sobre cada área pequeña

Teniendo en cuenta toda la información que podemos obtener del análisis realizado, nuestro principal interés recae sobre el estudio de los riesgos de cada una de las áreas pequeñas y la probabilidad que tiene cada una de dichas áreas de pertenecer a una de las clases que hacen parte del modelo.

Supuestos conocidos todos los parámetros del modelo y una vez observados los datos, la probabilidad de que el área pequeña  $a_i$  pertenezca a la clase  $c_j$  es:

$$P(a_i \in c_j|k, \eta, p, y, e) = P(z_{ij} = 1|k, \eta, p, y, e), \quad (2.8)$$

y esas probabilidades están dadas en la Proposición 3 y se calculan en cada paso del algoritmo de Gibbs.

Así pues, supuestos conocidos todos los parámetros, el riesgo asociado a cada área pequeña, tomando esperanzas respecto a la distribución de  $z$  es:

$$E(\theta_i|k, \eta, p, y, e) = \sum_{j=1}^k \eta_j P(a_i \in c_j|k, \eta, p, y, e), i = 1, \dots, m, \quad (2.9)$$

donde  $\theta_i$  es el riesgo relativo de cada una de las áreas pequeñas y su estimador bayesiano será la esperanza, respecto de la distribución posterior, de la expresión anterior. La estimación de esos riesgos, su esperanza con respecto a la distribución final dado  $k$ , puede aproximarse utilizando las cadenas de Markov generadas mediante Gibbs Sampling,

$$\hat{E}(\theta_i|k, y, e) = \frac{1}{N} \sum_{n=1}^N \theta_i^{(n)}, \quad (2.10)$$

siendo  $\theta_i^{(n)}$  el valor obtenido al sustituir en la ecuación (2.9) los resultados de la  $n$ -ésima iteración de Gibbs.

Finalmente, utilizando la distribución posterior sobre el número de clases:

$$\hat{E}(\theta_i|y, e) = \sum_{k=1}^K \hat{E}(\theta_i|k, y, e) f(k|y, e). \quad (2.11)$$

La expresión anterior es nuestra propuesta como estimación puntual del riesgo de cada área pequeña.

Además, si suponemos conocido  $k$ , o utilizamos una estimación puntual de dicho parámetro desconocido, podemos estimar la probabilidad de pertenencia de cada área pequeña a cada una de las clases de la siguiente forma:

Sea  $p_{ij}^{(n)}$  el valor obtenido al sustituir en la expresión 2.8 los valores de los parámetros obtenidos en la  $n$ -ésima etapa del algoritmo de Gibbs, entonces:

$$\hat{P}(a_i \in c_j|k, y, e) = \frac{1}{N} \sum_{n=1}^N p_{ij}^{(n)}. \quad (2.12)$$

## 2.4. Un ejemplo numérico

En este capítulo hemos estudiado cómo analizar el modelo de mixturas de Poisson propuesto, utilizando el algoritmo de Gibbs para el cálculo de la distribución posterior dado  $k$  y utilizando verosimilitudes marginales para el cálculo de

la distribución posterior marginal de  $k$ . Seguidamente en esta sección presentamos un ejemplo que hace parte del estudio de simulación que hemos realizado y cuyos detalles aparecen en el Capítulo 3. El objetivo de este ejemplo es doble: Por una parte mostrar cómo analizar un banco de datos. Por otra mostrar la bondad del modelo, pero eso no es posible con un sólo ejemplo, eso lo haremos en el siguiente capítulo. Estudiamos la capacidad de recuperación del modelo y el impacto que se tiene dentro del mismo cuando componentes latentes no observables existen en el proceso.

En este apartado vamos a trabajar con datos simulados, generados para los  $m = 87$  barrios de la ciudad de Valencia (España). El mapa de la Figura 2.3 muestra los riesgos relativos estimados bajo el modelo de conglomerados que proponemos en este capítulo. Los parámetros de la simulación se detallan en la Tabla 2.2. Se simularon datos del número de muertes por cierta enfermedad junto con las tasas de mortalidad esperadas las cuales fueron calculadas teniendo en cuenta el total de la población en cada barrio.

**Riesgos estimados con el modelo de conglomerados**

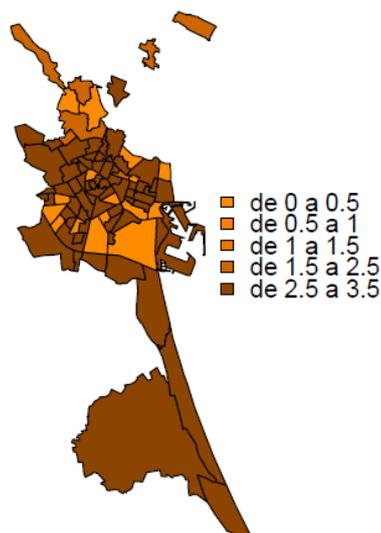


Figura 2.3: Riesgos relativos estimados bajo el modelo de conglomerados para los 87 barrios de la ciudad de Valencia, España.

El escenario estudiado es el siguiente:

Además de la información en la Tabla 2.2, los valores esperados  $e$  son calculados teniendo en cuenta el total de habitantes en cada uno de los 87 barrios que conforman la ciudad. En la Tabla 2.3 se encuentra un resumen numérico de los valores observados  $y_i$  que fueron simulados y que conforman el banco de datos

Tabla 2.2: Parámetros de la simulación con  $k = 2$ 

$\eta_1$	$\eta_2$	$p_1$	$p_2$	Incidencia
0.5	2.5	0.2	0.8	$30/10^4$ hab

que se analiza en este ejemplo.

Tabla 2.3: Resumen de los valores  $y_i$  generados en la simulación

Mín	Q1	Mediana	Media	Q3	Máy
0.00	15.00	35.00	53.55	72.50	192.00

A continuación mostramos el análisis realizado para un banco de datos que hace parte del escenario en estudio, de tal manera que revisemos la recuperación del modelo verdadero cuando  $k$  es conocido. Iniciamos estudiando la convergencia de las cadenas de Markov en la Figuras 2.4 y 2.5, y observamos que las series trazadas de los dos riesgos subyacentes convergen, sin necesidad de tener en cuenta un período de burn-in.

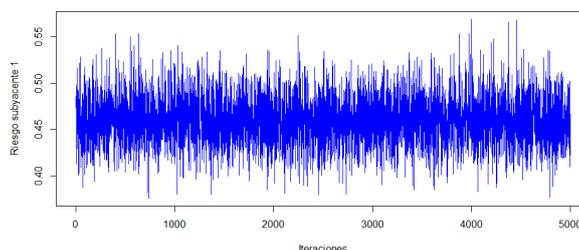


Figura 2.4: Traza del riesgo latente 1 para el banco de datos.

Seguidamente realizamos un análisis de la convergencia de las series relacionadas con cada uno de los riesgos medios estimados para cada una de las 87 áreas pequeñas. Hemos utilizado el diagnóstico de Geweke (1991). Los valores del diagnóstico para los riesgos medios de cada una de las 87 áreas indican que las series convergen. Por ejemplo, para el área 73 presentamos el valor del diagnóstico -1.24 y la Figura 2.6, ambos indican convergencia, mencionamos esta área porque mostró el mayor valor del rango asociado a los valores del riesgo estimado de cada área pequeña y debíamos tener cuidado con el estudio de la convergencia.

Suponiendo  $k$  conocido, estudiamos a continuación en la Tabla 2.4 el resumen de los valores obtenidos para la estimación de los riesgos latentes y de las probabilidades, además de los respectivos intervalos de probabilidad al 95%. Para las estimaciones puntuales de los parámetros del modelo hemos utilizado la mediana y el rango intercuartílico. Los resultados muestran que el modelo recupera bastante bien la estructura real que se ha fijado en el escenario estudiado.

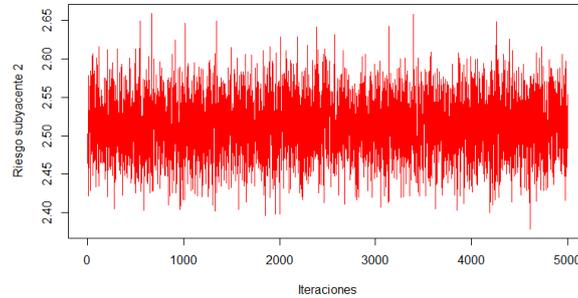


Figura 2.5: Traza del riesgo latente 2 para el banco de datos.

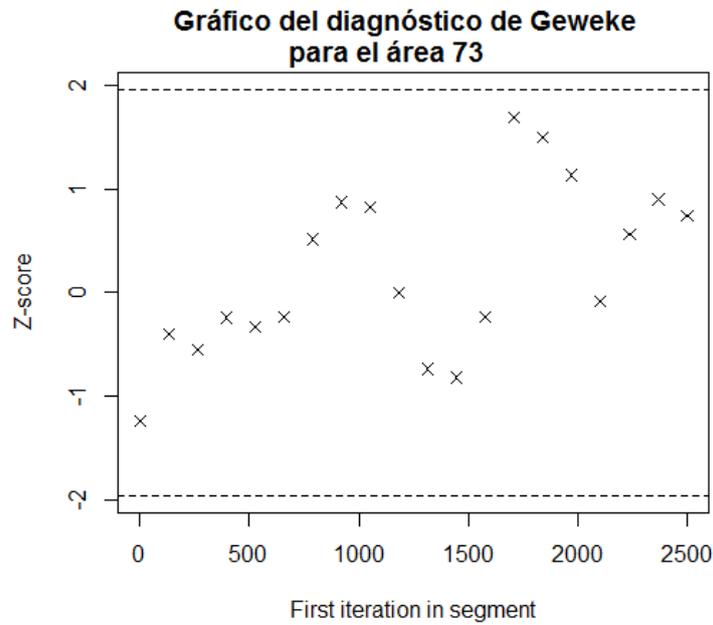


Figura 2.6: Gráfica del diagnóstico de Geweke para el área 73.

Tabla 2.4: Resumen numérico e intervalo de probabilidad al 95 % de los parámetros.

	Mín	Q1	Mediana	Media	Q3	Máx	I.P 95 %
$\hat{\eta}_1$	0.38	0.44	0.46	0.46	0.48	0.57	(0.41,0.51)
$\hat{\eta}_2$	2.38	2.49	2.51	2.51	2.54	2.66	(2.44,2.59)
$\hat{p}_1$	0.08	0.17	0.20	0.20	0.23	0.38	(0.12,0.30)
$\hat{p}_2$	0.62	0.77	0.80	0.80	0.83	0.92	(0.70,0.88)



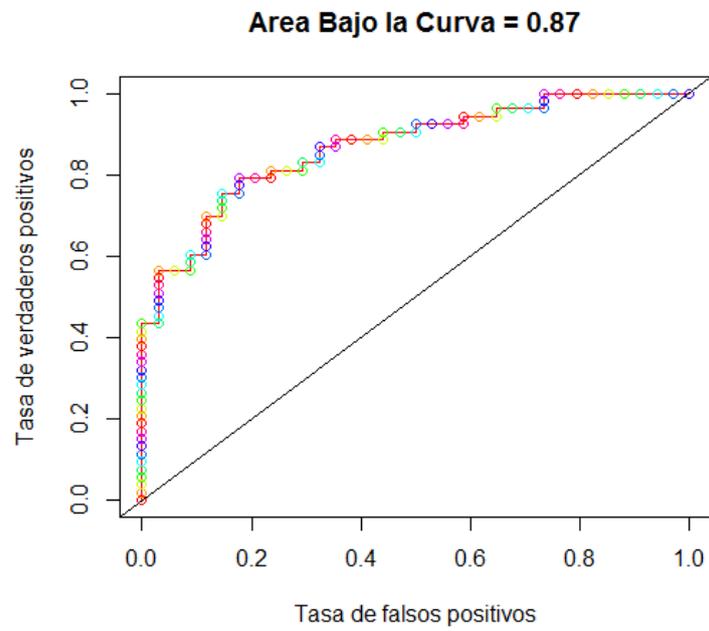


Figura 2.8: Curva ROC para las probabilidades de pertenencia a cada una de las clases con  $\eta = (0.8, 1.2)$ .

probabilidades que tienden a cero, detalles alusivos a este fenómeno se muestran en el Capítulo 3.

Tabla 2.5: Estimaciones de los parámetros con  $k$  real = 2 y  $k$  supuesto = 3

$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
0.34(0.26)	1.08(1.78)	5.84(169.69)	0.11(0.14)	0.20(0.62)	0.69(0.78)

Hasta aquí el análisis suponiendo  $k$  conocido. ¿Qué sucede cuando nos enfrentamos a la situación general donde  $k$  es desconocido? Abordaremos este análisis estudiando las probabilidades finales de  $k$  y luego con las estimaciones del riesgo en cada área pequeña cuando  $k$  es conocido, haremos uso de la mixtura para finalmente obtener las estimaciones de los riesgos en cada una de las áreas pequeñas que es nuestro objetivo principal.

En la Tabla 2.6 encontramos las probabilidades finales de  $k$  para el conjunto de datos estudiado en este escenario. Obsérvese que el valor aquí mayor se corresponde con el  $k$  real correspondiente, que en este caso es igual a 2. Con dichas probabilidades y teniendo en cuenta la estimación del riesgo cuando  $k$  es conocido, es posible estimar el riesgo en cada una de las áreas pequeñas teniendo en cuenta la mixtura, utilizando la expresión (2.11).

Tabla 2.6: Probabilidades finales de  $k$

$k$	1	2	3	4	5	6	7
	0.00	0.97	0.03	0.00	0.00	0.00	0.00

La Tabla 2.7 muestra el resumen de los valores correspondientes a la estimación del riesgo en cada área en el conjunto de datos que estamos estudiando. Notamos estabilidad en los valores estimados de los riesgos en cada área pequeña, además una buena recuperación de la estructura real del modelo.

Tabla 2.7: Resumen de los riesgos relativos estimados de cada área pequeña

Mín	Q1	Mediana	Media	Q3	Máy
0.46	2.51	2.51	2.10	2.51	2.51

Volviendo a la situación en la que suponemos un  $k$  diferente al  $k$  real, podemos observar en la Tabla 2.8 las pequeñas variaciones en el resumen de riesgos estimados en cada área pequeña bajo los diferentes  $k$  supuestos en comparación al  $k$  real. Cada vez que aumentamos el valor de  $k$  va en aumento el rango de variación de los valores del riesgo medio para cada área pequeña. En este escenario y con un valor de  $k$  real 2, resulta no ser tan grave suponer un valor de  $k$  incorrecto.

La Tabla 2.9 muestra el resumen de los riesgos estimados para las áreas pequeñas cuando conocemos el  $k$  y cuando es desconocido. Nótese que a través del

Tabla 2.8: Resumen de los riesgos relativos estimados de cada área pequeña con diferentes  $k$

	Mín	Q1	Mediana	Media	Q3	Máx
$k_{\text{real}}=2$	0.46	2.51	2.51	2.10	2.51	2.51
$k = 3$	0.46	2.50	2.51	2.10	2.51	2.56
$k = 5$	0.46	2.50	2.51	2.10	2.51	2.59
$k = 7$	0.46	2.50	2.50	2.11	2.52	2.65

Tabla 2.9: Resumen de los riesgos relativos de cada área pequeña.  $k$  conocido vs  $k$  desconocido.

	Mín	Q1	Mediana	Media	Q3	Máx
$k$ conocido	0.46	2.51	2.51	2.10	2.51	2.51
$k$ desconocido	0.46	2.51	2.51	2.10	2.51	2.52
Diferencias	0.02	0.09	0.10	0.09	0.10	0.11

modelo propuesto somos capaces de estimar el riesgo haciendo uso del modelo de mixtura garantizando una estabilidad en dichas estimaciones. Incluimos en la última fila de esta tabla las diferencias obtenidas al comparar los riesgos estimados cuando suponemos  $k$  conocido frente al riesgo estimado con el modelo propuesto cuando  $k$  es desconocido, evidentemente los valores indican cercanía y valores estables, siendo capaz el modelo de recuperar la estructura real que se ha fijado.

Los datos simulados son usados para evaluar la precisión del modelo y tienen la ventaja de que la verdadera superficie de riesgo es conocida. Para evaluar el ajuste del modelo hemos utilizado la medida RMSE y comparamos el riesgo real que se fijó para el estudio de simulación en cada escenario con el riesgo estimado con el modelo propuesto, por ejemplo para esta réplica el valor del RMSE corresponde a 0.20.

Tabla 2.10: Medidas de la bondad de los resultados

RMSE <sub>smr</sub>	RMSE <sub>mod</sub>	RMSE <sub>conoc</sub>
0.57	0.20	0.02

Finalmente las medidas de bondad de los resultados se muestran en la Tabla 2.10, aquí encontramos los valores del RMSE calculado cuando estimamos el riesgo basado en la tasa de mortalidad estandarizada, también calculado cuando se compara el riesgo real con la estimación realizada bajo el modelo propuesto y finalmente el RMSE comparado con el riesgo que hemos obtenido asumiendo que conocemos la clase a la cual pertenece cada área. Obsérvese como decrece el error cuando implementamos nuestra propuesta y el valor cuando suponemos conocidas las clases a las cuales pertenecen cada una de las áreas pequeñas.

---

En el siguiente capítulo encontraremos otros ejemplos en los cuales hacemos el análisis para las 100 réplicas generadas para los 246 escenarios, utilizando 5000 pasos en el algoritmo de Gibbs. Además en los Anexos de este trabajo se encuentran todos los resultados correspondientes al estudio de las 100 réplicas para cada uno de los 246 escenarios, los respectivos análisis y las conclusiones. En cada uno de los escenarios que se estudiaron revisamos los riesgos estimados a partir del modelo de mixturas para cada una de las áreas pequeñas, es decir para cada uno de los barrios que conforman la ciudad de Valencia. También hacemos un análisis de las probabilidades finales de  $k$ .



## Capítulo 3

# Resultados numéricos

### 3.1. Estudio de simulación: presentación

En esta sección presentamos un estudio de simulación para explorar el comportamiento del modelo desarrollado en el Capítulo 2. El principal objetivo de este estudio de simulación es investigar la capacidad de recuperación del modelo verdadero. Nosotros conducimos este estudio bajo diferentes escenarios, resultado de suponer distintos valores para el nivel de mortalidad, el número de clases subyacentes y los riesgos relativos y proporción en la que aparecen cada una de esas clases. En cada escenario simulamos 100 bancos de datos para tener en consideración la variabilidad inherente en la simulación. Cada banco de datos lo analizamos primero suponiendo que el número de clases subyacentes es conocido, luego estudiamos el caso general cuando el número de clases es desconocido.

Todo el estudio fue realizado mediante el software estadístico de libre disposición R y el código utilizado se muestra en los Anexos.

#### 3.1.1. Generación de los datos y diseño del estudio

Los datos simulados son generados para los  $m = 87$  barrios de la ciudad de Valencia (España) y hacen referencia al número de muertes por cierta enfermedad. El número de casos esperados en cada área fue calculado teniendo en cuenta la población de dicha área.

En la definición de los escenarios hemos considerado el número de clases subyacentes, desde  $k = 1$  hasta un  $K$  máximo igual a 7, en concreto hemos supuesto  $k = 1, 2, 3, 5$  y 7. Además, los tres valores de la tasa de mortalidad que hemos ido variando:  $r = 3, 30$  y 300 por cada  $10^4$  habitantes, (los cuales se asocian a prevalencias baja, media y alta respectivamente) y con ellos evaluamos el impacto que tienen en el rendimiento del modelo. Hemos variado los valores de los riesgos en

las clases subyacentes (Tabla 3.2) y en ellos, por ejemplo, revisamos escenarios en los que los riesgos subyacentes están distantes y escenarios en los que los riesgos subyacentes son cercanos. Para  $k = 2$  un caso de valores de riesgos subyacentes diferentes es  $\eta_1 = 0.5$  y  $\eta_2 = 2.5$ , mientras que un caso de riesgos similares es  $\eta_1 = 0.8$ ,  $\eta_2 = 1.2$ . Cabe anotar que en el caso de  $k = 1$  sólo definimos un riesgo subyacente  $\eta = 1$ . Con relación a  $p$ , la proporción de áreas en cada una de las  $k$  clases (Tabla 3.3), trabajamos en escenarios donde la proporción de áreas en cada clase son iguales y otros escenarios en los cuales dichas proporciones son diferentes.

Tabla 3.1: Número de escenarios en el estudio de simulación

$k$	1	2	3	5	7	Total
núm.esc	3	48	60	60	75	246

Tabla 3.2: Riesgos subyacentes supuestos

$k = 2$		$k = 3$			$k = 5$				
$\eta_1$	$\eta_2$	$\eta_1$	$\eta_2$	$\eta_3$	$\eta_1$	$\eta_2$	$\eta_3$	$\eta_4$	$\eta_5$
0.5	3.5	0.5	1.0	3.5	0.2	0.7	1.0	3.0	3.5
0.5	2.5	0.5	1.0	2.0	0.5	0.7	1.0	1.5	3.5
0.5	1.5	0.8	1.0	1.5	0.2	0.7	1.0	1.5	2.5
0.8	1.2	1.0	1.5	3.5	0.5	1.0	2.0	2.5	3.5
					0.5	1.0	1.5	2.5	3.5

$k = 7$						
$\eta_1$	$\eta_2$	$\eta_3$	$\eta_4$	$\eta_5$	$\eta_6$	$\eta_7$
0.2	0.5	0.8	1.0	1.5	2.5	3.5
0.5	0.7	0.9	1.0	1.5	2.0	2.5
0.2	0.4	0.6	1.0	1.3	1.5	2.0
0.1	0.4	0.6	1.0	2.7	3.0	3.5
0.5	1.0	1.5	2.0	2.5	3.0	3.5

Para cada número de clases,  $k$ , hemos combinado los distintos valores de riesgos en la Tabla 3.2 con las probabilidades de la Tabla 3.3 y los tres niveles de mortalidad. El número de escenarios resultante para cada valor de  $k$  se muestra en la Tabla 3.1. En total han sido 246 escenarios.

Con el fin de estudiar la variabilidad muestral de cada escenario, generamos 100 conjuntos de datos para cada uno de los 246 escenarios obtenidos. Así pues, vamos a trabajar con un total de 24600 bancos de datos generados teniendo en cuenta una semilla aleatoria.

Para generar los datos observados,  $y_i$ , hemos tenido en cuenta todas las características del escenario del que queríamos simular los:  $r, k, \eta$  y  $p$ . Primero

Tabla 3.3: Probabilidad de cada una de las clases supuestas

$k = 2$		$k = 3$			$k = 5$				
$p_1$	$p_2$	$p_1$	$p_2$	$p_3$	$p_1$	$p_2$	$p_3$	$p_4$	$p_5$
0.5	0.5	1/3	1/3	1/3	0.2	0.2	0.2	0.2	0.2
0.2	0.8	0.2	0.3	0.5	0.1	0.2	0.3	0.3	0.1
0.8	0.2	0.2	0.7	0.1	0.6	0.1	0.1	0.1	0.1
0.1	0.9	0.1	0.1	0.8	0.1	0.1	0.1	0.1	0.6
		0.6	0.2	0.2					
$k = 7$									
		$p_1$	$p_2$	$p_3$	$p_4$	$p_5$	$p_6$	$p_7$	
		1/7	1/7	1/7	1/7	1/7	1/7	1/7	
		0.4	0.1	0.1	0.1	0.1	0.1	0.1	
		0.1	0.1	0.1	0.1	0.1	0.1	0.4	
		0.3	0.2	0.1	0.1	0.1	0.1	0.1	
		0.1	0.1	0.1	0.4	0.1	0.1	0.1	

asignamos las áreas pequeñas a cada una de las clases, simulando de una distribución discreta con probabilidades dadas por el vector  $p$ ; así, para cada área pequeña obtenemos su vector de asignación  $z_i$ , vector de ceros con un único 1 indicando la clase a la que pertenece. De este modo cada área pequeña queda en una única clase y el riesgo que toma es el riesgo subyacente de la clase asignada. Luego los valores esperados  $e_i$  son calculados teniendo en cuenta el tamaño de la población en cada uno de los 87 barrios y variando la incidencia,  $r$ , de tal modo que analizamos cada escenario bajo diferentes volúmenes de información. Con lo mencionado anteriormente, los conteos de la enfermedad  $y_i$  son simulados de un modelo Poisson con tasa esperada  $e_i z_i' \eta$ , donde  $z_i' \eta$  es el riesgo correspondiente a la clase a la que se ha asignado el área pequeña  $i$ -ésima.

### 3.1.2. Análisis estadístico y evaluación de los resultados

Con cada uno de los bancos generados procedemos a la implementación del modelo, el análisis y evaluación de los resultados obtenidos. Primero estudiamos las simulaciones realizadas a través del algoritmo de Gibbs. Cinco mil puntos fueron simulados de la distribución a posteriori mediante el método descrito en detalle en la sección 2.1.1. De aquí obtuvimos las estimaciones de los riesgos de cada una de las áreas pequeñas cuando suponemos  $k$  conocido, expresión (2.10), estudiamos la variabilidad y el error de dichas estimaciones. Luego, nos ubicamos en la situación general, es decir, cuando  $k$  es desconocido. Entonces obtenemos las probabilidades finales de  $k$  estudiadas en el Capítulo 2 haciendo uso de la expresión (2.7), las cuales nos permiten estimar el riesgo relativo en cada una

de las áreas pequeñas haciendo uso de la mixtura, en esta parte del proceso se realizaron 100000 simulaciones para la integración Monte-Carlo.

Resulta entonces interesante estudiar el error de dichas estimaciones comparando esta medida con los valores de los riesgos reales fijados en el estudio. La medida de error que hemos utilizado es la raíz del error cuadrático medio (RMSE):

$$RMSE = \sqrt{\frac{1}{m} \sum_{i=1}^m (\theta_i - \hat{\theta}_i)^2},$$

donde  $\theta_i$  y  $\hat{\theta}_i$  son el riesgo verdadero y el estimado de la  $i$ -ésima área pequeña, respectivamente. Para analizar el error hemos utilizado el RMSE calculado para distintos métodos de estimación. Primero obtenemos el  $RMSE_{smr}$  basado en la tasa de mortalidad estandarizada, la cual hace una estimación del riesgo con mucha variabilidad. Para un segundo análisis calculamos el  $RMSE_{mod}$ , el error basado en la estimación de los riesgos con nuestra propuesta y hacemos la comparación con el riesgo real del modelo. Finalmente suponemos conocida la clase a la cual pertenecen cada una de las áreas pequeñas y basados en esto obtenemos el error que hemos denotado  $RMSE_{conoc}$ . Para el análisis de estos errores hemos utilizado la media y desviación típica de los errores obtenidos de los 100 bancos de datos en cada uno de los escenarios.

En el caso de  $k$  desconocido que es el habitual, también estamos interesados en estimar el número desconocido de clases a partir de su distribución a posteriori marginal. Para ello, hemos realizado un análisis del porcentaje de correspondencia, es decir, hemos calculado el porcentaje de veces que la mayor probabilidad final de  $k$  se corresponde con el  $k$  verdadero, fijado en la simulación. De este modo estudiamos la recuperación de la estructura real del modelo y además analizamos las probabilidades de pertenencia de las áreas pequeñas a cada una de las diferentes clases.

### 3.2. Estudio de simulación: Análisis de algunos escenarios

En esta sección mostramos un resumen de los resultados obtenidos en los distintos escenarios. Todos los resultados del estudio de simulación se encuentran en documentos adjuntos que aparecen en el soporte digital que acompaña a esta memoria.

El análisis de convergencia fue realizado para cada uno de los bancos de datos estudiados. Para ello se trazaron los gráficos de las series y se obtuvo el diagnóstico de Geweke (Geweke, 1991). Las series convergen tanto para los riesgos latentes estimados, probabilidades y riesgos medios en cada área pequeña.

A continuación se describen los resultados de algunos de los escenarios estudiados.

### 3.2.1. Análisis del escenario 1: $k = 1$

En este caso sólo existe una única clase, por ello, nuestro análisis se centra en las estimaciones de los riesgos obtenidas para cada una de las áreas pequeñas. La Tabla 3.4 muestra los valores obtenidos para las probabilidades finales de  $k$ , el mayor valor aquí se corresponde con  $k = 1$ . Y acorde a lo anterior, la Tabla 3.5 muestra que el porcentaje de veces que el mayor valor de la probabilidad se asocia al  $k$  fijado es efectivamente el que corresponde con el  $k = 1$ .

Tabla 3.4: Probabilidades finales de  $k$  en escenario 1

$k$	1	2	3	4	5	6	7
media	0.99	0.01	0.00	0.00	0.00	0.00	0.00
desv. típica	0.02	0.02	0.00	0.00	0.00	0.00	0.00

Tabla 3.5: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  en escenario 1

							$k$
1	2	3	4	5	6	7	
100 %	0 %	0 %	0 %	0 %	0 %	0 %	0 %

En la Tabla 3.6 observamos cómo varía el error calculado bajo diferentes estimaciones del riesgo. Dados los diferentes valores de la incidencia, el error decrece cada vez que aumenta la información. En cuanto al cambio de los valores de los errores dentro de cada uno de los diferentes valores de la incidencia, el cambio es relevante entre el error obtenido con la estimación de la tasa de mortalidad estandarizada y el obtenido con nuestra propuesta. Evidentemente el menor valor del error es el que corresponde a la estimación cuando conocemos a qué clase pertenece cada una de las áreas pequeñas.

Tabla 3.6: Errores de ajuste del modelo bajo diferentes incidencias en escenario 1

Incidencia	RMSE <sub>smr</sub>	RMSE <sub>mod</sub>	RMSE <sub>conoc</sub>
$3/10^4$ hab	1.06(0.64)	0.06(0.04)	0.05(0.04)
$30/10^4$ hab	0.40(0.17)	0.02(0.01)	0.02(0.01)
$300/10^4$ hab	0.13(0.04)	0.01(0.01)	0.01(0.01)

### 3.2.2. Análisis del escenario 2: $k = 3$ y riesgos lejanos

A manera de ejemplo estudiamos el escenario donde  $\eta = (0.5, 1.0, 3.5)$ ,  $p = (0.2, 0.3, 0.5)$  y el valor de la prevalencia de la enfermedad es  $r = 30/10^4$  habitantes. De este escenario hemos analizado los 100 bancos de datos, de tal modo que tengamos una visión general de la simulación. La Tabla 3.7 muestra los valores de la mediana y el rango intercuartílico (con las 100 réplicas) para las estimaciones de  $\eta$  y  $p$  realizadas bajo el modelo propuesto y suponiendo  $k$  conocido. Los resultados obtenidos evidencian que el modelo recupera bastante bien la estructura real del modelo fijado en la simulación.

Tabla 3.7: Mediana (rango intercuartílico) de los valores de  $\eta$  y  $p$  estimados en el escenario 2

$k = 3$		
$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$
0.49(0.05)	1.00(0.05)	3.51(0.09)
$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
0.20 (0.07)	0.30 (0.08)	0.50(0.08)

Hasta aquí hemos supuesto  $k$  conocido, ahora iniciamos el análisis con la situación habitual, es decir, cuando  $k$  es desconocido.

En la situación en la que  $k$  no se conoce analizamos los resultados sobre nuestro objetivo: ¿cuántas clases hay? La Figura 3.1 y la Tabla 3.8 muestran los valores medios y la desviación correspondiente a las probabilidades finales de  $k$  para los 100 bancos de datos estudiados en este escenario. Aquí la probabilidad final con mayor media se corresponde con el  $k$  verdadero del escenario estudiado. Existe cierta probabilidad de que las áreas pequeñas pertenezcan a las clases 4, 5 y 6.

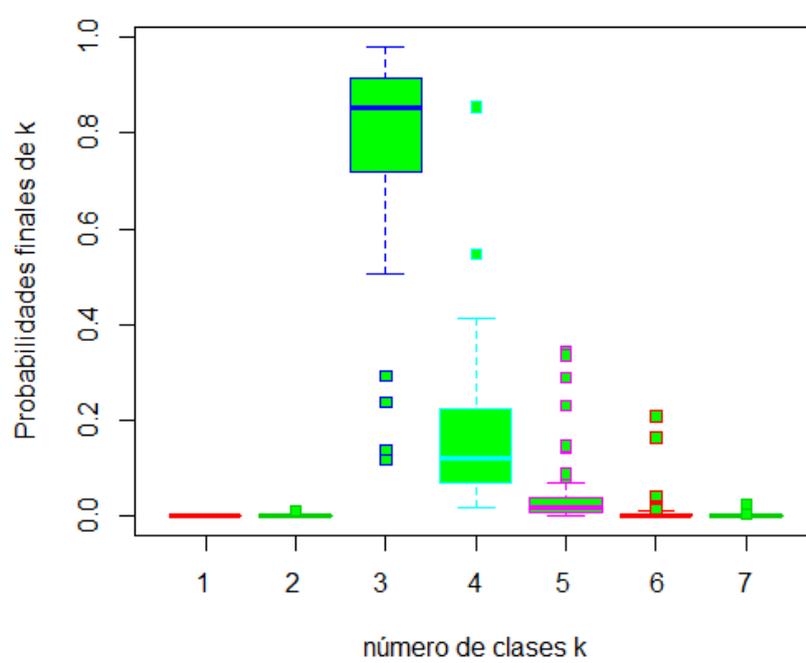
Tabla 3.8: Probabilidades finales de  $k$  en escenario 2

$k$	1	2	3	4	5	6	7
media	0.00	0.00	0.80	0.16	0.04	0.01	0.00
desv. típica	0.00	0.00	0.17	0.12	0.06	0.03	0.00

Tabla 3.9: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  en escenario 2

$k$	1	2	3	4	5	6	7
	0%	0%	96%	3%	1%	0%	0%

La Tabla 3.9 muestra el porcentaje de veces que la probabilidad mayor se asocia con cada  $k$ . Como podemos observar hay una alta correspondencia del

Figura 3.1: Probabilidades finales de  $k$  en escenario 2

valor máximo de la probabilidad con el  $k$  real en el escenario estudiado. Nótese como decrece el error cuando implementamos el modelo propuesto.

El objetivo principal del estudio es estimar el riesgo de las áreas pequeñas. A continuación mostramos el error cometido en la estimación del riesgo relativo de cada una de las áreas pequeñas. En concreto, la Tabla 3.10 muestra las medias y entre paréntesis las desviaciones típicas respectivas, del  $\text{RMSE}_{\text{smr}}$  (con una estimación del riesgo basada en la tasa de mortalidad estandarizada),  $\text{RMSE}_{\text{mod}}$  (con la estimación del riesgo basada en nuestra propuesta) y  $\text{RMSE}_{\text{conoc}}$  (con la estimación del riesgo basada en el caso en que conocemos a qué clase pertenece cada una de las áreas pequeñas).

Tabla 3.10: Errores de ajuste del modelo

$\text{RMSE}_{\text{smr}}$	$\text{RMSE}_{\text{mod}}$	$\text{RMSE}_{\text{conoc}}$
0.56(0.19)	0.26(0.07)	0.04(0.02)

Es interesante revisar cómo se comportan estos errores si variamos la tasa de la enfermedad. Para el análisis que presentamos a continuación mostramos tres escenarios que se derivan del escenario estudiado, lo único que cambiamos es la tasa de mortalidad. Medimos el impacto que tiene cada uno de los diferentes valores de la mortalidad en la recuperación del modelo verdadero y hemos calculado el porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con cada uno de los tres valores propuestos ( $r = 3/10^4\text{hab}$ ,  $r = 30/10^4\text{hab}$  y  $r = 300/10^4\text{hab}$ ). En la Tabla 3.11 observamos que a medida que aumenta la incidencia, obtenemos un número de éxitos mayor asociado con el  $k$  verdadero.

Tabla 3.11: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  variando las incidencias.

	$k$						
Incidencia	1	2	3	4	5	6	7
$3/10^4$ hab	0%	68%	25%	5%	1%	1%	0%
$30/10^4$ hab	0%	0%	96%	3%	1%	0%	0%
$300/10^4$ hab	0%	0%	99%	1%	0%	0%	0%

Al analizar la Tabla 3.12 encontramos interesantes resultados: al comparar el  $\text{RMSE}_{\text{smr}}$  con el  $\text{RMSE}_{\text{mod}}$  se observa como caen los valores del error con el modelo de conglomerados; nuestro modelo es capaz de estimar en forma suavizada el riesgo en cada una de las áreas pequeñas. Como era de esperar, el  $\text{RMSE}_{\text{conoc}}$  es menor que el encontrado con nuestro modelo, pues aquí la tarea de buscar a qué clase es asignada cada área ya está realizada. Nótese que los valores de la incidencia logran tener un efecto sobre las estimaciones y por ende sobre los errores calculados. Obsérvese cómo los errores van disminuyendo a medida que

tenemos más información, es decir, a medida que aumenta la incidencia de la enfermedad.

Tabla 3.12: Errores de ajuste bajo diferentes valores de la incidencia

Incidencia	RMSE <sub>smr</sub>	RMSE <sub>mod</sub>	RMSE <sub>conoc</sub>
$3/10^4$ hab	1.67(1.18)	0.76(0.08)	0.13(0.07)
$30/10^4$ hab	0.56(0.19)	0.26(0.07)	0.04(0.02)
$300/10^4$ hab	0.19(0.07)	0.07(0.06)	0.01(0.01)

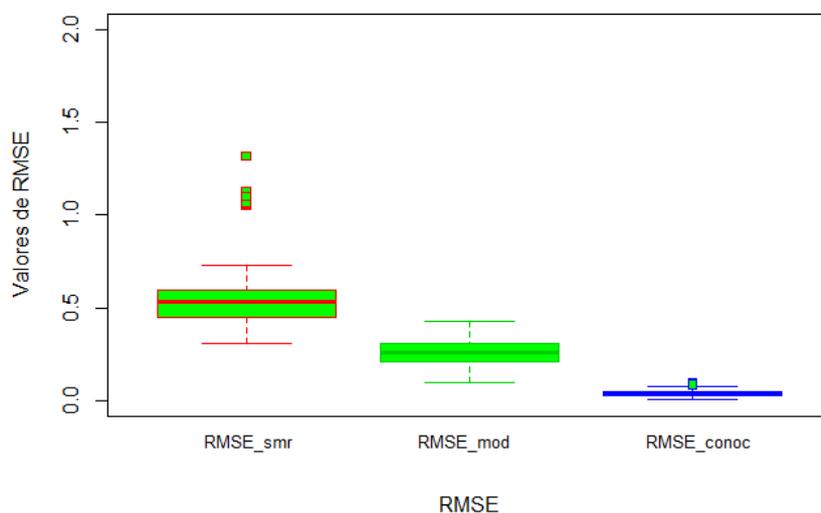


Figura 3.2: Errores de ajuste del modelo en el escenario 2

### 3.2.3. Análisis del escenario 3: $k = 3$ y riesgos parecidos

A continuación mostramos los resultados obtenidos en el escenario donde los riesgos y las probabilidades son:  $\eta = (0.8, 1.0, 1.5)$ ,  $p = (1/3, 1/3, 1/3)$  y el valor de la tasa de mortalidad de la enfermedad es  $r = 30/10^4$  habitantes. Con los riesgos aquí expuestos pretendemos revisar si el modelo propuesto es capaz de separar e identificar clases a pesar de que los riesgos están cerca unos de otros. De este escenario hemos analizado los 100 bancos de datos, de tal modo que revisemos el rendimiento de la simulación. La Tabla 3.13 muestra los valores de la mediana y el rango intercuartílico, obtenidas con las 100 réplicas, para las estimaciones de  $\eta$  y  $p$  realizadas bajo el modelo propuesto y suponiendo  $k$  conocido.

Tabla 3.13: Mediana (rango intercuartílico) de los valores de  $\eta$  y  $p$  estimados en el escenario 3

$k = 3$		
$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$
0.71 ( 0.33)	1.25(0.35)	73.33 (109.26)
$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$
0.44 (0.32)	0.44 (0.10)	0.14 (0.23)

En la Tabla 3.13 se observa que las estimaciones correspondientes a  $\eta_1$  y  $\eta_2$  están muy cercanas a los valores reales de los riesgos fijados en el escenario estudiado. Sin embargo la estimación de  $\eta_3$  se aleja del valor real con una desviación importante. Resulta interesante comentar que, en el estudio que hemos realizado observamos que cuando las estimaciones de los riesgos subyacentes toman valores altos, las probabilidades asociadas a dichos valores son prácticamente cero. Este escenario, con  $k$  conocido, muestra que la situación en la que los riesgos latentes están cercanos unos de otros resulta más complejo distinguir las diferentes clases.

Hasta aquí hemos supuesto  $k$  conocido, ahora nos ubicamos en la situación en que  $k$  es desconocido. Para ello estudiamos la media y la desviación típica de las probabilidades finales de  $k$  para las 100 réplicas, de esta manera también revisamos cuántas clases están ocupadas. Dichos resultados se muestran en la Tabla 3.14 y de estos podemos decir que el porcentaje de ocupación de las diferentes clases está repartido entre las componentes 2,3,4,5 y 6. Sin embargo, el mayor valor medio le corresponde a la clase 2 y le sigue la clase 3, podrían estos valores darnos un indicio de que en este escenario donde los valores de los riesgos latentes fijados se encuentran muy cercanos uno del otro, quizá dos o tres clases serían suficientes para formar grupos con riesgo similar y así estudiar los riesgos.

Tabla 3.14: Probabilidades finales de  $k$  en el escenario 3

$k$	1	2	3	4	5	6	7
media	0.00	0.69	0.23	0.05	0.02	0.01	0.00
desv. típica	0.00	0.27	0.19	0.09	0.08	0.08	0.00

Tabla 3.15: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$ 

$k$	1	2	3	4	5	6	7
	0%	74%	23%	1%	1%	1%	0%

Al analizar las 100 replicas de este escenario en la Tabla 3.15, encontramos que el 74 % de las veces el valor de la probabilidad máxima se corresponde con  $k = 2$  y el 23 % está asociado con el  $k$  verdadero fijado en el estudio.

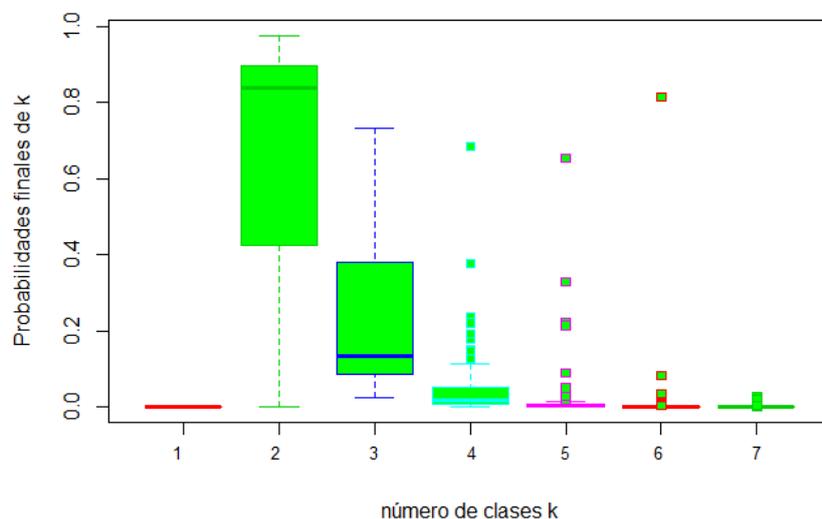


Figura 3.3: Probabilidad final de  $k$  en el escenario 3

En el análisis de este escenario estudiamos las estimaciones de los riesgos medios de cada área pequeña, abordando así el principal objetivo de nuestro trabajo. Inicialmente revisamos la Tabla 3.16 en la cual se muestra cómo varía el error. El error de estimación bajo nuestra propuesta es mucho menor que el error generado por la estimación de la tasa de mortalidad estandarizada. Y el valor del error de estimación cuando se conoce a qué clase pertenece cada área pequeña es menor que el nuestro. En la Figura 3.4 podemos visualizar cómo decrecen los errores.

Tabla 3.16: Errores de ajuste del modelo en el escenario 3

RMSE <sub>smr</sub>	RMSE <sub>mod</sub>	RMSE <sub>conoc</sub>
0.42(0.18)	0.19(0.02)	0.04(0.02)

Para finalizar con el análisis de este escenario revisamos cómo se comportan estos errores si variamos la tasa de mortalidad de la enfermedad. En el análisis que presentamos a continuación mostramos tres escenarios que resultan de cambiar la tasa de mortalidad. Medimos el impacto que tiene cada uno de los diferentes valores de la mortalidad en la recuperación del modelo verdadero y hemos calculado el porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con cada uno de los tres valores propuestos ( $r = 3/10^4$ hab,  $r = 30/10^4$ hab y  $r = 300/10^4$ hab). Al variar la incidencia (Tabla 3.17), se observa que a medida que aumentan dichos valores, el porcentaje de éxitos asociados con el  $k$  verdadero

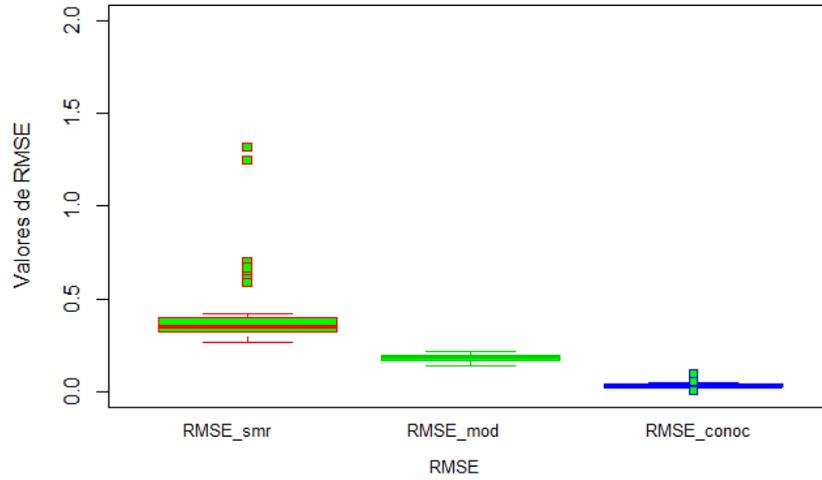


Figura 3.4: Errores de ajuste del modelo en el escenario 3

va en aumento.

Tabla 3.17: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con diferentes incidencias

	$k$						
Incidencia	1	2	3	4	5	6	7
$3/10^4$ hab	0%	76%	17%	3%	3%	1%	0%
$30/10^4$ hab	0%	74%	23%	1%	1%	1%	0%
$300/10^4$ hab	0%	0%	99%	1%	0%	0%	0%

Además resulta interesante comentar el impacto que tiene la variación de la incidencia en un mismo escenario, logrando así que a mayor información suministrada, los diferentes escenarios se van viendo afectados y disminuyen sus respectivos valores (ver Tabla 3.18).

### 3.2.4. Análisis del escenario 4: $k = 5$

Analizamos a continuación un escenario para  $k = 5$ , donde se observa que a medida que aumentamos el número de clases se va complicando la identificación de las mismas. Estudiamos el escenario con  $\eta = (0.5, 1.0, 1.5, 2.5, 3.5)$ ,  $p = (0.1, 0.1, 0.1, 0.1, 0.6)$  y  $r = 30/10^4$  habitantes, los resultados que se presentan a continuación se basan en los 100 bancos de datos generados para el escenario

Tabla 3.18: Errores de ajuste bajo diferentes valores de la incidencia

Incidencia	RMSE <sub>smr</sub>	RMSE <sub>mod</sub>	RMSE <sub>conoc</sub>
3/10 <sup>4</sup> hab	0.97(0.20)	0.88(0.15)	0.11(0.05)
30/10 <sup>4</sup> hab	0.42(0.18)	0.19(0.02)	0.04(0.02)
300/10 <sup>4</sup> hab	0.13(0.04)	0.10(0.05)	0.01(0.01)

en cuestión. En la Tabla 3.19 se encuentran las estimaciones de los riesgos subyacentes y probabilidades cuando consideramos  $k$  conocido. Obsérvese que ahora, cuando tenemos más clases las estimaciones se distancian un poco más de los valores verdaderos del escenario.

Tabla 3.19: Mediana (rango intercuartílico) de los valores de  $\eta$  y  $p$  estimados en el escenario 4

$k = 5$				
$\hat{\eta}_1$	$\hat{\eta}_2$	$\hat{\eta}_3$	$\hat{\eta}_4$	$\hat{\eta}_5$
0.27 (0.14)	0.73 (0.23)	1.41(0.20)	2.49 (0.28)	3.51 (0.11)
$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$
0.04(0.04)	0.12 (0.03)	0.13 (0.06)	0.13 (0.07)	0.59 (0.10)

Tabla 3.20: Probabilidades finales de  $k$  con  $k = 5$  en el escenario 4

$k$	1	2	3	4	5	6	7
media	0.00	0.00	0.05	0.41	0.29	0.18	0.07
desv. típica	0.00	0.00	0.17	0.31	0.24	0.20	0.12

Considerando a  $k$  desconocido estudiamos el número de clases que existen, es decir, el número de clases en la que al menos un área es asignada. En la Tabla 3.20 las probabilidades finales de  $k$  muestran que las áreas están repartidas en las clases 3,4,5,6 y 7. Además si analizamos dichas probabilidades, hay una probabilidad media de 0.29 que se corresponde con el  $k$  real  $k = 5$ , y una probabilidad media de 0.48 con  $k = 4$ . Igualmente en la Tabla 3.21 se muestra que el mayor número de áreas se reparte entre las clases 4 y 5, en concordancia con el escenario estudiado. Finalmente, comentar que este porcentaje de correspondencia se afecta positivamente por la variación en la incidencia, si nos situamos en la Tabla 3.22 observamos que a medida que aumentan los valores de la incidencia, el porcentaje de correspondencia está más acorde con el  $k$  real.

Como nuestro interés recae principalmente sobre la estimación del riesgo en cada una de las áreas pequeñas, continuamos el análisis de este escenario con el estudio del error obtenido a partir de las estimaciones obtenidas para los riesgos de dichas áreas. Obsérvese en las Tablas 3.23 y 3.24 el comportamiento del error

Tabla 3.21: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  en el escenario 4

$k$	1	2	3	4	5	6	7
	0%	0%	6%	48%	29%	15%	3%

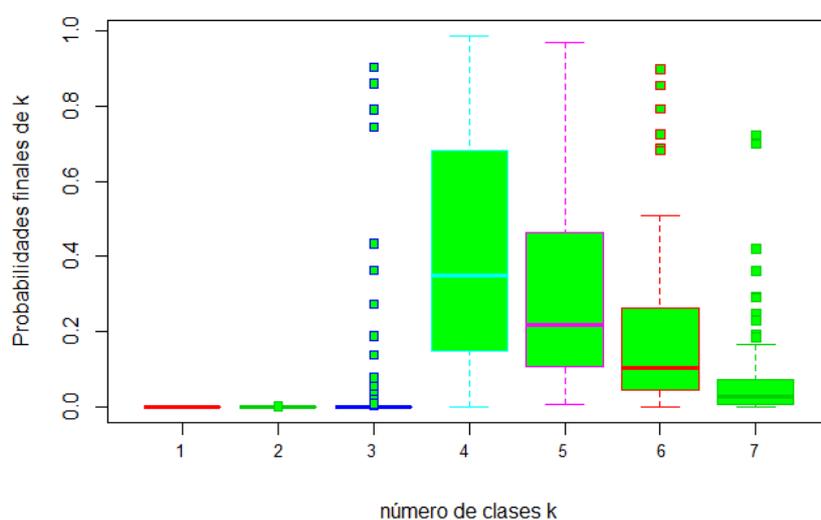


Figura 3.5: Probabilidad final de  $k$  en el escenario 4

Tabla 3.22: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con varias incidencias.

	$k$						
Incidencia	1	2	3	4	5	6	7
$3/10^5$ hab	0%	72%	25%	3%	0%	0%	0%
$30/10^5$ hab	0%	0%	6%	48%	29%	15%	3%
$300/10^5$ hab	0%	0%	0%	1%	65%	28%	6%

Tabla 3.23: Errores de ajuste del modelo en el escenario 4

$RMSE_{smr}$	$RMSE_{mod}$	$RMSE_{conoc}$
0.64(0.25)	0.34(0.05)	0.06(0.02)

Tabla 3.24: Errores de ajuste bajo diferentes valores de la incidencia

Incidencia	RMSE <sub>smr</sub>	RMSE <sub>mod</sub>	RMSE <sub>conoc</sub>
3/10 <sup>4</sup> hab	0.97(0.20)	0.88(0.15)	0.11(0.05)
30/10 <sup>4</sup> hab	0.64(0.25)	0.34(0.05)	0.06(0.02)
300/10 <sup>4</sup> hab	0.13(0.04)	0.10(0.05)	0.01(0.01)

de ajuste para los diferentes valores de la incidencia, se muestra de manera contundente el impacto que tiene ésta sobre las estimaciones del riesgo que hemos obtenido con el modelo propuesto.

A continuación mostramos un resumen en el que se comparan los resultados obtenidos en los distintos escenarios.

### 3.3. Estudio de simulación: comparación de escenarios

Queriendo dar al lector una visión más general de los resultados obtenidos en el estudio de simulación realizado, hacemos un análisis del impacto de los diferentes parámetros en la estimación del riesgo en áreas pequeñas. En la Tabla 3.25 encontramos tres tipos de los escenarios que han hecho parte del estudio de simulación: un escenario que hemos denominado *cercano*, en el cual los valores de los riesgos están muy cerca uno del otro, creando así cierta complejidad para distinguir las clases; un segundo escenario llamado *intermedio*, en el cual los valores de los riesgos conservan cierta distancia y el rango de valores que se proponen para los riesgos está más bien centrado; y finalmente el escenario *lejano* en el cual tratamos de dar la mayor amplitud entre riesgos. En esta tabla cruzada ubicamos en las filas los diferentes valores de probabilidad que se han contemplado en cada uno de los diferentes escenarios y los cuales se encuentran detallados en la Tabla 3.3. El resultado a analizar es el RMSE<sub>mod</sub>. Cada uno de estos escenarios son estudiados para diferentes incidencias, revisando así como varía el error cuando confluyen los diferentes valores de los parámetros que hacen parte del modelo. Cabe resaltar que además de estudiar las tablas en conjunto por incidencia, también las estudiaremos variando  $k$ .

Tabla 3.25: RMSE<sub>mod</sub> con  $r = 3/10^4$  y  $k = 2$ 

Prob.	Cercano	Intermedio	Lejano
1	0.21(0.02)	0.60(0.06)	0.70(0.11)
2	0.17(0.02)	0.53(0.07)	0.63(0.10)
3	0.17(0.02)	0.49(0.08)	0.58(0.11)
4	0.14(0.03)	0.46(0.08)	0.53(0.12)

En las Tablas 3.25, 3.26 y 3.27 en las cuales estudiamos  $k = 2$ , los valores de los riesgos subyacentes correspondientes al escenario ‘cercano’ son (0.8, 1.2), para el escenario ‘intermedio’ los valores son (0.5, 2.5) y en el escenario ‘lejano’ los valores de los riesgos son (0.5, 3.5). Si hacemos una lectura horizontal de la Tabla 3.25, podemos observar que el error de estimación aumenta cada vez que ampliamos el rango de valores para los riesgos subyacentes, esto es posible porque la variabilidad es mayor teniendo un intervalo más amplio para dichos riesgos. Sin embargo, al hacer una lectura de la tabla en forma vertical observamos que dentro de cada situación (cercano, intermedio, lejano) el error varía muy poco, pero hay que señalar que disminuye cuando las probabilidades se van distanciando lo que recuerda a la varianza de la Binomial, que también disminuye al alejarse  $p$  de 0.5; teniendo en cuenta lo anterior podemos afirmar que las probabilidades no afectan de manera importante las estimaciones de interés. Obsérvese que a nivel individual en las Tablas 3.26 y 3.27 sucede lo mismo, nótese que lo único que hemos cambiado es la incidencia. En los resultados de estas tablas vemos que los valores del error por columnas permanecen casi constantes, reflejando así la estabilidad de la estimación del riesgo bajo diferentes vectores de probabilidades. Ahora, si hacemos la lectura de las tablas de manera conjunta podemos observar como disminuye el error a medida que aumenta la información a través de la incidencia de la enfermedad estudiada.

Tabla 3.26:  $\text{RMSE}_{\text{mod}}$  con  $r = 30/10^4$  y  $k = 2$

Prob.	Cercano	Intermedio	Lejano
1	0.15(0.01)	0.19(0.06)	0.20(0.08)
2	0.13(0.02)	0.16(0.07)	0.15(0.09)
3	0.13(0.02)	0.15(0.07)	0.15(0.10)
4	0.11(0.02)	0.13(0.07)	0.12(0.09)

Tabla 3.27:  $\text{RMSE}_{\text{mod}}$  con  $r = 300/10^4$  y  $k = 2$

Prob.	Cercano	Intermedio	Lejano
1	0.05(0.01)	0.06(0.06)	0.04(0.06)
2	0.05(0.01)	0.04(0.05)	0.04(0.07)
3	0.05(0.01)	0.03(0.04)	0.03(0.05)
4	0.04(0.02)	0.03(0.03)	0.02(0.04)

Para el análisis de  $k = 3$ , los valores de los riesgos subyacentes correspondientes al escenario ‘cercano’ son (0.8, 1.0, 1.5), para el escenario ‘intermedio’ los valores son (0.5, 1.0, 2.0) y en el escenario ‘lejano’ los valores de los riesgos son (0.5, 1.0, 3.5). Al hacer una lectura horizontal de la Tabla 3.28, podemos observar que el error de estimación aumenta cada vez que ampliamos el rango de valores para los riesgos subyacentes, esto sucede porque la variabilidad es mayor teniendo un intervalo más amplio para dichos riesgos. Sin embargo, al hacer una lectura

de la tabla en forma vertical observamos que dentro de cada situación (cercano, intermedio, lejano) el error varía muy poco.

Tabla 3.28:  $\text{RMSE}_{\text{mod}}$  con  $r = 3/10^4$  y  $k = 3$ 

Prob	Cercano	Intermedio	Lejano
1	0.30(0.02)	0.49(0.04)	0.73(0.09)
2	0.30(0.02)	0.50(0.04)	0.76(0.08)
3	0.19(0.03)	0.36(0.04)	0.53(0.09)
4	0.25(0.03)	0.44(0.05)	0.66(0.10)
5	0.27(0.02)	0.45(0.05)	0.63(0.09)

Tabla 3.29:  $\text{RMSE}_{\text{mod}}$  con  $r = 30/10^4$  y  $k = 3$ 

Prob	Cercano	Intermedio	Lejano
1	0.19(0.02)	0.24(0.03)	0.26(0.08)
2	0.19(0.02)	0.88(0.06)	0.26(0.07)
3	0.14(0.02)	0.19(0.03)	0.19(0.07)
4	0.16(0.02)	0.20(0.04)	0.19(0.08)
5	0.17(0.02)	0.21(0.04)	0.20(0.07)

Tabla 3.30:  $\text{RMSE}_{\text{mod}}$  con  $r = 300/10^4$  y  $k = 3$ 

Prob	Cercano	Intermedio	Lejano
1	0.10(0.01)	0.19(0.04)	0.07(0.06)
2	0.42(0.03)	0.17(0.03)	0.07(0.06)
3	0.08(0.02)	0.16(0.05)	0.05(0.04)
4	0.07(0.02)	0.12(0.02)	0.04(0.05)
5	0.08(0.02)	0.16(0.06)	0.05(0.04)

Obsérvese que a nivel individual en las Tablas 3.29 y 3.30 sucede lo mismo, note que lo único que hemos cambiado es la incidencia. En los resultados de estas tablas vemos que los valores del error por columnas permanecen casi constantes, reflejando así la estabilidad de la estimación del riesgo bajo diferentes vectores de probabilidades. Ahora, si hacemos la lectura de las tablas de manera conjunta podemos observar como disminuye el error a medida que aumenta la información a través de la incidencia de la enfermedad estudiada.

Las Tablas 3.31, 3.32 y 3.33 muestran algunos escenarios con  $k = 5$ , el escenario ‘cercano’ con valores (0.2, 0.7, 1.0, 1.5, 2.5), para el escenario intermedio los valores son: (0.5, 0.7, 1.0, 1.5, 3.5) y el lejano (0.5, 1.0, 1.5, 2.5, 3.5), además las probabilidades se especifican en la Tabla 3.3. Al analizar dichas tablas encontramos que se comportan de manera similar a las estudiadas con  $k = 2$  y  $k = 3$ , el error va en aumento a medida que tenemos riesgos subyacentes más lejanos pero

el riesgo en cada uno de esos escenarios es relativamente el mismo, las probabilidades no afectan considerablemente. Sin embargo, note que en comparación con los escenarios en los que contemplamos  $k = 2$  y  $k = 3$  el error ahora es mayor en general, esto debido a que tenemos más clases a las cuales asignar las áreas pequeñas y esto vuelve más complejo el problema.

Tabla 3.31:  $\text{RMSE}_{\text{mod}}$  con  $r = 3/10^4$  y  $k = 5$

Prob.	Cercano	Intermedio	Lejano
1	0.58(0.05)	0.66(0.09)	0.73(0.06)
2	0.49(0.05)	0.59(0.08)	0.67(0.05)
3	0.48(0.06)	0.56(0.10)	0.62(0.08)
4	0.60(0.05)	0.76(0.07)	0.76(0.07)

Tabla 3.32:  $\text{RMSE}_{\text{mod}}$  con  $r = 30/10^4$  y  $k = 5$

Prob.	Cercano	Intermedio	Lejano
1	0.27(0.03)	0.28(0.06)	0.34(0.04)
2	0.25(0.03)	0.25(0.05)	0.33(0.04)
3	0.22(0.04)	0.22(0.07)	0.28(0.06)
4	0.28(0.04)	0.28(0.07)	0.34(0.05)

Tabla 3.33:  $\text{RMSE}_{\text{mod}}$  con  $r = 300/10^4$  y  $k = 5$

Prob.	Cercano	Intermedio	Lejano
1	0.11(0.03)	0.09(0.04)	0.10(0.04)
2	0.11(0.03)	0.10(0.04)	0.10(0.04)
3	0.08(0.04)	0.07(0.03)	0.08(0.05)
4	0.10(0.04)	0.08(0.05)	0.11(0.06)

Finalmente el estudio de  $k = 7$  en las Tablas 3.34, 3.35 y 3.36 cuyos escenarios son: ‘cercano’ con valores de riesgos (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0), escenario ‘intermedio’ (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5) y por último el escenario ‘lejano’ (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5). Los resultados nos confirman que aunque el comportamiento general del error es el mismo, la complejidad del modelo va ligada al aumento del número de clases.

Bajo la situación habitual en que desconocemos  $k$  resulta interesante estudiar su distribución de probabilidad en diversos escenarios. A continuación estudiamos de manera global el porcentaje en el que la mayor probabilidad corresponde a cada  $k$ . Para ello tenemos en cuenta las probabilidades finales de  $k$  obtenidas en cada uno de los diferentes escenarios, el número de veces que éstas toman el valor máximo y dicho valor se corresponde con el  $k$  verdadero. Los primeros estudios

Tabla 3.34:  $\text{RMSE}_{\text{mod}}$  con  $r = 3/10^4$  y  $k = 7$ 

Prob.	Cercano	Intermedio	Lejano
1	0.47(0.03)	0.72(0.07)	0.71(0.06)
2	0.44(0.03)	0.67(0.09)	0.70(0.07)
3	0.51(0.04)	0.76(0.09)	0.76(0.06)
4	0.45(0.04)	0.66(0.09)	0.70(0.07)
5	0.42(0.03)	0.68(0.07)	0.66(0.05)

Tabla 3.35:  $\text{RMSE}_{\text{mod}}$  con  $r = 30/10^4$  y  $k = 7$ 

Prob.	Cercano	Intermedio	Lejano
1	0.25(0.02)	0.31(0.05)	0.36(0.04)
2	0.23(0.03)	0.27(0.06)	0.34(0.05)
3	0.26(0.03)	0.32(0.06)	0.36(0.05)
4	0.22(0.03)	0.27(0.06)	0.34(0.04)
5	0.24(0.02)	0.29(0.06)	0.35(0.03)

Tabla 3.36:  $\text{RMSE}_{\text{mod}}$  con  $r = 300/10^4$  y  $k = 7$ 

Prob.	Cercano	Intermedio	Lejano
1	0.12(0.03)	0.11(0.04)	0.16(0.04)
2	0.10(0.02)	0.10(0.04)	0.15(0.04)
3	0.11(0.02)	0.12(0.05)	0.15(0.04)
4	0.10(0.02)	0.10(0.04)	0.14(0.03)
5	0.11(0.02)	0.11(0.04)	0.15(0.03)

corresponden a un  $k = 2$ , los escenarios son los mismos que hemos definido anteriormente, pero para el análisis de las probabilidades nos hemos quedado con un vector con valores iguales  $p_1 = (0.5, 0.5)$  y otro vector con probabilidades diferentes  $p_2 = (0.2, 0.8)$ , dado que las probabilidades no influyen demasiado en los resultados como hemos comprobado antes.

Tabla 3.37: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 3/10^4$  y  $k = 2$ ,

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	95	4	1	0	0	0	0
	$p_2$	97	3	0	0	0	0	0
Intermedio	$p_1$	0	85	11	4	0	0	0
	$p_2$	0	98	2	0	0	0	0
Lejano	$p_1$	0	88	7	4	0	1	0
	$p_2$	0	95	3	2	0	0	0

En la Tabla 3.37 se observa que en el escenario donde los riesgos latentes están cercanos el porcentaje de correspondencia mayor no es el del  $k$  real sino el de  $k = 1$ , esto debido a la cercanía de los riesgos. Quizá bastaría una clase para hacer la asignación de áreas.

Tabla 3.38: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 30/10^4$  y  $k = 2$

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	1	97	2	0	0	0	0
	$p_2$	28	71	1	0	0	0	0
Intermedio	$p_1$	0	98	2	0	0	0	0
	$p_2$	0	99	1	0	0	0	0
Lejano	$p_1$	0	98	2	0	0	0	0
	$p_2$	0	100	0	0	0	0	0

Nótese que al estudiar conjuntamente las Tablas 3.37, 3.38 y 3.39 los porcentajes de correspondencia se asocian con el  $k$  real a medida que aumenta la incidencia. Obsérvese que con  $r = 300$  la asignación es perfecta.

Para el análisis de los siguientes escenarios en los que  $k = 3$ , los escenarios son los mismos que hemos fijado para el análisis anterior, los vectores de probabilidades que hemos fijado son los siguientes:  $p_1 = (1/3, 1/3, 1/3)$  y  $p_2 = (0.2, 0.7, 0.1)$ . Nuevamente se observa que al estudiar conjuntamente las Tablas 3.40, 3.41 y 3.42

Tabla 3.39: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 300/10^4$  y  $k = 2$ 

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	100	0	0	0	0	0
	$p_2$	0	100	0	0	0	0	0
Intermedio	$p_1$	0	100	0	0	0	0	0
	$p_2$	0	100	0	0	0	0	0
Lejano	$p_1$	0	100	0	0	0	0	0
	$p_2$	0	100	0	0	0	0	0

los porcentajes de correspondencia se asocian con el  $k$  real a medida que aumenta la incidencia.

Tabla 3.40: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 3/10^4$  y  $k = 3$ 

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	85	15	0	0	0	0	0
	$p_2$	98	2	0	0	0	0	0
Intermedio	$p_1$	1	80	12	3	2	2	0
	$p_2$	55	39	4	1	0	1	0
Lejano	$p_1$	0	55	36	5	3	0	1
	$p_2$	6	68	17	5	1	3	0

En las tablas asociadas al estudio de  $k = 5$  los escenarios fijados son como en el análisis anterior y la probabilidades para el análisis son las siguientes:  $p_1 = (0.2, 0.2, 0.2, 0.2, 0.2)$  y  $p_2 = (0.6, 0.1, 0.1, 0.1, 0.1)$  es claro ver que el porcentaje de correspondencia asociado al  $k$  correcto aumenta considerablemente cada vez que se incrementa el valor de la incidencia. Lo mismo sucede con  $k = 7$  (Tablas 3.46, 3.47 y 3.48) donde hemos tomado los mismos escenarios (cercano, intermedio, lejano) definidos antes y las probabilidades son las siguientes:  $p_1 = (1/7, 1/7, 1/7, 1/7, 1/7, 1/7, 1/7)$  y  $p_2 = (0.1, 0.1, 0.1, 0.4, 0.1, 0.1, 0.1)$ .

A la vista de los resultados obtenidos en este estudio de simulación (ver Anexos para resultados adicionales) podemos concluir que el modelo propuesto permite estimar el riesgo en áreas pequeñas y clasificar las áreas sin necesidad de forzar desde el inicio la inclusión de una estructura de vecindad. Además, a lo largo de este estudio de simulación observamos que los riesgos subyacentes y las probabilidades estimadas muestran que el modelo es capaz de recuperar la es-

Tabla 3.41: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 30/10^4$  y  $k = 3$

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	74	23	1	1	1	0
	$p_2$	17	75	7	1	0	0	0
Intermedio	$p_1$	0	0	85	12	3	0	0
	$p_2$	0	0	85	9	5	1	0
Lejano	$p_1$	0	0	85	9	3	1	2
	$p_2$	0	0	91	6	2	1	0

Tabla 3.42: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 300/10^4$  y  $k = 3$

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	0	94	6	0	0	0
	$p_2$	0	1	95	4	0	0	0
Intermedio	$p_1$	0	0	98	1	1	0	0
	$p_2$	0	0	99	1	0	0	0
Lejano	$p_1$	0	0	100	0	0	0	0
	$p_2$	0	0	98	2	0	0	0

Tabla 3.43: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 3/10^4$  y  $k = 5$

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	28	42	19	5	5	1
	$p_2$	0	39	25	21	6	8	1
Intermedio	$p_1$	0	35	41	8	9	3	4
	$p_2$	1	46	31	11	9	2	1
Lejano	$p_1$	0	45	32	15	5	1	3
	$p_2$	0	42	38	9	4	2	5

Tabla 3.44: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 30/10^4$  y  $k = 5$ 

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	0	0	16	41	32	11
	$p_2$	0	0	6	50	23	13	8
Intermedio	$p_1$	0	0	5	47	34	10	4
	$p_2$	0	0	29	44	12	11	4
Lejano	$p_1$	0	0	0	5	35	32	28
	$p_2$	0	0	7	38	31	15	9

Tabla 3.45: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 300/10^4$  y  $k = 5$ 

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	0	0	0	58	32	10
	$p_2$	0	0	0	0	72	23	5
Intermedio	$p_1$	0	0	0	0	42	45	13
	$p_2$	0	0	0	3	63	27	7
Lejano	$p_1$	0	0	0	0	57	30	13
	$p_2$	0	0	0	0	64	27	9

Tabla 3.46: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 3/10^4$  y  $k = 7$ 

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	64	22	7	6	1	0
	$p_2$	0	75	13	8	1	1	2
Intermedio	$p_1$	0	21	41	19	13	4	2
	$p_2$	0	15	56	15	9	3	3
Lejano	$p_1$	0	46	42	7	1	3	1
	$p_2$	0	70	19	6	2	2	1

Tabla 3.47: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 30/10^4$  y  $k = 7$

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	0	0	16	32	36	16
	$p_2$	0	0	4	35	34	19	8
Intermedio	$p_1$	0	0	0	11	26	37	26
	$p_2$	0	0	0	16	37	27	21
Lejano	$p_1$	0	0	0	4	34	34	28
	$p_2$	0	0	1	20	33	31	15

Tabla 3.48: Porcentaje en el que la mayor probabilidad corresponde a cada  $k$  con  $r = 300/10^4$  y  $k = 7$

$k$		1	2	3	4	5	6	7
Cercano	$p_1$	0	0	0	0	0	9	91
	$p_2$	0	0	0	0	1	19	80
Intermedio	$p_1$	0	0	0	0	0	17	83
	$p_2$	0	0	0	0	0	21	79
Lejano	$p_1$	0	0	0	0	0	4	96
	$p_2$	0	0	0	0	0	14	86

estructura real que se ha fijado. Resulta interesante el hecho de que la distribución final de  $k$  permite estimar el número desconocido de clases y aunque éste en sí no es el objetivo principal de este modelo, usando estas probabilidades y los riesgos latentes estimados, podemos obtener el riesgo relativo estimado para cada una de las áreas pequeñas.

### 3.4. Estudio de un caso real: varicela en la ciudad de Valencia en el período 2008-2013.

A continuación mostramos los resultados obtenidos en el análisis del banco de datos correspondiente a casos declarados de varicela en 70 barrios de la ciudad de Valencia (Figura 3.6) durante los años 2008 y 2013.

La varicela es una enfermedad vírica aguda y altamente contagiosa, de comienzo repentino, con fiebre moderada, síntomas generales mínimos y una erupción cutánea de tipo maculopapular durante pocas horas y vesicular durante tres o cuatro días, que deja costras granulosas. Las complicaciones de la varicela ocurren entre las personas inmunocomprometidas, neonatos, niños menores de un año de edad y adultos. El virus que causa la varicela es el varicela zoster, un miembro de la familia del herpesvirus, que causa también el herpes zóster en los adultos (Peña-Rey et al., 2009).

En los últimos años ha ido en aumento el interés en el estudio de esta enfermedad con el fin de intensificar la vigilancia y realizar la identificación oportuna de casos para la toma de medidas de control que contribuyan a disminuir la incidencia de la enfermedad. Entre los factores que más se estudian están la edad de los pacientes y el estado vacunal. En la actualidad se están modificando los calendarios vacunales de varicela en España, esto con el fin de aumentar la cobertura de la aplicación de dicha vacuna, pues se ha observado que la incidencia ha disminuido considerablemente. La incidencia anual de varicela en la ciudad de Valencia para los años 2008 a 2013 está alrededor de 500 por cada 100000 habitantes (Informe varicela - Vigilancia epidemiológica, año 2008, GV). La Tabla 3.49 resume algunos recuentos importantes publicados en el Informe anual de varicela elaborado por la Conselleria de Sanitat. Nótese que en este estudio sólo vamos a trabajar con los 70 barrios mostrados en la gráfica y no con los 87 barrios que conforman la ciudad de Valencia, esto debido a que nos compararemos con otros estudios que se han realizado utilizando dichos datos. Una de las posibles razones por la que se haya delimitado esta información puede estar asociada al hecho de que en las zonas eliminadas la población es poca.



### 3.4.1. Resultados y comparación con estudios previos.

Los trabajos de Iftimi et al. (2015a,b) se han ocupado de estudiar el comportamiento de la varicela en la ciudad de Valencia para los años 2008 a 2013. En el primer trabajo aplican un modelo de conglomerados y evalúan la detección de existencias de conglomerados así como interacciones espacio-temporales. Realizan el análisis temporal a través de la regresión armónica que modeliza el logaritmo de la intensidad y el análisis espacial lo hacen a través de un método no paramétrico para confrontar segregación espacial en un proceso espacial. Para el análisis puramente espacial toman como población de referencia la del año 2010 y realizan el contraste de hipótesis donde el riesgo es constante en todas las localizaciones, frente a una alternativa en la que al menos un riesgo es diferente. Observan que existe una mayor probabilidad de contraer la enfermedad en la periferia de la ciudad, ya que en este sitio la población es más joven. También concluyen que hay una menor probabilidad de adquirir la enfermedad en el área central debido a que allí la mayoría de la población está conformada por adultos mayores y a esto se asocia una incidencia menor. Por último, destacan que en el sur de Valencia hay una alta concentración de áreas con nivel socio-económico medio bajo y en consecuencia es posible suponer que en muchos hogares de esta zona los niños no están vacunados.

En Iftimi et al. (2015b) estudian herramientas para un análisis exploratorio espacio-temporal para detectar el comportamiento específico de la varicela en Valencia para los años 2008 a 2013. El objetivo de dicho trabajo es encontrar una relación entre la componente espacial y temporal del proceso de la enfermedad, para ello estudian e implementan el modelo propuesto por Knorr-Held (2000). Dentro del análisis realizado observan que de los posibles factores asociados a altos riesgos relativos está el hecho de vivir en zonas económicamente desaventajadas y áreas menos involucradas en programas de vacunación. Las Figuras 3.7 y 3.8 muestran los mapas con los riesgos estimados por estos autores para cada uno de los barrios en la ciudad de Valencia para los años 2008 y 2013 respectivamente.

Hemos aplicado el modelo de conglomerados con estructura de clases latentes para estimar el riesgo relativo en cada uno de los 70 barrios que conforman nuestra zona de estudio y simultáneamente clasificar áreas con riesgo similar sin forzar a que ellas compartan algún límite geográfico. Hemos utilizado un ordenador portátil con las siguientes características: procesador Intel(R) Core(TM) i5-4200U CPU 1.60 GHz 1.60 GHz, memoria instalada (RAM) 6 GB, sistema operativo de 64 bits. El tiempo computacional promedio que tardó la implementación del modelo propuesto con estos datos fue de aproximadamente 4.25 minutos en realizar los 5000 puntos del algoritmo de Gibbs explicado en el Capítulo 2 y 6.15 minutos en las 100000 simulaciones por el método Monte-Carlo para estimar las probabilidades finales de  $k$ .

La Tabla 3.50 presenta un resumen de los riesgos relativos estimados para

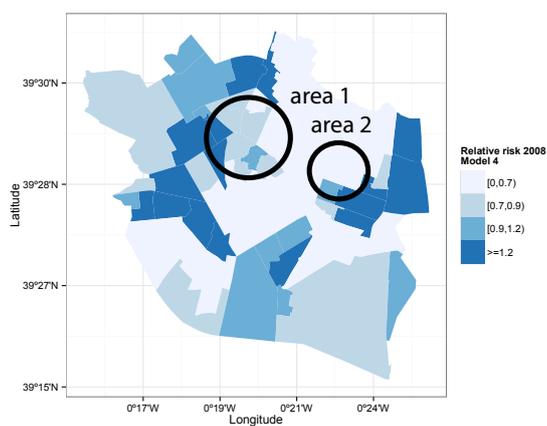


Figura 3.7: Mapa de riesgos estimados de varicela año 2008 con el modelo expuesto en Iftimi et al. (2015b). Fuente: Extraído del artículo de Iftimi et al. (2015b).

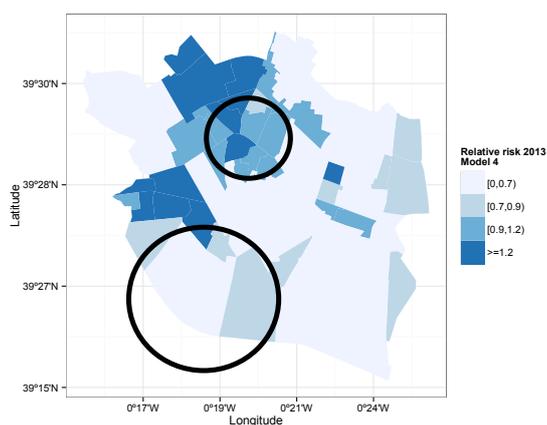


Figura 3.8: Mapa de riesgos estimados de varicela año 2013 con el modelo expuesto en Iftimi et al. (2015b). Fuente: Extraído del artículo de Iftimi et al. (2015b).

cada barrio. Los datos analizados en la última fila corresponden a la suma de los datos anuales desde 2008 a 2013.

Tabla 3.50: Resumen de medidas para el riesgo estimado en cada barrio

Año	Mín	Media	Desviación	Máx
2008	0.21	0.94	0.44	2.31
2009	0.36	0.96	0.57	3.11
2010	0.38	0.99	0.58	2.26
2011	0.37	0.89	0.52	2.20
2012	0.33	0.94	0.70	4.22
2013	0.45	1.05	0.78	5.19
2008-2013	0.33	0.95	0.37	1.74

Las probabilidades finales de  $k$  para los datos de la varicela en el año 2008 y 2013 se presentan en la tabla 3.51. Para el año 2008 la mayor probabilidad se asocia con  $k = 5$  y el resto de probabilidades se reparte entre las clases 4, 6 y 7. Nótese que para los datos de la varicela en el año 2013 las probabilidades están más repartidas. La mayor probabilidad se asocia con  $k = 7$  y el resto de probabilidades indican que las clases 3, 4, 5 y 6 también tienen cierta probabilidad de estar ocupadas.

Tabla 3.51: Probabilidades finales de  $k$ . Años 2008 y 2013

$k$	1	2	3	4	5	6	7
2008	0.00	0.00	0.00	0.01	0.45	0.21	0.33
2013	0.00	0.00	0.02	0.23	0.26	0.13	0.37

Las Figuras 3.9 y 3.10 muestran los mapas de los riesgos estimados bajo el modelo propuesto para cada uno de los dos años que hacen parte del estudio. En la representación del riesgo para el año 2008 se observa que los valores de mayor riesgo se encuentran en la periferia y los de menor riesgo en la zona central de la ciudad. Un comportamiento similar pero quizá un poco más suavizado lo observamos en los resultados obtenidos en el año 2013.

Tabla 3.52: Resumen de los riesgos relativos de cada área pequeña

	Mín	Q1	Mediana	Media	Q3	Máx
2008	0.21	0.60	0.86	0.94	1.19	2.31
2013	0.45	0.73	0.89	1.05	1.06	5.19

La Tabla 3.52 presenta el resumen de las medidas para las estimaciones de los riesgos relativos en cada área pequeña para los años 2008 y 2013.

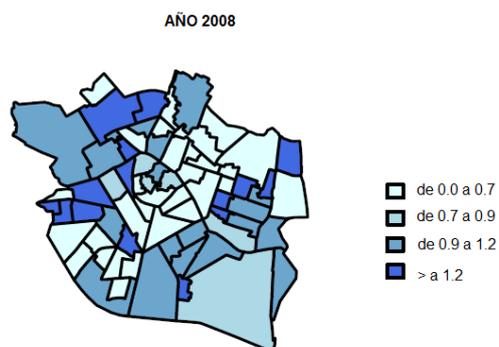


Figura 3.9: Mapa de riesgos estimados para la varicela en el año 2008 con el modelo propuesto.

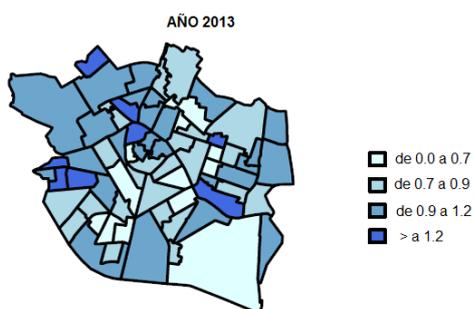


Figura 3.10: Mapa de riesgos estimados para la varicela en el año 2013 con el modelo propuesto.

---

Al compararnos con el trabajo de Iftimi et al. (2015b) observamos que los riesgos estimados para cada uno de los barrios son similares, con la ventaja de que con el modelo propuesto no imponemos una estructura de vecindad y sin embargo somos capaces de recuperar la relación espacial existente. Si comparamos la Figura 3.9 con la Figura 3.7 vemos como recuperamos la relación espacial en la zona de estudio, coincidimos con la ubicación de las zonas de alto y bajo riesgo y también observamos que en la periferia es donde se ubican las zonas con mayor valor en el riesgo, quizá se esté observando este fenómeno debido a que en dichas zonas suelen vivir personas más jóvenes con niveles socio-económicos medio bajos y algunos de ellos sin acceso a programas de vacunación. Del mismo modo, si comparamos la Figura 3.10 con la Figura 3.8 observamos una cantidad importante de áreas con riesgos similares a los representados en el mapa de referencia .

En el Capítulo 4 continuaremos con el estudio del comportamiento del riesgo de la varicela en los años 2008 y 2013, pero incluyendo covariables.



## Capítulo 4

# Análisis del modelo de conglomerados incluyendo covariables

En este capítulo presentamos una extensión del modelo de conglomerados con estructura de clases latentes para mapeo de enfermedades, que nos permitirá estimar el riesgo de una cierta enfermedad en una región de estudio dividida en áreas pequeñas teniendo en cuenta el efecto de una o más covariables, así como identificar áreas que tienen un riesgo similar.

### 4.1. Modelo de conglomerados con estructura de clases latentes incluyendo covariables y $k$ conocido.

Como se detalló en el Capítulo 2 vamos a suponer que la zona de estudio tiene  $m$  áreas pequeñas. El vector de datos observados es  $y = (y_1, \dots, y_m)'$ , donde  $y_i$  representa el número de casos en el área  $i$ . Además  $e_i$  es el número de casos esperados en la  $i$ -ésima área bajo el modelo en el que todas las áreas tienen igual riesgo de contraer cierta enfermedad, el vector  $e = (e_1, \dots, e_m)'$  siempre se considera conocido.

En este nuevo análisis incorporamos las covariables en el modelo a través del predictor lineal  $x_i'\beta$  donde  $x_i$  es la  $i$ -ésima fila de la matriz de diseño que incluye los valores de las covariables para cada área  $i$  que se consideren en el estudio,  $x_i = (x_{1i}, \dots, x_{Li})$ .  $Y$ ,  $\beta$  es el vector de coeficientes de cada una de las covariables,  $\beta = (\beta_1, \dots, \beta_L)$  con  $L$  el número de covariables incluidas en el modelo.

Vamos a seguir utilizando como modelo una mixtura de Poissons en el cual se tiene un número de clases  $k$  conocido y cada una de ellas tiene asociado un riesgo

subyacente. La probabilidad de pertenecer a cada una de las diferentes clases existentes en el modelo viene dada por el vector  $p = (p_1, \dots, p_k)'$  con  $\sum_{j=1}^k p_j = 1$  y  $p_j \geq 0$ . Además nuestro modelo incluye una estructura latente a través de vectores de variables latentes (o variables de asignación no observadas)  $z_i = (z_{i1}, \dots, z_{ik})'$  con  $i = 1, \dots, m$ , donde  $z_{ij} = 1$  identifica la clase  $j$  a la cual pertenece cada área  $i$ . La matriz de variables de asignación  $Z$  y el vector de riesgos latentes  $\eta$  es igual que antes.

Habiendo especificado lo anterior, asumimos que los conteos  $y_i$  provienen de una mixtura de distribuciones Poisson, de manera que

$$y_i | z_i \sim \text{Po}(e_i \theta_i),$$

donde  $\theta_i$ , el riesgo de cada una de las áreas pequeñas, ahora es modelizado teniendo en cuenta los conglomerados y las covariables.

$$\theta_i = \exp\{x_i' \beta\} z_i' \eta, \quad (4.1)$$

con  $z_i' \eta$  indicando el riesgo latente de cada una de las componentes o clases.

Con la modelización presentada en la ecuación (4.1) pretendemos recoger en el primer término el mayor efecto de las covariables a estudiar y en el segundo término  $z_i' \eta$  la variación asociada al riesgo residual.

Finalmente, el modelo lo podemos representar así:

$$y_i | k, z_i, \eta, \beta, x_i, e_i \sim \text{Po}(y_i | e_i \exp\{x_i' \beta\} z_i' \eta) \quad (4.2)$$

$$z_i | k, p \sim \text{Mult}_k(z_i | n = 1, p) \quad (4.3)$$

Como es habitual la elección estándar para la distribución de  $z_i$  es la Multinomial ( $\text{Mult}_k$ ). Seguimos considerando que los vectores  $(y_i, z_i)$ , con  $i = 1, \dots, m$ , son independientes dados los parámetros. Así, la distribución conjunta de los datos observados y las variables latentes es expresada como una productoria de Poissons y Multinomiales, facilitando el análisis de la función de verosimilitud:

$$\begin{aligned} f(y, z | k, \eta, p, \beta, x, e) &\propto f(y | k, z, \eta, \beta, x, e) f(z | k, p) \\ f(y, z | k, \eta, p, \beta, x, e) &\propto \prod_{i=1}^m \text{Po}(y_i | e_i \exp\{x_i' \beta\} z_i' \eta) z_i' p \\ &\propto \left[ \prod_{i=1}^m (\exp\{x_i' \beta\} z_i' \eta)^{y_i} \exp\{-e_i \exp\{x_i' \beta\} z_i' \eta\} \right] \left[ \prod_{i=1}^m z_i' p \right] \end{aligned}$$

A continuación mostramos el análisis bayesiano del modelo propuesto anteriormente.

#### 4.1.1. Análisis bayesiano del modelo con covariables: Distribuciones previas y a posteriori

Iniciamos con la previa para los riesgos latentes de cada una de las clases  $\eta$ , dichos valores se encuentran en la semirecta real positiva y como se expuso en el Capítulo 2 la previa que proponemos permite que estos riesgos estén ordenados garantizando de este modo identificabilidad de las clases. Recordemos que para este fin establecimos unos puntos de corte, de manera que cada uno de los riesgos  $\eta_j$  oscilará entre un punto de corte y otro. La distribución inicial de los riesgos  $\eta$  es como en el Corolario 1 del Capítulo 2:

$$f(\eta|v) \propto \prod_{j=1}^k \frac{1}{(v_j - v_{j-1})} \frac{1}{(1 + \eta_j)^2} \mathbf{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j),$$

donde  $\mathbf{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j)$  es la función indicatriz o indicadora, que indica la pertenencia o no de cada  $\eta_j$  al intervalo  $\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)$ , de modo que asignará el valor de 1 a todos los elementos que pertenezcan a dicho intervalo y 0 en caso contrario.

Basados en la distancia que existe entre un punto de corte y el siguiente, es decir entre  $d_j = v_j - v_{j-1}$  asignamos una distribución previa Dirichlet para  $d$ . La previa sobre  $v$  se obtiene haciendo el cambio de variables y resulta ser la misma que en el estudio sin covariables:

$$f(v) \propto \prod_{j=1}^k (v_j - v_{j-1})^{\gamma_j - 1} \mathbf{I}_{(v_{j-1}, v_{j+1})}(v_j).$$

La previa sobre el vector de probabilidades  $p$  es una Dirichlet de parámetro  $\alpha$ ,

$$f(p) \propto \prod_{j=1}^k p_j^{\alpha_j - 1}.$$

Como previa para el parámetro  $\beta = (\beta_1, \dots, \beta_L)'$ , que corresponde al vector de coeficientes de las covariables, proponemos una distribución normal multivariante con vector de medias cero y matriz de varianzas y covarianzas  $\Sigma$ .

$$f(\beta|\Sigma) \propto |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}\beta'\Sigma^{-1}\beta\right\}.$$

Así mismo para la previa de la matriz de varianzas y covarianzas trabajamos con una distribución Wishart inversa con matriz de escala  $\psi$ , de dimensión  $L \times L$ , y  $\nu$  grados de libertad.

$$f(\Sigma) \propto |\Sigma|^{-\frac{\nu+L+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\psi\Sigma^{-1})\right\}.$$

Haciendo uso de la verosimilitud conjunta de  $z$  e  $y$  junto con las previas descritas anteriormente construimos la distribución a posteriori conjunta:

$$\begin{aligned} f(z, \eta, p, v, \beta, \Sigma | k, y, x, e) &\propto f(y, z | k, \eta, p, v, \beta, \Sigma, x, e) f(\eta, p, v, \beta, \Sigma) \\ &\propto f(y | k, z, \eta, \beta, x, e) f(z | p) f(p) f(\eta | v) f(v) f(\beta | \Sigma) f(\Sigma) \\ &\propto \prod_{i=1}^m f(y_i | k, z_i, \eta, \beta, x_i, e_i) \prod_{i=1}^m f(z_i | p) \\ &\quad f(p) f(\eta | v) f(v) f(\beta | \Sigma) f(\Sigma) \\ &\propto \prod_{i=1}^m (\exp\{x_i' \beta\} z_i' \eta)^{y_i} \exp\{-e_i \exp\{x_i' \beta\} z_i' \eta\} \prod_{i=1}^m z_i' p \\ &\quad \prod_{j=1}^k p_j^{\alpha_j - 1} \\ &\quad \prod_{j=1}^k \frac{1}{(v_j - v_{j-1})} \frac{1}{(1 + \eta_j)^2} \mathbf{I}_{\left(\frac{v_{j-1}}{1-v_{j-1}}, \frac{v_j}{1-v_j}\right)}(\eta_j) \\ &\quad \prod_{j=1}^k (v_j - v_{j-1})^{\gamma_j - 1} \mathbf{I}_{(v_{j-1}, v_{j+1})}(v_j) \\ &\quad |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}\beta' \Sigma^{-1} \beta\right\} \\ &\quad |\Sigma|^{-\frac{\nu+L+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\psi\Sigma^{-1})\right\} \end{aligned}$$

Hacemos uso del método de Gibbs sampler (Casella, 1992) para muestrear de la distribución conjunta y para ello estudiaremos la distribución condicional completa para cada parámetro.

#### 4.1.2. Distribuciones condicionales completas

Exceptuando las distribuciones condicionales completas de los parámetros asociados a las covariables  $\beta$  y  $\Sigma$ , las demás condicionales completas se obtienen de forma parecida a como se hizo en el Capítulo 2. Las incluimos a continuación indicando sus similitudes y discrepancias.

La distribución condicional completa del vector de variables latentes de asignación  $z_i$  dado los datos y los demás parámetros se escribe a continuación.

**Proposición 11.** *Distribución condicional completa de  $z_i$*

$$f(z_i = E_l | k, \eta, p, \beta, \Sigma, x_i, y_i, e_i) = C \eta_l^{y_i} \exp\{-e_i \exp\{x_i' \beta\} \eta_l\} p_l,$$

$\forall l = 1, \dots, k$ , siendo

$$C = \frac{1}{\sum_{l=1}^k (\eta_l^{y_i} \exp\{-e_i \exp\{x_i' \beta\} \eta_l\} p_l)}.$$

La demostración es igual que en el Capítulo 2, aquí debemos tener en cuenta que la media de la Poisson con la cual estamos trabajando es la que incluye las covariables a través del término del predictor lineal. Esto se reduce a sustituir cada  $e_i$  por  $e_i \exp\{x_i' \beta\}$ .

Por otro lado, al revisar la condicional completa para los riesgos de cada clase  $\eta_j$  encontramos que en este análisis en el modelo existirán tantos riesgos como clases. Para estudiar dicha distribución debemos tener en cuenta que la matriz  $Z$  es conocida, por eso reescribimos los valores observados y esperados de la siguiente manera:

$$y_{(j)} = \sum_{i=1}^m y_i(z_i == E_j)$$

$$e_{(j)}^* = \sum_{i=1}^m e_i \exp\{x_i' \beta\}(z_i == E_j).$$

De esta forma  $z_i == E_j$  actúa como un operador lógico donde el sumando sólo aparecerá cuando el área  $i$  vaya a la clase  $j$  adecuada. Con  $E_j$  la columna  $j$ -ésima de la matriz identidad  $k$ -dimensional.

Con lo anterior, la distribución condicional completa del riesgo en cada clase se muestra en la siguiente proposición:

**Proposición 12.**

$$f(\eta_j | k, v, z, \beta, \Sigma, \eta_{-j}, x, y, e) \propto \text{Ga}[\eta_j | y_{(j)} + 1, e_{(j)}^*] \frac{1}{(1 + \eta_j)^2}.$$

Encontramos la demostración correspondiente en el Capítulo 2. La única diferencia es, como en el caso anterior, que sustituimos cada  $e_i$  por  $e_i \exp\{x_i' \beta\}$ .

**Proposición 13.** *La distribución condicional completa sobre  $p$  es como sigue:*

$$f(p|k, \eta, z, v, \alpha, \beta, \Sigma, x, y, e) \propto \prod_{j=1}^k p_j^{(\sum_{i=1}^m z_{ij}) + \alpha_j - 1},$$

lo que significa que sigue una distribución Dirichlet  $(p|\alpha + \sum_{i=1}^m z_i)$ .

Este es exactamente el mismo resultado obtenido en el Capítulo 2.

**Proposición 14.** *La distribución condicional completa sobre cada punto de corte  $v_j$ , con  $j = 1, \dots, k - 1$ , es como sigue:*

$$f(v_j|k, \eta, p, z, v_{-j}, y, e) \propto (v_j - v_{j-1})^{\eta_j - 2} (v_{j+1} - v_j)^{\eta_{j+1} - 2} \\ \text{si } \frac{\eta_j}{1 + \eta_j} < v_j < \frac{\eta_{j+1}}{1 + \eta_{j+1}}.$$

Este también es el mismo resultado obtenido en el Capítulo 2.

**Proposición 15.** *La distribución condicional completa del vector de coeficientes de las covariables es como sigue:*

$$f(\beta|k, \eta, p, z, \Sigma, x, y, e) \propto \exp\left\{\left(\sum_{i=1}^m y_i x_i\right)' \beta\right\} \exp\left\{-\frac{1}{2} \beta' \Sigma^{-1} \beta\right\} \\ \exp\left\{\sum_{i=1}^m -e_i \exp\{x_i' \beta\} z_i' \eta\right\}.$$

*Demostración.*

$$\begin{aligned}
 f(\beta|k, \eta, p, z, \Sigma, x, y, e) &\propto \prod_{i=1}^m (\exp\{x'_i \beta\} z'_i \eta)^{y_i} \exp\{-e_i \exp\{x'_i \beta\} z'_i \eta\} \\
 &\quad |\Sigma|^{-1/2} \exp\{-\frac{1}{2} \beta' \Sigma^{-1} \beta\} \\
 &\propto \prod_{i=1}^m (\exp\{x'_i \beta\})^{y_i} \exp\{-e_i \exp\{x'_i \beta\} z'_i \eta\} \\
 &\quad \exp\{-\frac{1}{2} \beta' \Sigma^{-1} \beta\} \\
 &\propto \left[ \prod_{i=1}^m (\exp\{x'_i \beta\})^{y_i} \right] \left[ \exp\{-\frac{1}{2} \beta' \Sigma^{-1} \beta\} \right] \\
 &\quad \left[ \prod_{i=1}^m \exp\{-e_i \exp\{x'_i \beta\} z'_i \eta\} \right] \\
 &\propto \exp\left\{ \left( \sum_{i=1}^m y_i x_i \right)' \beta \right\} \exp\left\{ -\frac{1}{2} \beta' \Sigma^{-1} \beta \right\} \\
 &\quad \exp\left\{ \sum_{i=1}^m -e_i \exp\{x'_i \beta\} z'_i \eta \right\}
 \end{aligned}$$

■

En el caso particular en el que haya una única covariable,  $L=1$ ,  $\Sigma$  ya no es una matriz, es la varianza de una distribución Normal univariante, y la representaremos como  $\sigma^2$ . En ese caso se obtiene el siguiente corolario:

**Corolario 2.**

$$\begin{aligned}
 f(\beta|k, \eta, p, z, \sigma, x, y, e) &\propto \exp\left\{ \sum_{i=1}^m x_i y_i \beta \right\} \exp\left\{ -\frac{\beta^2}{2\sigma^2} \right\} \\
 &\quad \exp\left\{ \sum_{i=1}^m -e_i \exp\{x_i \beta\} z'_i \eta \right\}.
 \end{aligned}$$

Por último, la distribución condicional completa sobre  $\Sigma$  es:

**Proposición 16.** *La distribución condicional completa de la matriz de varianzas y covarianzas  $\Sigma$  es Wishart inversa con matriz de escala  $(\beta\beta' + \psi)$  y  $\nu + 1$  grados de libertad.*

*Demostración.*

$$\begin{aligned}
 f(\Sigma|k, \eta, p, z, \beta, \psi, v, x, y, e) &\propto |\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}\beta'\Sigma^{-1}\beta\right\} |\Sigma|^{-\frac{v+L+1}{2}} \\
 &\quad \exp\left\{-\frac{1}{2}\text{tr}(\psi\Sigma^{-1})\right\} \\
 &\propto |\Sigma|^{-\frac{v+1+L+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}(\beta\beta'\Sigma^{-1}) + \text{tr}(\psi\Sigma^{-1})\right\} \\
 &\propto |\Sigma|^{-\frac{v+1+L+1}{2}} \exp\left\{-\frac{1}{2}\text{tr}((\beta\beta' + \psi)\Sigma^{-1})\right\}.
 \end{aligned}$$

■

En el caso particular en el que haya una única covariable, no utilizaremos una distribución inicial Wishart inversa, en su lugar utilizaremos una  $\text{Ga}(a, b)$  sobre la precisión  $h = \sigma^{-2}$ . En ese caso, la distribución condicional completa sobre la precisión es:

**Corolario 3.** *Sea  $h = \sigma^{-2}$ , entonces*

$$f(h) \sim \text{Ga}\left(a + \frac{1}{2}, \frac{\beta^2}{2} + b\right)$$

*Demostración.*

$$\begin{aligned}
 f(h|k, \eta, p, z, \beta, a, b, x, y, e) &\propto (h)^{1/2} \exp\left\{-\frac{\beta^2}{2}h\right\} (h)^{a-1} \exp\{-bh\} \\
 &\propto \exp\left\{-h\left(\frac{\beta^2}{2} + b\right)\right\} (h)^{a+\frac{1}{2}-1}.
 \end{aligned}$$

■

### 4.1.3. Muestreo de las distribuciones condicionales completas

El muestreo de las distribuciones condicionales completas de  $p$  y  $z$  es inmediato pues conocemos la forma analítica de sus distribuciones. Para muestrear de la distribución condicional completa de los riesgos de las clases no existe una rutina de generación de variables aleatorias en forma automática, sin embargo, podemos simular de ella estudiando los casos que fueron explicados en el Capítulo 2.

Para el proceso de la simulación de la distribución condicional completa de los puntos de corte  $v$  utilizaremos el mismo proceso que se explicó en detalle en el Capítulo 2.

Para muestrear de la distribución condicional completa del vector de coeficientes  $\beta$  proponemos un esquema de Metropolis.

Dado un valor inicial  $\beta_{ini}$  seleccionamos un candidato  $\beta^*$  según una distribución uniforme en una bola de  $\mathfrak{R}^L$  centrada en  $\beta_{ini}$  y con radio  $l$ . Utilizaremos a  $l$  como parámetro de sintonización. El valor de  $\beta$  en el nuevo paso del algoritmo será  $\beta_{fin} = \beta^*$  con probabilidad de aceptación:

$$P(\text{aceptación}) = \min \left( 1, \frac{f(\beta^*|k, \eta, p, z, \Sigma, x, y, e)}{f(\beta_{ini}|k, \eta, p, z, \Sigma, x, y, e)} \right),$$

o  $\beta_{fin} = \beta_{ini}$  en otro caso. Donde  $f(\beta|k, \eta, p, z, \Sigma, x, y, e)$  es el núcleo de la función de densidad obtenida en la Proposición 15.

En cada etapa del algoritmo de Gibbs proponemos hacer un paso de Metropolis para actualizar  $\beta$ , con lo que el algoritmo es un ejemplo de Metropolis-within-Gibbs.

La distribución condicional completa sobre  $\Sigma$  es una distribución para la que existen métodos de simulación eficientes y bien conocidos.

## 4.2. Situación general: número desconocido de clases

Ahora suponemos que el número de clases o componentes  $k$  es desconocido, fijamos un número máximo de clases  $K$  conocido y al igual que en el Capítulo 2 asignamos a  $k$  una previa discreta sobre los enteros:

$$f(k) = q_k,$$

si  $k = 1, \dots, K$ .

Cuando se introducen covariables, la verosimilitud de los datos supuesto  $k$  conocido viene dada por:

$$\begin{aligned} f(y|k, \eta, p, \beta, e) &\propto \prod_{i=1}^m \sum_{j=1}^k \eta_j^{y_i} \exp\{y_i \beta' x_i\} \exp\{-e_i \eta_j \exp\{\beta' x_i\}\} p_{(j)} \\ &\propto \exp\left\{\left(\sum_{i=1}^m y_i x_i'\right) \beta\right\} \prod_{i=1}^m \sum_{j=1}^k \eta_j^{y_i} \exp\{-e_i \exp\{x_i' \beta\} \eta_j\} p_{(j)}. \end{aligned}$$

Nótese que, supuesto conocido  $\beta$ , la expresión anterior se reduce a la expresión (2.4) del Capítulo 2, sin más que sustituir los valores  $e_i$  que aparecían en aquella por  $e_i \exp\{x_i' \beta\}$ , y aparece un nuevo factor que sólo depende de los parámetros a través del vector  $\beta$ .

**Definición 2.** *La verosimilitud marginal de los datos dado  $k$  es:*

$$f(y|k, x, e) = \int_{\Theta} f(y|k, \eta, p, \beta, x, e) f(\eta, p, \beta|k) d\theta.$$

siendo  $\theta = (\eta', p', \beta)'$ .

Nuestra propuesta, en este caso para obtener una aproximación a la verosimilitud marginal  $\hat{f}(y|k, x, e)$  es utilizar la distribución inicial de  $\beta$  como su función importante y, dado  $\beta$ , construir la función importante sobre los otros parámetros ( $\eta$  y  $p$ ) siguiendo el esquema del apartado anterior.

Una vez obtenida una aproximación a la verosimilitud marginal,  $\hat{f}(y|k, x, e)$ , la distribución final sobre  $k$  la aproximaremos utilizando el Teorema de Bayes:

$$f(k|y, x, e) \propto \hat{f}(y|k, x, e) f(k) \text{ con } k = 1, \dots, K.$$

Por último, la distribución final completa deseada es:

$$f(\eta, p, \beta, k|x, y, e) = f(\eta, p, \beta|x, y, e, k) f(k|x, y, e).$$

Con el primer factor, en la parte derecha de la igualdad anterior, obtenemos la distribución final cuando  $k$  es conocido. Y si nos quedamos con la expresión completa obtenemos la distribución final sobre todos los parámetros, incluido  $k$ .

Al considerar  $k$  desconocido, este parámetro podría tomar el valor de 1, situación que no ha sido considerada antes en este apartado. Bajo esa situación,  $k = 1$ , existe una única clase, trabajamos con un único riesgo latente  $\eta$  y el vector de coeficientes de las variables  $\beta$ , el vector de variables de asignación  $z_i$  y el vector de probabilidades  $p$  no hacen parte de este caso porque existe una única componente.

Así partiendo del supuesto de que:

$$y_i|\eta, \beta, x_i, e_i \sim \text{Po}(y_i|e_i \exp\{x_i'\beta\}\eta).$$

La distribución condicional completa del riesgo  $\eta$  dado los datos  $x, y$  y  $e$  (la cual se deduce de la Proposición 5 del Capítulo 2) es:

$$\begin{aligned} f(\eta|k, \beta, x, y, e) &\propto f(y|k, \eta, \beta, x, e) f(\eta) \\ &\propto \left[ \prod_{i=1}^m \exp\{-e_i \exp\{x_i'\beta\}\eta\} \frac{(e_i \exp\{x_i'\beta\}\eta)^{y_i}}{y_i!} \right] \frac{1}{(1+\eta)^2} \\ &\propto \exp\left\{ \left[ -\sum_{i=1}^m e_i \exp\{x_i'\beta\} \right] \eta \right\} (\exp\{x_i'\beta\}\eta)^{y_i} \frac{1}{(1+\eta)^2} \\ &\propto \exp\left\{ \left[ -\sum_{i=1}^m e_i \exp\{x_i'\beta\} \right] \eta \right\} (\eta)^{y_i} \frac{1}{(1+\eta)^2} \end{aligned}$$

siendo  $0 < \eta < +\infty$ .

Obsérvese que aunque nos enfrentamos a un habitual modelo de Poisson, la inicial asumida para  $\eta$  es  $\frac{1}{(1+\eta)^2}$ , de tal manera que dicha estructura es compatible con la previa asumida para cuando  $k > 1$  explicada en el Capítulo 2. Para el estudio de la distribución condicional de  $\beta$  hacemos uso de las Proposiciones 15 y 16 desarrolladas en este capítulo y para el estudio en una dimensión utilizamos los respectivos corolarios.

### 4.3. Conclusiones sobre cada área pequeña

Teniendo en cuenta toda la información que podemos obtener del análisis realizado y teniendo en cuenta que en este capítulo hemos incluido covariables, nuestro principal interés recae en el estudio de los riesgos de cada una de las áreas pequeñas y la probabilidad que tiene cada una de dichas áreas de pertenecer a una de las clases que hacen parte del modelo.

Supuestos conocidos todos los parámetros del modelo y una vez observados los datos, la probabilidad de que el área pequeña  $a_i$  pertenezca a la clase  $c_j$  es:

$$P(a_i \in c_j | k, \eta, \beta, p, x, y, e) = P(z_{ij} = 1 | k, \eta, \beta, p, x, y, e), \quad (4.4)$$

y esas probabilidades están dadas en la Proposición 11 de este capítulo y se calculan en cada paso del algoritmo de Gibbs.

Así pues, supuestos conocidos todos los parámetros, el riesgo asociado a cada área pequeña, tomando esperanzas respecto a la distribución de  $z$ , es:

$$E(\theta_i | k, \eta, \beta, p, x, y, e) = \sum_{j=1}^k \eta_j \exp\{x'_i \beta\} P(a_i \in c_j | k, \eta, \beta, p, x, y, e), \quad (4.5)$$

con  $i = 1, \dots, m$ .

La estimación de esos riesgos, su esperanza con respecto a la distribución final, dado  $k$ , puede aproximarse utilizando las cadenas de Markov generadas mediante Gibbs sampling,

$$\hat{E}(\theta_i | k, x, y, e) = \frac{1}{N} \sum_{n=1}^N \theta_i^{(n)},$$

siendo  $\theta_i^{(n)}$  el valor obtenido al sustituir en la ecuación (4.5) los resultados de la  $n$ -ésima iteración de Gibbs.

Finalmente, utilizando la distribución posterior sobre el número de clases:

$$\hat{E}(\theta_i|x, y, e) = \sum_{k=1}^K \hat{E}(\theta_i|k, x, y, e) f(k|x, y, e).$$

La expresión anterior es nuestra propuesta como estimación puntual del riesgo de cada área pequeña. Además, si suponemos conocido  $k$ , o utilizamos una estimación puntual de dicho parámetro desconocido, podemos estimar la probabilidad de pertenencia de cada área pequeña a cada una de las clases de la siguiente forma:

Sea  $p_{ij}^{(n)}$  el valor obtenido al sustituir en la ecuación (4.4) los valores de los parámetros obtenidos en la  $n$ -ésima etapa del algoritmo de Gibbs, entonces:

$$\hat{P}(a_i \in c_j|k, x, y, e) = \frac{1}{N} \sum_{n=1}^N p_{ij}^{(n)}.$$

## 4.4. Ejemplos numéricos

En esta sección continuamos estudiando los datos de la varicela en Valencia, los cuales fueron descritos en el ejemplo de aplicación del Capítulo 3. Revisamos tres ejemplos, cada uno relacionado con una covariable. La idea es estudiar si cada una de las covariables de manera individual tiene algún efecto o influencia importante sobre la estimación del riesgo en cada una de las áreas pequeñas. La primera covariable (‘tamaño familiar’) que estudiamos es el número medio de personas por hoja familiar en el año 2008 en la ciudad de Valencia, la segunda covariable (‘distancia al centro’) se refiere a la distancia que hay desde el centroide de cada uno de los barrios al centroide de la ciudad y la última covariable (‘vecinos’) corresponde al riesgo medio de cada área obtenido a partir de los riesgos de los vecinos de primer orden, es decir, en esta tercera covariable incluimos una estructura de vecindad en la estimación del riesgo.

En todos los casos hemos analizado la convergencia estudiando el comportamiento de las cadenas de Markov y las series trazadas para los diferentes parámetros convergen.

### 4.4.1. Análisis de la covariable ‘tamaño familiar’

Esta covariable presenta datos que corresponden al número promedio de personas que habitan en cada uno de los 70 barrios que hacen parte del estudio. En la hoja familiar están anotados el número de personas que residen en un determinado hogar y luego se ha tomado el valor medio del tamaño familiar por

barrio. Para hacernos una idea de los valores que estudiamos en esta covariable presentamos el resumen numérico en la Tabla 4.1.

Tabla 4.1: Resumen de la covariable ‘tamaño familiar’

Mín	Q1	Mediana	Media	Q3	Máx
2.02	2.44	2.52	2.50	2.61	2.84

Iniciamos el análisis de esta covariable estudiando la estimación de los parámetros bajo el modelo propuesto. Presentamos los intervalos de probabilidad al 95 % sobre la distribución posterior de los nuevos parámetros que hacen parte del modelo:  $\beta$  el coeficiente de la covariable y  $\sigma$  su desviación.

Tabla 4.2: Estimaciones, por intervalo al 95 %, de  $\beta$  y  $\sigma$  (‘tamaño familiar’)

$k$	$\beta$	$\sigma$
1	(-0.03,0.31)	(0.05,0.47)
2	(-0.21,0.06)	(0.05,0.35)
3	(-0.04,0.10)	(0.05,0.33)
4	(-0.18,0.09)	(0.05,0.36)
5	(-0.35,0.07)	(0.06,0.49)
6	(-0.24,0.17)	(0.05,0.42)
7	(-0.19,0.06)	(0.05,0.36)

Obsérvese que en la Tabla 4.2 el intervalo de probabilidad al 95 % del parámetro  $\beta$  para cada uno de los diferentes  $k$  considerados incluye el cero, lo que indica que dicho parámetro puede tomar este valor. En este caso el riesgo relativo de cada área pequeña estaría explicado por el riesgo residual.

Al estimar los riesgos de cada una de las áreas pequeñas con el modelo de mixturas que proponemos, obtenemos el resumen que aparece en la Tabla 4.3. Si lo comparamos con la estimación obtenida sin covariables realizada en el ejemplo numérico del Capítulo 2, encontramos que el rango de variación de los riesgos estimados se reduce cuando hemos incluido esta covariable, los riesgos parecen más homogéneos, están más suavizados.

Tabla 4.3: Resumen de la estimación de los riesgos en áreas pequeñas (‘tamaño familiar’)

Mín	Q1	Mediana	Media	Q3	Máx
0.94	0.99	1.00	0.99	1.01	1.05

La Figura 4.1 representa los riesgos estimados para cada uno de los barrios de la ciudad de Valencia incluyendo la covariable ‘tamaño familiar’. Nótese que el riesgo está suavizado de tal manera que lo que podemos observar son dos grupos de riesgo.

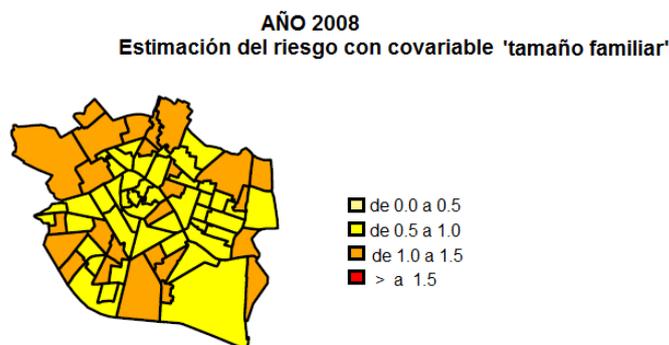


Figura 4.1: Mapa de riesgos estimados para la varicela en el año 2008 con la covariable ‘tamaño familiar’

La covariable ‘tamaño familiar’ suaviza la estimación del riesgo obtenida y al parecer no aporta información adicional.

#### 4.4.2. Análisis de la covariable ‘distancia al centro’

Continuamos ahora con el estudio de la covariable ‘distancia al centro’ que hace referencia a la distancia (euclídea) entre cada uno de los centroides de cada área pequeña al centroide de la ciudad de Valencia. El resumen presentado en la Tabla 4.4 contiene los valores que describen los datos relacionados con esta covariable.

Tabla 4.4: Resumen de la covariable ‘distancia al centro’

Mín	Q1	Mediana	Media	Q3	Máx
0.13	1.59	2.30	2.33	2.98	4.53

La Tabla 4.5 muestra los intervalos de probabilidad al 95 % de las estimaciones de los parámetros que hacen parte del término que mide el efecto de la covariable. Para los diferentes valores de  $k$  encontramos que el parámetro  $\beta$  es claramente positivo y los valores de los intervalos varían poco entre los diferentes valores de  $k$ .

Nuestro parámetro de interés es la estimación de los riesgos para cada uno de los 70 barrios que hacen parte de la zona de estudio. Obsérvese en la Tabla 4.6 que si comparamos el rango de valores en los cuales se mueve el riesgo estimado

Tabla 4.5: Estimaciones, por intervalo al 95 %, de  $\beta$  y  $\sigma$  ('distancia al centro')

$k$	$\beta$	$\sigma$
1	(0.05,0.14)	(0.05,0.37)
2	(0.05,0.14)	(0.05,0.37)
3	(0.03,0.12)	(0.05,0.35)
4	(0.06,0.25)	(0.08,0.55)
5	(0.02,0.28)	(0.06,0.46)
6	(0.03,0.25)	(0.06,0.45)
7	(0.03,0.21)	(0.06,0.39)

cuando incluimos la covariable 'tamaño familiar' en comparación con la covariable 'distancia al centro', encontramos que dicho rango se amplía un poco, dando más variabilidad en los valores de los riesgos relativos estimados. La Figura 4.2 muestra los riesgos estimados incluyendo esta covariable, notar que ahora el mapa muestra el efecto de la covariable, es notable cómo se van incrementando los valores de los riesgos a medida de que el barrio se aleja del centro de la ciudad. Nuevamente las zonas con valores mayores en el riesgo quedan ubicadas en la periferia de Valencia.

Tabla 4.6: Resumen de la estimación de los riesgos en áreas pequeñas ('distancia al centro')

Mín	Q1	Mediana	Media	Q3	Máx
0.70	0.88	0.98	1.00	1.09	1.38

#### 4.4.3. Análisis de la covariable 'vecinos'

Por último, estudiamos la covariable 'vecinos' relacionada con la estructura de vecindad y su impacto en la estimación del riesgo en cada uno de los barrios que conforman la ciudad de Valencia. Teniendo en cuenta las áreas vecinas (las que comparten algún límite geográfico) de cada uno de los barrios y sus riesgos estimados a través de la tasa de mortalidad estandarizada, hemos calculado un riesgo medio para cada barrio teniendo en cuenta el riesgo de cada uno de sus vecinos. Los valores de dicha covariable se resumen en la Tabla 4.7.

Tabla 4.7: Resumen de la covariable 'vecinos'

Min	Q1	Mediana	Media	Q3	Max
0.36	0.71	0.91	0.93	1.07	1.92

Al revisar en la Tabla 4.8 los valores obtenidos para los intervalos de probabilidad al 95 % de cada uno de los parámetros que conforman el término asociado

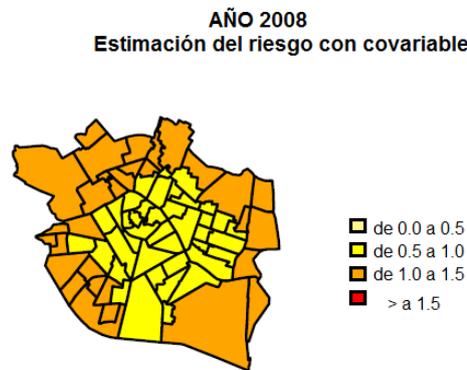


Figura 4.2: Mapa de riesgos estimados para la varicela en el año 2008 con la covariable 'distancia al centro'.

al efecto de la covariable, encontramos que el parámetro relacionado con el coeficiente de la covariable resulta positivo cuando  $k$  toma valores 1 y 2. Para los otros valores de  $k$  dichos intervalos incluyen el cero por lo que podríamos considerarlo como cero.

Tabla 4.8: Estimaciones, por intervalo al 95 %, de  $\beta$  y  $\sigma$  ('vecinos')

$k$	$\beta$	$\sigma$
1	(0.10,0.37)	(0.08,0.63)
2	(0.01,0.43)	(0.06,0.62)
3	(-0.03,0.27)	(0.05,0.44)
4	(-0.03,1.12)	(0.06,1.62)
5	(-0.06,0.50)	(0.05,0.57)
6	(-0.06,0.91)	(0.06,0.93)
7	(-0.09,0.48)	(0.05,0.50)

Haciendo uso de las probabilidades finales de  $k$  y de las estimaciones del riesgo cuando  $k$  es desconocido, hemos estimado los riesgos para cada una de las áreas pequeñas, el resumen de dichas estimaciones aparece en la Tabla 4.9. Ahora los riesgos estimados tienen un rango de variación menor que cuando hacemos el análisis sin covariables. La Figura 4.3 muestra cómo se distribuyen los riesgos estimados en la zona de estudio. El hecho de tener en cuenta para la estimación del riesgo a los vecinos de primer orden suaviza la estimación del riesgo. Sin embargo, con los valores fijados para los intervalos en el mapa podemos distinguir tres clases.

Tabla 4.9: Resumen de la estimación de los riesgos en áreas pequeñas ('vecinos')

Mín	Q1	Mediana	Media	Q3	Máx
0.87	0.95	1.00	1.01	1.04	1.28

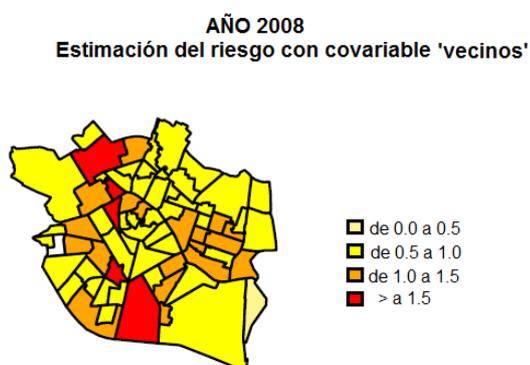


Figura 4.3: Mapa de riesgos estimados para la varicela en el año 2008 con la covariable 'vecinos'.

Cada una de las covariables que hemos incluido en el estudio aporta información que en ocasiones no es relevante y/o significativa para estimar los riesgos. Deseamos que dichos resultados sean un punto de partida para continuar estudiando el efecto de las covariables y por ende la estimación de los riesgos en áreas pequeñas. En un trabajo futuro podemos estudiar el impacto de estas covariables simultáneamente, de tal manera que revisemos el efecto conjunto que tienen sobre la estimación del riesgo en cada una de las áreas pequeñas.

## Capítulo 5

# Conclusiones y futuras líneas de investigación

### 5.1. Conclusiones

En esta tesis presentamos una nueva propuesta para modelizar el riesgo de una enfermedad en áreas pequeñas. En concreto, un modelo de conglomerados con estructura de clases latentes para el análisis bayesiano de datos epidemiológicos en áreas pequeñas. Dicho modelo nos permite obtener estimaciones del riesgo en cada una de las áreas que conforman la zona de estudio, como también estudiar el número de clases o conglomerados que pueden existir en una zona geográfica. La novedad de nuestra propuesta es que a través de este modelo podemos estimar el riesgo y clasificar áreas sin tener en cuenta algún criterio de vecindad. El modelo propuesto incluye ideas de modelos de mixturas, de conglomerados y de modelos ocultos.

El modelo desarrollado en esta tesis permite que se integren los dos objetivos del mapeo de enfermedades: por un lado la estimación del riesgo relativo en cada una de las áreas pequeñas que conforman la zona de estudio y por otra parte la clasificación de áreas con riesgo similar para formar conglomerados. Durante el desarrollo de este trabajo consideramos que hemos profundizado en esta línea de investigación, porque trabajamos simultáneamente en la consecución de estas dos finalidades. Los Capítulos 2, 3 y 4 son interesantes ejemplos del desarrollo en esta temática de investigación.

A través de la revisión de la literatura de mapeo de enfermedades expuestas en el Capítulo 1, hemos introducido la temática y algunas problemáticas abordadas en el transcurso de la tesis, siempre apuntando a los dos objetivos que nos interesan. Con esta introducción presentamos los fundamentos para el desarrollo de nuestra propuesta.

Estudiar este problema desde una perspectiva bayesiana nos ha permitido profundizar y explorar herramientas matemáticas y computacionales para llegar a una solución teórica que a nivel numérico es razonable y a nivel informático permite obtener resultados satisfactorios.

En el Capítulo 2 propusimos un nuevo modelo para la estimación del riesgo y clasificación de áreas con riesgo similar en el ámbito de mapeo de enfermedades. Este modelo aporta una nueva metodología capaz de recuperar las relaciones espaciales existentes entre las observaciones. A nuestro parecer es una formulación novedosa y quizá más flexible que las propuestas con mayor difusión en la literatura, debido a que no estamos forzando a las áreas vecinas a tener un riesgo similar incluyendo una estructura de vecindad.

Con la transformación propuesta en el Capítulo 2 para la previa de los riesgos subyacentes, somos capaces de ordenar dichos riesgos, y de este modo evitar el problema de identificabilidad de las clases. Hemos trabajado con distribuciones previas propias, poco informativas, para obtener la distribución a posteriori para los parámetros del modelo que, aunque no es analíticamente tratable, puede ser estimada mediante métodos de simulación. En concreto, proponemos un algoritmo de Gibbs para simular de las condicionales completas a posteriori.

Hemos diseñado y ejecutado un estudio de simulación de grandes dimensiones, con 246 escenarios y en cada uno de ellos se han analizado 100 bancos de datos. A través de dicho estudio encontramos que somos capaces de recuperar la estructura real del modelo fijado sin incluir desde el inicio algún tipo de dependencia espacial, lo cual se evidencia en los resultados del estudio que se encuentran en el Capítulo 3 y en la aplicación con datos reales ejemplificados en los Capítulos 2 y 4. Las estimaciones de los riesgos y las probabilidades muestran que nuestro modelo tiene la capacidad de recuperar el modelo verdadero, además es posible estimar los riesgos relativos en cada área pequeña haciendo uso del modelo de mixtura presentado en el Capítulo 2.

Dentro del estudio de simulación sucede lo siguiente con los riesgos subyacentes: a medida que los riesgos fijados en cada escenario se acercan uno al otro, la tarea de identificar las clases se vuelve más compleja. Lo mismo sucede cuando vamos aumentando el número de clases  $k$ . Sin embargo, el modelo es capaz de identificar y separar las clases en los ejercicios realizados. El porcentaje de correspondencia de la máxima probabilidad con el verdadero  $k$  fijado, se incrementa cada vez que aumenta la incidencia de la enfermedad, algo que era de esperar pues a mayor incidencia los datos son más informativos.

La distribución final de  $k$  contiene toda la información acerca de la asignación de las áreas pequeñas a cada clase y la ocupación de dichas clases. Para estudiar esta distribución hemos utilizado la integración Monte-Carlo, así tenemos una alternativa diferente al salto reversible (Reversible jump) frecuentemente utilizado para estudiar el número de clases que existen en una zona de estudio. Usando esas probabilidades y los riesgos latentes estimados, podemos obtener un riesgo

relativo estimado para cada una de las áreas pequeñas. La propuesta que realizamos mediante el estudio de las probabilidades finales de  $k$  amplía el uso de eficientes alternativas en simulación bayesiana.

Al medir la bondad de ajuste de los resultados obtenidos con el modelo, encontramos que el error RMSE se ve afectado principalmente por la incidencia de la enfermedad, esto es, a medida que incorporamos más información el error disminuye notablemente. Por otro lado, el impacto de las probabilidades fijadas en el estudio para cada escenario, sobre la estimación de los parámetros no es relevante. Los valores asignados no afectan de manera importante dichas estimaciones.

Hemos realizado una extensión del modelo incluyendo covariables a través de un término asociado a un predictor lineal. De este modo descomponemos el riesgo de cada área pequeña en dos componentes, una que estudia el efecto de la(s) covariable(s) y el otro término que estudia el riesgo residual. Así, este nuevo modelo puede ser visto como un proceso de suavización del riesgo que al mismo tiempo permite estudiar el efecto de una o más covariables. Aunque en el trabajo con covariables es necesaria una profundización, consideramos que el término incluido en esta extensión recoge información que, según sea el caso, será o no relevante en el modelo.

Resulta interesante ver la similaridad obtenida en las estimaciones de los riesgos para cada uno de los barrios en la ciudad de Valencia, cuando trabajamos con los datos de varicela y nos comparamos con los trabajos desarrollados por Iftimi et al. (2015a) e Iftimi et al. (2015b). Podemos decir que el modelo propuesto supone una alternativa capaz de competir con los principales modelos que se han propuesto en la literatura y que trabajan simultáneamente con los dos objetivos de mapeo de enfermedades.

## 5.2. Líneas futuras de investigación

El desarrollo de esta tesis nos ha permitido despejar algunas incógnitas sobre el tema tratado, pero al mismo tiempo, nos ha generado nuevas preguntas, ideas y/o posibles líneas futuras de trabajo en las que seguir investigando. En esta sección se presentan algunas líneas de investigación que pueden ser objeto de interés.

Como primera posibilidad, hemos pensado en incluir una componente espacial en el modelo de conglomerados y así obtener las estimaciones del riesgo en cada una de las áreas pequeñas. Sería de interés comparar los resultados obtenidos con este nuevo modelo con los obtenidos con los modelos propuestos en los Capítulos 2 y 3.

Puede ser útil realizar un análisis de sensibilidad de los resultados. En los Capítulos 2 y 4 proponemos como distribución inicial sobre el parámetro  $p$  una

distribución Dirichlet con hiperparámetro  $\alpha$ . Según hemos podido comprobar el modelo resulta particularmente sensible a este hiperparámetro. Hemos analizado los resultados con valores de  $\alpha = a1$ , siendo 1 el vector de unos y tomando  $a = 1, 2, 3$  y 5. Nosotros nos hemos quedado con el valor 2, que supone un valor intermedio en la variabilidad del parámetro. Partiendo del análisis de las distribuciones previas expuestas en los Capítulos 2 y 4, sería interesante estudiar la posibilidad de trabajar con otras previas y así estudiar la sensibilidad de los parámetros frente a las nuevas propuestas. Por otro lado, resultaría interesante buscar una reparametrización del modelo para reducir el error Monte-Carlo tal como sugieren Salmerón et al. (2015).

Otro tema en el que se podría profundizar es en la modelización espaciotemporal del problema. Lee and Lawson (2014) abordan esta situación introduciendo una componente temporal donde el modelo es capaz de detectar conglomerados dinámicamente; es decir, la relación de pertenencia al conglomerado y su nivel de riesgo promedio evolucionan en el tiempo. Sería interesante trabajar en la inclusión de dicha componente para potenciar el análisis temporal en estas aplicaciones.

También sería importante continuar trabajando en la extensión del modelo incluyendo otras covariables. Las covariables que hemos incluido en el estudio aportan información que en ocasiones no es relevante y/o significativa para estimar los riesgos. Deseamos que dichos resultados sean un punto de partida para continuar estudiando el efecto de las covariables y por ende la estimación de los riesgos en áreas pequeñas. En un trabajo futuro podemos estudiar el impacto de estas covariables simultáneamente, de tal manera que revisemos el efecto conjunto que tienen sobre la estimación del riesgo en cada una de las áreas pequeñas.

Siguiendo con posibles líneas de trabajo y a la vista de los resultados obtenidos en los diferentes capítulos, pensamos que podría resultar interesante compararnos con otros modelos de estimación del riesgo relativo en áreas pequeñas y de detección de conglomerados, por ejemplo, con el modelo de particiones de Knorr-Held and Rasser (2000), el modelo de conglomerados con suavización espacial a través de la distribución CAR propuesto por Lee and Lawson (2014) y el modelo de conglomerados con suavización espacial que también es propuesto en el mismo trabajo.

Otra posible e interesante línea de trabajo sería analizar datos relacionados con otras enfermedades. Sería conveniente estudiar el rendimiento del modelo cuando los datos corresponden a enfermedades raras, donde los casos observados en cada una de las áreas pequeñas pueden tomar el valor de cero, lo que nos llevaría a tener sobredispersión, pues en este tipo de modelos en el primer nivel de jerarquía los datos observados se distribuyen con una Poisson. También resultaría interesante estudiar la estimación del riesgo y la detección de discontinuidades en zonas de estudio de gran tamaño, es decir, compuestas por gran cantidad de áreas pequeñas.

Otro campo de actuación sería el desarrollo de una herramienta informática con el software R, que sea fácil de implementar y que sea una alternativa para analizar un banco de datos, cuando inicialmente no se tiene información de los vecinos.



# Capítulo 6

## Anexos

### A.1 Código para el modelo de conglomerados

```
### Funciones para el análisis de datos

library(gtools)
library(lattice)
library(coda)

##PROCESO GIBBS

rzeta <- function(etarisk,prisk,e,y){
  EXP <- - e*etarisk + y * log(etarisk) + log(prisk)
  pzeta<-exp(EXP - max(EXP))
  pzeta<- pzeta/sum(pzeta)
  psum <- cumsum(pzeta)
  u <- runif(1)
  z <- u<psum & u>c(0,psum[-length(psum)])
  return(list(z=z*1,p=pzeta))}

v.corte <- function(eta,corte){
  corte<-c(0,corte,1)
  eta <- eta/(1+eta)
  vcorte <- NULL
  for (j in 1:(length(eta)-1))
  {
    a <- corte[j]
    A <- eta[j]
    B <- eta[j+1]
```

```

    b <- corte[j+2]
    u <- runif(1)
    Ku <- (A-a)/(b-A)*((B-a)*(b-A)/(b-B)/(A-a))**u
    corte[j+1] <- (a + b*Ku)/(1+Ku)
    vcorte <- c(vcorte,corte[j+1])
  }
  return(vcorte)}

simulador_eta <- function(alfa,beta2,corte1,corte2){
  if (alfa<3 | alfa<beta2){
    x <- corte1
    u <- 2
    while (u > (1+corte1)^2/(1+x)^2){
      x <- -1
      while (!(x>corte1 & x<corte2)) x <- rgamma(1,alfa,beta2)
      u <- runif(1)}
  }
  else{
    a<-pgamma(corte1, alfa-2,beta2)
    b<- pgamma(corte2, alfa-2,beta2)
    if (b-a > 0.01){
      x <- corte2
      u <- 2
      while (u > (1+1/corte2)^2/(1/x+1)^2){
        x <- -1
        while (!(x>corte1 & x<corte2)) x <- rgamma(1,alfa-2,beta2)
        u <- runif(1)}
    }
    else {
      x <- corte2
      u <- 2
      while (u > (1+1/corte2)^2/(1/x+1)^2){
        x <- qgamma(runif(1,a,b),alfa-2,beta2)
        if(x==Inf) x <- corte1
        u <- runif(1)}}
  }
  return(x)
}

eta.riesgo<- function(zetanueva, y, e,corte){

  # parámetros de las gamma
  alfas <- matrix(y,nrow=1) %*% zetanueva + 1
  betas <- matrix(e,nrow=1) %*% zetanueva

```

```

# Simulación
corte <- c(0,corte,1)/(1-c(0,corte,1))
etanuevo <- NULL
for (i in 1:length(alfas)){
  if (betas[i]==0) x <- (1/(1+corte[i])-(1/(1+corte[i])-
    1/(1+corte[i+1])))*runif(1)**(-1)-1
  if (betas[i]>0) x <- simulador_eta(alfas[i],betas[i],
    corte[i],corte[i+1])
  etanuevo <- c(etanuevo,x)}
return(etanuevo) }

beta_metropolis <- function(eta,zeta,beta_ini,l,np,sigma_ini,
  datos){
  y <- datos$y
  e <- datos$e
  x <- datos$x
  salto = T
  for (i in 1:np) {
    beta_fin <- beta_ini -l + runif(1)*2*l
    A <- (beta_fin^2 - beta_ini^2)/2/sigma_ini^2 -
      (beta_fin - beta_ini) * sum(x*y)
    B <- sum((zeta %*% eta) * e * (exp(x*beta_fin) -
      exp(x*beta_ini)))
    u <- runif(1)
    if (A + B > -log(u)) {
      beta_fin <- beta_ini
      salto = F}
  }
  return(list(beta=beta_fin,salto=salto))
}

beta_metropolis_k_1 <- function(eta.ini,beta.ini,l,np,sigma.ini,
  datos){
  y <- datos$y
  e <- datos$e
  x <- datos$x
  for (i in 1:np) {
    beta_fin <- beta.ini -l + runif(1)*2*l
    A <- (beta_fin^2 - beta.ini^2)/2/sigma.ini^2 - (beta_fin -
      beta.ini)* sum(x*y)
    B <- sum( eta.ini * e * (exp(x*beta_fin) - exp(x*beta.ini)))
    u <- runif(1)
    if (A + B > -log(u)) beta_fin <- beta.ini
  }
}

```

```

    }
    return(beta_fin)
}

sigma_nuevo <- function(beta_nuevo,sigma_ini,a,b,cmin){
  cota <- pgamma(cmin,a+(1/2),(beta_nuevo^2/2)+b)
  u <- runif(1,cota,1)
  prec <- qgamma(u,a+(1/2),(beta_nuevo^2/2)+b)
  return(1/sqrt(prec))
}

unaetapa <- function(ini,viejos,par.metropolis,datos) {
  y <- datos$y
  e <- datos$e
  x <- datos$x
  m <- length(y)

  # Parámetros de las distribuciones iniciales
  alfa <- ini$p_v$alfas
  a <- ini$beta$a
  b <- ini$beta$b
  cmin <- ini$beta$cmin
  l <- par.metropolis$l
  np <- par.metropolis$np

  # Punto de arranque de esta etapa del algoritmo de Gibbs
  eta <- viejos$eta_start
  corte <- viejos$v_start
  beta_ini <- viejos$beta_start
  sigma_ini <- viejos$sigma_start
  pinicial <- viejos$p_start

  ### simulamos Z
  znuevo <- NULL
  pmat <- NULL
  media <- NULL
  e_cov <- e*exp(x*beta_ini)
  for (i in 1:m) {
    zn <- rzeta(eta,pinicial,e_cov[i],y[i])
    pmat <- rbind(pmat,zn$p)
    media <- c(media,sum(zn$p*eta))
    znuevo <- rbind(znuevo,zn$z)
  }
}

```

```
media<- media*exp(x*beta_ini)

## Simulamos P
zetacol <-apply(znuevo,2,sum)
pnuevo <-rdirichlet(1,zetacol+alfa)

## Simulamos V
vnuevo <- v.corte(eta,corte)

## Simulamos eta
etanuevo <- eta.riesgo(znuevo,y,e_cov,vnuevo)
if (length(datos$x)>1) {
  betafin <- beta_metropolis(etanuevo,znuevo,beta_ini,l,np,
                             sigma_ini,datos)
  beta_ini <- betafin$beta
  beta_salto <- betafin$salto
  sigma_ini <-sigma_nuevo(beta_ini,sigma_ini,a,b,cmin)
}

## Salida
sal <- list(z=znuevo,p=pnuevo,v=vnuevo,eta=etanuevo,
           areas=media,pmat=pmat,beta_est=beta_ini,
           sigma_est=sigma_ini)
return(sal) }

gibbs_k_1 <- function(Nsim,ini,punto,par.metropolis,datos){
  y <- datos$y
  e <- datos$e
  x <- datos$x
  a <- ini$beta$a
  b <- ini$beta$b
  cmin <- ini$beta$cmin
  l <- par.metropolis$l
  np <- par.metropolis$np
  m <- length(y)
  eta.ini <- punto$eta_start
  beta.ini <- punto$beta_start
  sigma.ini <- punto$sigma_start
  corte1 <- 0
  corte2 <- Inf
  sum_e <- sum(e)
  alfa <- sum(y)+1
  eta_k_1 <- beta_k_1 <- sigma_k_1 <- media <- NULL
  for (ii in 1:Nsim){
```

```

beta2 <- sum(e*exp(x*beta.ini))
eta.ini <- simulador_eta(alfa,beta2,corte1,corte2)
med <- rep(eta.ini,m)
eta_k_1 <- c(eta_k_1,eta.ini)
if(length(x)>1) {
  beta.ini <- beta_metropolis_k_1(eta.ini,beta.ini,l,np,
                                sigma.ini,datos)
  sigma.ini <-sigma_nuevo(beta.ini,sigma.ini,ini.beta$a,
                        ini.beta$b,ini.beta$cmin)
  med <- exp(beta.ini*x)*eta.ini
  beta_k_1 <- c(beta_k_1,beta.ini)
  sigma_k_1 <- c(sigma_k_1,sigma.ini)}
media <- rbind(media,med)}
return(list(eta=eta_k_1,beta=beta_k_1,sigma=sigma_k_1,
          areas=media))
}

```

```

gibbs_k_conocida <- function(Nsim,ini,punto,par.metropolis,
                             datos) {
  if (length(punto$eta_start)>1) {
    viejos <- punto

    # Inicialización de los objetos de salida
    zsim <- vsim <- etasim <- psim <- areas <- NULL
    pmatsim <- 0
    betasim <- NULL
    sigmasim <- NULL

    # Algoritmo de Gibbs
    for (j in 1:Nsim){

      nuevospara<- unaetapa(ini,viejos,par.metropolis,datos)

      znueva <- nuevospara$z
      pnuevo <- nuevospara$p
      etanuevo<-nuevospara$eta
      vnuevo<- nuevospara$v
      betanuevo <- nuevospara$beta_est
      sigmanuevo <- nuevospara$sigma_est

      znueva <- znueva %*% matrix(1:k,ncol=1)

      zsim <- rbind(zsim,as.vector(znueva))
    }
  }
}

```

```

etasim <- rbind(etasim,etanuevo)
psim <- rbind(psim,pnuevo)
vsim <- rbind(vsim,vnuevo)
if (length(datos$x)>0) {
  betasim <- c(betasim,betanuevo)
  sigmasim <-c(sigmasim,sigmanuevo)
}
areas <- rbind(areas,nuevospara$areas)
pmatsim <- pmatsim + nuevospara$pmat

viejos <- list(p_start=pnuevo,v_start=vnuevo,
              eta_start=etanuevo,beta_start=betanuevo,
              sigma_start=sigmanuevo)
}
salidafinal<-list(eta=etasim,p=psim,v=vsim,z=zs,
                 pmat=pmatsim/Nsim,areas=areas,beta=betasim,
                 sigma=sigmasim)
}
if (length(punto$eta_start)==1) salidafinal <-
  gibbs_k_1(Nsim,ini,punto,par.metropolis,datos)
return(salidafinal)
}

```

#FUNCIONES PARA EL CALCULO DE LA DISTRIBUCIÓN FINAL DE K

# Función que calcula el logaritmo de la verosimilitud en  
#un punto dado

```

lvero.un.punto <- function(theta,datos,k) {
  e <- datos$e
  y <- datos$y
  x <- datos$x
  eta <- theta[1:k]
  p <- theta[k + 1:k]
  bet <- 0
  if (length(datos$x) > 1) bet <- theta[2*k+1]
  ee <- e*exp(bet*x)
  matr <- matrix(y,ncol=1) %*% matrix(log(eta),nrow=1) -
    matrix(ee,ncol=1) %*% matrix(eta,nrow=1)
  maxim <- apply(matr,1,max)
  matr <- matr - maxim
  vero <- exp(matr)
}

```

```

lvero <- sum(log(vero %*% matrix(p,ncol=1))) +sum(maxim)
+ bet*sum(x*y)
return(lvero)
}

```

# Función que simula de la función importante  $g(\nu|d)$

```

fun_nus <- function(adn,k){
  a <- adn[1:k]
  d <- cumsum(adn[k + 1:(k-1)])
  nus <- adn[-(1:(2*k-1))]
  nus <- sort(c(nus,d))
  xx <- cbind(c(0,nus),c(nus,1) - c(0,nus))
  u <- runif(k) * (a+1) + c(0,cumsum(a[-k])) + (0:(k-1))
  ui <- ceiling(u)
  uu <- 1+u-ui
  nus <- xx[ui,1] + uu*xx[ui,2]
  log_nus <- sum(-log(xx[ui,2])-log(a+1))
  return(list(nus=nus/(1-nus),log_nus=log_nus))
}

```

# Función que calcula los estimadores naive del riesgo,  
#los ordena y evita empates

```

fun_naive_nus <- function(e,y) {
  nus <- y/(y+e)
  nus <- sort((abs(jitter(nus))))
  if (min(nus)<=0) nus <- nus + min(nus) + 1.0E-06
  if (max(nus)>=1) nus <- nus / max(nus) - 1.0E-07
  return(nus)
}

```

```

final_k <- function(datos,previas,N){
  K <- length(previas$k)
  y <- datos$y
  e <- datos$e
  x <- datos$x
  e_betas_x <- matrix(rep(e,N),ncol = N)
  betas <- NULL
  if (length(x)>1) {
    cota <- pgamma(previas$beta$cmin,previas$beta$a,
                  previas$beta$b)
    u <- runif(N,cota,1)

```

---

```

    betas <- rnorm(N) /sqrt(qgamma(u,previas$beta$a,
                                previas$beta$b))
    e_betas_x <- e*exp(outer(x,betas))
  }
nus_naives <- apply(e_betas_x,2,fun_naive_nus,y=y)
const <- NULL
vero_mean <- NULL
vero_se <- NULL

# Para k == 1

e_sum <- apply(e_betas_x,2,sum)
constante <- min(e_sum)
etas <- rgamma(N,sum(y)+1,e_sum)
const <- cbind(const, - (sum(y)+1) * log(constante) +
               lgamma(sum(y)+1))
valores <- (constante/e_sum)**(sum(y)+1)/(1+etas)**2
vero_mean <- cbind(vero_mean,mean(valores))
vero_se <- cbind(vero_se,sd(valores)/sqrt(N))

# Para los k > 1

for (k in 2:K) {
  gammas <- previas$p_v[[k]]$gammas
  alfas <- previas$p_v[[k]]$alfas

  # Simulamos de g(d), que coincide con la previa de d
  ds <- rdirichlet(N,gammas)

  # Simulamos de g(p|d) y calculamos su log_densidad
  #en esos puntos
  As <- apply(cbind(ds,t(nus_naives)),1,fun <-function(d)
             apply(outer(d[-(1:k)],cumsum(d[1:k]), FUN='<'),2,sum))
  As <- t(rbind(As[1,],As[-1,]-As[-k,]) + alfas)
  ps <- t(apply(As,1,rdirichlet,n=1))
  log_den_p <- apply(((As - 1) * log(ps)),1,sum) -
  apply(lgamma(As),1,sum) + lgamma(apply(As,1,sum))
  As <- t(t(As) - alfas)

  # Simulamos g(nu_i|d), para i=1,...,k
  nus <- matrix(unlist(apply(cbind(As,ds[, -k],t(nus_naives)),
                          1,fun_nus,k=k)),ncol=k+1,byrow=T)
  log_den_nus <- nus[,k+1]
  nus <- nus[, -(k+1)]

```

```

# Calculamos el logaritmo de la función a integrar en
# todos los puntos
lvero <- apply(cbind(nus,ps,betas),1,lvero.un.punto,
              datos=datos,k=k)
lvero <- lvero -log_den_p -log_den_nus - apply(log(ds),
        1,sum)
lvero <- lvero + apply((alfas - 1) * t(log(ps)),2,sum) -
sum(lgamma(alfas)) + lgamma(sum(alfas))
lvero_max <- max(lvero)
const <- cbind(const,lvero_max)
vero_mean <- cbind(vero_mean,mean(exp(lvero-lvero_max)))
vero_se <- cbind(vero_se,sd(exp(lvero-lvero_max))/
                sqrt(length(lvero)))
}
constante <- max(const)
const <- exp(const - constante)
vero_mean <- vero_mean * const
vero_se <- (vero_se * const)

prob_finales_k<- t(vero_mean)*previas$k
suma <- apply(prob_finales_k,2,sum)
prob_finales_k<- t(prob_finales_k )/ suma

return(list(const = constante, vero_mean=vero_mean,
            vero_se=vero_se, prob_finales_k=prob_finales_k))
}

```

## A.2 Estimaciones de riesgos latentes y probabilidades

Presentamos las tablas correspondientes a las estimaciones de los riesgos latentes y probabilidades bajo el modelo propuesto. Los valores mostrados corresponden a la mediana y el rango intercuartílico de las 100 réplicas estudiadas para cada uno de los 246 escenarios. Si bien es cierto que las estimaciones bajo el modelo propuesto recuperan bastante bien el modelo verdadero en la mayoría de los escenarios, algunas estimaciones se ajustan bastante menos. En el capítulo 4 se analizaron simulaciones de algunas réplicas en las cuales el valor estimado del riesgo latente se disparaba a valores muy altos o tomaba valores muy bajos, sin embargo se comprobó que dichos valores están asociados a probabilidades que son prácticamente cero, mostrando así que existe una posibilidad casi nula de que un riesgo tome valores extremadamente altos o bajos. También se hicieron los análisis de convergencia correspondientes a cada una de las diferentes trazas.

Tabla 6.1: Estimaciones de los riesgos latentes y probabilidades  $k = 2$ . Incidencia  $3/10^4$ hab.

Riesgos		Probabilidades	
0.5 0.49(0.09)	3.5 3.52(0.26)	0.5 0.51(0.09)	0.5 0.49(0.09)
0.5 0.51(0.14)	3.5 3.52(0.17)	0.2 0.21(0.05)	0.8 0.79(0.05)
0.5 0.51(0.08)	3.5 3.56(0.37)	0.8 0.79(0.06)	0.2 0.21(0.06)
0.5 0.48(0.22)	3.5 3.53(0.18)	0.1 0.11(0.05)	0.9 0.89(0.05)
0.5 0.49(0.11)	2.5 2.51(0.20)	0.5 0.49(0.10)	0.5 0.51(0.10)
0.5 0.48(0.14)	2.5 2.52(0.16)	0.2 0.20(0.07)	0.8 0.80(0.07)
0.5 0.50(0.09)	2.5 2.48(0.46)	0.8 0.79(0.06)	0.2 0.21(0.06)
0.5 0.45(0.27)	2.5 2.53(0.19)	0.1 0.11(0.07)	0.9 0.89(0.07)

Tabla 6.2: Estimaciones de los riesgos latentes y probabilidades  $k = 2$ . Incidencia  $3/10^4$ hab. Continuación.

Riesgos		Probabilidades	
0.5 0.49(0.15)	1.5 1.57(0.25)	0.5 0.49(0.14)	0.5 0.51(0.14)
0.5 0.47(0.25)	1.5 1.72(3.44)	0.2 0.26(0.17)	0.8 0.75(0.17)
0.5 0.43(0.09)	1.5 1.63(0.66)	0.8 0.69(0.21)	0.2 0.31(0.21)
0.5 0.53(0.26)	1.5 4.36(14.32)	0.1 0.22(0.18)	0.9 0.78(0.18)
0.8 0.56(0.27)	1.2 13.55(24.28)	0.5 0.43(0.27)	0.5 0.57(0.27)
0.8 0.63(0.28)	1.2 16.25(25.27)	0.2 0.41(0.29)	0.8 0.59(0.29)
0.8 0.51(0.19)	1.2 14.11(23.55)	0.8 0.46(0.25)	0.2 0.54(0.25)
0.8 0.62(0.26)	1.2 22.62(34.18)	0.1 0.40(0.26)	0.9 0.60(0.26)

Tabla 6.3: Estimaciones de los riesgos latentes y probabilidades  $k = 2$ . Incidencia  $30/10^4\text{hab}$ .

Riesgos		Probabilidades	
0.5 0.50(0.03)	3.5 3.51(0.08)	0.5 0.50(0.07)	0.5 0.51(0.07)
0.5 0.50(0.04)	3.5 3.5(0.05)	0.2 0.21(0.05)	0.8 0.79(0.05)
0.5 0.50(0.02)	3.5 3.51(0.12)	0.8 0.79(0.06)	0.2 0.21(0.06)
0.5 0.50(0.06)	3.5 3.50(0.05)	0.1 0.11(0.05)	0.9 0.90(0.05)
0.5 0.50(0.02)	2.5 2.5(0.08)	0.5 0.50(0.06)	0.5 0.50(0.06)
0.5 0.50(0.04)	2.5 2.50(0.06)	0.2 0.21(0.05)	0.8 0.79(0.05)
0.5 0.50(0.02)	2.5 2.50(0.08)	0.8 0.79(0.05)	0.2 0.21(0.05)
0.5 0.49(0.06)	2.5 2.50(0.06)	0.1 0.11(0.05)	0.9 0.90(0.05)

Tabla 6.4: Estimaciones de los riesgos latentes y probabilidades  $k = 2$ . Incidencia  $30/10^4$ hab. Continuación.

Riesgos		Probabilidades	
0.5 0.50(0.03)	1.5 1.5(0.05)	0.5 0.51(0.07)	0.5 0.50(0.07)
0.5 0.50(0.05)	1.5 1.51(0.05)	0.2 0.21(0.06)	0.8 0.79(0.06)
0.5 0.50(0.02)	1.5 1.52(0.07)	0.8 0.80(0.05)	0.2 0.20(0.05)
0.5 0.52(0.08)	1.5 1.55(1.62)	0.1 0.12(0.05)	0.9 0.88(0.05)
0.8 0.79(0.05)	1.2 1.26(0.49)	0.5 0.50(0.11)	0.5 0.50(0.11)
0.8 0.80(0.07)	1.2 2.93(8.90)	0.2 0.28(0.12)	0.8 0.73(0.12)
0.8 0.79(0.03)	1.2 1.31(0.53)	0.8 0.78(0.13)	0.2 0.22(0.13)
0.8 0.80(0.21)	1.2 12.01(40.99)	0.1 0.27(0.34)	0.9 0.73(0.34)

Tabla 6.5: Estimaciones de los riesgos latentes y probabilidades  $k = 2$ . Incidencia  $300/10^4$ hab.

Riesgos		Probabilidades	
0.5 0.50(0.01)	3.5 3.50(0.02)	0.5 0.50(0.07)	0.5 0.51(0.07)
0.5 0.50(0.01)	3.5 3.50(0.02)	0.2 0.21(0.05)	0.8 0.79(0.05)
0.5 0.50(0.00)	3.5 3.50(0.04)	0.8 0.79(0.05)	0.2 0.21(0.05)
0.5 0.50(0.02)	3.5 3.50(0.02)	0.1 0.11(0.04)	0.9 0.89(0.04)
0.5 0.50(0.01)	2.5 2.50(0.02)	0.5 0.50(0.07)	0.5 0.50(0.07)
0.5 0.50(0.02)	2.5 2.50(0.02)	0.2 0.21(0.05)	0.8 0.79(0.05)
0.5 0.50(0.00)	2.5 2.5(0.03)	0.8 0.79(0.05)	0.2 0.21(0.05)
0.5 0.50(0.02)	2.5 2.50(0.02)	0.1 0.11(0.05)	0.9 0.90(0.05)

Tabla 6.6: Estimaciones de los riesgos latentes y probabilidades  $k = 2$ . Incidencia  $300/10^4$ hab. Continuación.

Riesgos		Probabilidades	
0.5 0.50(0.00)	1.5 1.50(0.03)	0.5 0.50(0.08)	0.5 0.51(0.08)
0.5 0.50(0.01)	1.5 1.52(0.08)	0.2 0.21(0.06)	0.8 0.79(0.06)
0.5 0.50(0.00)	1.5 1.50(0.03)	0.8 0.79(0.05)	0.2 0.21(0.05)
0.5 0.51(0.03)	1.5 1.85(2.15)	0.1 0.12(0.05)	0.9 0.88(0.05)
0.8 0.80(0.02)	1.2 1.22(0.19)	0.5 0.49(0.07)	0.5 0.51(0.07)
0.8 0.81(0.02)	1.2 1.55(1.88)	0.2 0.23(0.05)	0.8 0.77(0.05)
0.8 0.80(0.00)	1.2 1.21(0.07)	0.8 0.80(0.06)	0.2 0.21(0.06)
0.8 0.83(0.06)	1.2 6.09(29.08)	0.1 0.16(0.10)	0.9 0.84(0.10)

Tabla 6.7: Estimaciones de los riesgos latentes y probabilidades  $k = 3$ . Incidencia  $3/10^4$ hab.

Riesgos			Probabilidades		
0.5 0.35(0.19)	1.0 1.12(0.67)	3.5 8.13(39.41)	0.33 0.24(0.18)	0.33 0.43(0.10)	0.33 0.28(0.13)
0.5 0.34(0.13)	1.0 1.21(0.57)	3.5 4.74(42.24)	0.2 0.15(0.10)	0.3 0.39(0.08)	0.5 0.45(0.15)
0.5 0.39(0.16)	1.0 1.30(0.46)	3.5 21.45(37.81)	0.2 0.26(0.19)	0.7 0.62(0.21)	0.1 0.10(0.06)
0.5 0.28(0.13)	1.0 1.14(0.58)	3.5 3.68(31.72)	0.1 0.07(0.04)	0.1 0.18(0.17)	0.8 0.75(0.19)
0.5 0.32(0.12)	1.0 1.16(0.50)	3.5 15.93(49.29)	0.6 0.34(0.19)	0.2 0.47(0.16)	0.2 0.18(0.05)
0.5 0.35(0.13)	1.0 1.09(0.26)	2.0 25.61(40.32)	0.33 0.23(0.12)	0.33 0.47(0.10)	0.33 0.26(0.15)
0.5 0.36(0.17)	1.0 1.19(0.35)	2.0 37.16(53.39)	0.2 0.17(0.13)	0.3 0.45(0.08)	0.5 0.38(0.19)
0.5 0.34(0.15)	1.0 1.03(0.34)	2.0 42.99(51.07)	0.2 0.21(0.14)	0.7 0.55(0.13)	0.1 0.20(0.16)
0.5 0.32(0.14)	1.0 1.29(0.51)	2.0 40.15(87.28)	0.1 0.09(0.05)	0.1 0.44(0.28)	0.8 0.47(0.29)
0.5 0.30(0.11)	1.0 1.01(0.28)	2.0 29.85(48.35)	0.6 0.33(0.15)	0.2 0.47(0.11)	0.2 0.16(0.09)

Tabla 6.8: Estimaciones de los riesgos latentes y probabilidades  $k = 3$ . Incidencia  $3/10^4$ hab. Continuación.

Riesgos			Probabilidades		
0.8 0.36(0.15)	1.0 1.17(0.30)	1.5 61.22(75.81)	0.33 0.17(0.13)	0.33 0.58(0.17)	0.33 0.22(0.18)
0.8 0.31(0.14)	1.0 1.17(0.33)	1.5 66.0(100.89)	0.2 0.11(0.12)	0.3 0.57(0.24)	0.5 0.28(0.28)
0.8 0.34(0.19)	1.0 1.20(0.73)	1.5 91.87(121.65)	0.2 0.19(0.21)	0.7 0.59(0.21)	0.1 0.16(0.26)
0.8 0.34(0.15)	1.0 1.24(0.39)	1.5 69.94(100.53)	0.1 0.10(0.12)	0.1 0.55(0.21)	0.8 0.32(0.26)
0.8 0.33(0.14)	1.0 1.09(0.35)	1.5 65.01(79.78)	0.6 0.19(0.15)	0.2 0.56(0.18)	0.2 0.20(0.17)
1.0 0.49(0.26)	1.5 1.54(0.51)	3.5 8.25(39.75)	0.33 0.16(0.19)	0.33 0.51(0.12)	0.33 0.31(0.13)
1.0 0.49(0.30)	1.5 1.62(0.56)	3.5 8.81(55.06)	0.2 0.10(0.12)	0.3 0.43(0.06)	0.5 0.46(0.15)
1.0 0.51(0.34)	1.5 1.63(0.49)	3.5 29.03(62.41)	0.2 0.18(0.23)	0.7 0.64(0.21)	0.1 0.12(0.11)
1.0 0.39(0.16)	1.5 1.48(0.63)	3.5 3.73(48.26)	0.1 0.04(0.04)	0.1 0.24(0.17)	0.8 0.71(0.20)
1.0 0.51(0.28)	1.5 1.52(0.64)	3.5 24.88(46.20)	0.6 0.25(0.23)	0.2 0.56(0.16)	0.2 0.17(0.09)

Tabla 6.9: Estimaciones de los riesgos latentes y probabilidades  $k = 3$ . Incidencia  $30/10^4$ hab.

Riesgos			Probabilidades		
0.5 0.50(0.04)	1.0 1.00(0.06)	3.5 3.50(0.06)	0.33 0.32(0.06)	0.33 0.35(0.08)	0.33 0.33(0.07)
0.5 0.49(0.05)	1.0 1.00(0.05)	3.5 3.51(0.09)	0.2 0.20(0.07)	0.3 0.30(0.08)	0.5 0.50(0.08)
0.5 0.49(0.06)	1.0 1.01(0.04)	3.5 3.53(0.21)	0.2 0.20(0.07)	0.7 0.69(0.09)	0.1 0.10(0.04)
0.5 0.48(0.06)	1.0 1.02(0.11)	3.5 3.51(0.06)	0.1 0.10(0.04)	0.1 0.11(0.05)	0.8 0.79(0.06)
0.5 0.50(0.03)	1.0 1.00(0.07)	3.5 3.50(0.12)	0.6 0.61(0.07)	0.2 0.19(0.06)	0.2 0.20(0.06)
0.5 0.54(0.14)	1.0 1.13(0.42)	2.0 25.01(101.62)	0.33 0.39(0.17)	0.33 0.35(0.09)	0.33 0.26(0.16)
0.5 0.69(0.24)	1.0 1.72(0.81)	2.0 114.40(207.07)	0.2 0.38(0.22)	0.3 0.43(0.18)	0.5 0.15(0.42)
0.5 0.55(0.10)	1.0 1.10(0.27)	2.0 11.25(46.66)	0.2 0.29(0.16)	0.7 0.62(0.19)	0.1 0.09(0.06)
0.5 0.61(0.18)	1.0 1.93(0.75)	2.0 125.85(180.26)	0.1 0.16(0.09)	0.1 0.73(0.50)	0.8 0.02(0.58)
0.5 0.50(0.03)	1.0 1.05(0.13)	2.0 4.02(21.3)	0.6 0.62(0.10)	0.2 0.21(0.08)	0.2 0.18(0.08)

Tabla 6.10: Estimaciones de los riesgos latentes y probabilidades  $k = 3$ . Incidencia  $30/10^4$ hab. Continuación.

Riesgos			Probabilidades		
0.8 0.71(0.33)	1.0 1.25(0.35)	1.5 73.33(109.26)	0.33 0.44(0.32)	0.33 0.44(0.10)	0.33 0.14(0.23)
0.8 0.84(0.26)	1.0 1.42(0.29)	1.5 107.85(90.19)	0.2 0.39(0.22)	0.3 0.49(0.08)	0.5 0.06(0.24)
0.8 0.71(0.35)	1.0 1.29(0.38)	1.5 95.72(129.58)	0.2 0.55(0.48)	0.7 0.41(0.32)	0.1 0.05(0.10)
0.8 0.79(0.21)	1.0 1.47(0.19)	1.5 125.61(144.36)	0.1 0.16(0.08)	0.1 0.75(0.24)	0.8 0.01(0.25)
0.8 0.69(0.32)	1.0 1.26(0.37)	1.5 87.03(102.75)	0.6 0.54(0.43)	0.2 0.39(0.26)	0.2 0.07(0.13)
1.0 0.99(0.06)	1.5 1.51(0.08)	3.5 3.50(0.08)	0.33 0.33(0.08)	0.33 0.34(0.09)	0.33 0.33(0.06)
1.0 0.96(0.09)	1.5 1.51(0.08)	3.5 3.50(0.07)	0.2 0.18(0.08)	0.3 0.31(0.10)	0.5 0.50(0.08)
1.0 0.98(0.10)	1.5 1.50(0.06)	3.5 3.49(0.16)	0.2 0.19(0.09)	0.7 0.71(0.11)	0.1 0.10(0.04)
1.0 0.81(0.33)	1.5 1.46(0.14)	3.5 3.50(0.05)	0.1 0.07(0.07)	0.1 0.13(0.06)	0.8 0.79(0.06)
1.0 0.99(0.04)	1.5 1.53(0.09)	3.5 3.50(0.11)	0.6 0.60(0.10)	0.2 0.21(0.07)	0.2 0.20(0.06)

Tabla 6.11: Estimaciones de los riesgos latentes y probabilidades  $k = 3$ . Incidencia  $300/10^4\text{hab.}$

Riesgos			Probabilidades		
0.5 0.50(0.02)	1.0 1.00(0.02)	3.5 3.50(0.02)	0.33 0.32(0.06)	0.33 0.34(0.06)	0.33 0.33(0.07)
0.5 0.50(0.01)	1.0 1.00(0.02)	3.5 3.50(0.02)	0.2 0.20(0.05)	0.3 0.29(0.07)	0.5 0.50(0.08)
0.5 0.50(0.02)	1.0 1.00(0.02)	3.5 3.51(0.06)	0.2 0.20(0.05)	0.7 0.69(0.07)	0.1 0.10(0.04)
0.5 0.50(0.01)	1.0 1.01(0.03)	3.5 3.50(0.01)	0.1 0.11(0.05)	0.1 0.10(0.04)	0.8 0.80(0.06)
0.5 0.50(0.00)	1.0 1.00(0.02)	3.5 3.50(0.05)	0.6 0.60(0.07)	0.2 0.20(0.05)	0.2 0.20(0.06)
0.5 0.75(0.07)	1.0 1.99(0.03)	2.0 193.52(197.65)	0.33 0.66(0.08)	0.33 0.33(0.06)	0.33 0.00(0.00)
0.5 0.80(0.06)	1.0 2.00(0.02)	2.0 183.09(151.36)	0.2 0.49(0.08)	0.3 0.50(0.07)	0.5 0.00(0.00)
0.5 0.87(0.13)	1.0 1.97(0.33)	2.0 156.33(159.32)	0.2 0.87(0.21)	0.7 0.13(0.19)	0.1 0.00(0.03)
0.5 0.73(0.13)	1.0 2.00(0.01)	2.0 174.12(104.50)	0.1 0.20(0.05)	0.1 0.79(0.05)	0.8 0.00(0.00)
0.5 0.61(0.08)	1.0 1.98(0.44)	2.0 160.87(136.32)	0.6 0.76(0.09)	0.2 0.21(0.05)	0.2 0.00(0.06)

Tabla 6.12: Estimaciones de los riesgos latentes y probabilidades  $k = 3$ . Incidencia  $300/10^4$ hab. Continuación.

Riesgos			Probabilidades		
0.8 0.90(0.02)	1.0 1.50(0.02)	1.5 157.13(114.62)	0.33 0.66(0.07)	0.33 0.34(0.06)	0.33 0.00(0.00)
0.8 0.92(0.03)	1.0 1.50(0.02)	1.5 157.73(145.40)	0.2 0.48(0.08)	0.3 0.49(0.08)	0.5 0.00(0.00)
0.8 0.95(0.04)	1.0 1.50(0.13)	1.5 134.90(131.90)	0.2 0.87(0.15)	0.7 0.13(0.13)	0.1 0.00(0.02)
0.8 0.89(0.06)	1.0 1.50(0.01)	1.5 164.54(115.71)	0.1 0.20(0.05)	0.1 0.80(0.06)	0.8 0.00(0.00)
0.8 0.84(0.03)	1.0 1.49(0.07)	1.5 127.50(110.73)	0.6 0.76(0.06)	0.2 0.21(0.06)	0.2 0.00(0.03)
1.0 1.00(0.02)	1.5 1.50(0.02)	3.5 3.50(0.03)	0.33 0.32(0.07)	0.33 0.34(0.07)	0.33 0.33(0.07)
1.0 1.00(0.02)	1.5 1.50(0.02)	3.5 3.50(0.02)	0.2 0.20(0.05)	0.3 0.30(0.08)	0.5 0.50(0.08)
1.0 1.00(0.02)	1.5 1.50(0.02)	3.5 3.50(0.05)	0.2 0.20(0.05)	0.7 0.69(0.07)	0.1 0.10(0.04)
1.0 1.00(0.03)	1.5 1.49(0.03)	3.5 3.50(0.02)	0.1 0.11(0.04)	0.1 0.10(0.04)	0.8 0.80(0.05)
1.0 1.00(0.01)	1.5 1.50(0.02)	3.5 3.50(0.04)	0.6 0.60(0.07)	0.2 0.20(0.05)	0.2 0.20(0.06)



Tabla 6.14: Estimaciones de los riesgos latentes y probabilidades  $k = 5$ . Incidencia  $3/10^4$  hab. Continuación.

		Riesgos					Probabilidades													
0.2	0.7	1.0	1.5	2.5	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	
0.10(0.05)	0.41(0.11)	1.03(0.24)	2.71(0.92)	179.47(134.27)	0.09(0.05)	0.21(0.10)	0.33(0.10)	0.28(0.07)	0.28(0.07)	0.33(0.10)	0.28(0.07)	0.28(0.07)	0.33(0.10)	0.28(0.07)	0.28(0.07)	0.33(0.10)	0.28(0.07)	0.28(0.07)	0.33(0.10)	0.28(0.07)
0.2	0.7	1.0	1.5	2.5	0.1	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
0.11(0.07)	0.46(0.17)	1.11(0.34)	3.36(1.44)	206.25(162.98)	0.06(0.04)	0.20(0.12)	0.38(0.11)	0.28(0.12)	0.28(0.12)	0.38(0.11)	0.28(0.12)	0.28(0.12)	0.38(0.11)	0.28(0.12)	0.28(0.12)	0.38(0.11)	0.28(0.12)	0.28(0.12)	0.38(0.11)	0.28(0.12)
0.2	0.7	1.0	1.5	2.5	0.6	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.08(0.03)	0.32(0.08)	0.92(0.21)	2.76(0.82)	170.63(197.87)	0.22(0.09)	0.35(0.09)	0.22(0.08)	0.16(0.06)	0.16(0.06)	0.22(0.08)	0.16(0.06)	0.16(0.06)	0.22(0.08)	0.16(0.06)	0.16(0.06)	0.22(0.08)	0.16(0.06)	0.16(0.06)	0.22(0.08)	0.16(0.06)
0.2	0.7	1.0	1.5	2.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.10(0.04)	0.42(0.15)	1.06(0.40)	2.46(1.23)	161.40(228.52)	0.04(0.03)	0.12(0.06)	0.22(0.13)	0.42(0.17)	0.42(0.17)	0.22(0.13)	0.42(0.17)	0.42(0.17)	0.22(0.13)	0.42(0.17)	0.42(0.17)	0.22(0.13)	0.42(0.17)	0.42(0.17)	0.22(0.13)	0.42(0.17)
0.5	1.0	2.0	2.5	3.5	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
0.14(0.06)	0.56(0.15)	1.32(0.42)	3.02(1.50)	171.60(195.60)	0.04(0.03)	0.17(0.08)	0.29(0.10)	0.34(0.09)	0.34(0.09)	0.29(0.10)	0.34(0.09)	0.34(0.09)	0.29(0.10)	0.34(0.09)	0.34(0.09)	0.29(0.10)	0.34(0.09)	0.34(0.09)	0.29(0.10)	0.34(0.09)
0.5	1.0	2.0	2.5	3.5	0.1	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
0.14(0.05)	0.57(0.17)	1.40(0.37)	3.31(1.73)	187.30(189.43)	0.03(0.02)	0.13(0.09)	0.33(0.12)	0.37(0.11)	0.37(0.11)	0.33(0.12)	0.37(0.11)	0.37(0.11)	0.33(0.12)	0.37(0.11)	0.37(0.11)	0.33(0.12)	0.37(0.11)	0.37(0.11)	0.33(0.12)	0.37(0.11)
0.5	1.0	2.0	2.5	3.5	0.6	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.14(0.05)	0.48(0.15)	1.21(0.44)	3.28(1.76)	198.75(287.28)	0.09(0.08)	0.32(0.09)	0.29(0.10)	0.18(0.08)	0.18(0.08)	0.29(0.10)	0.18(0.08)	0.18(0.08)	0.29(0.10)	0.18(0.08)	0.18(0.08)	0.29(0.10)	0.18(0.08)	0.18(0.08)	0.29(0.10)	0.18(0.08)
0.5	1.0	2.0	2.5	3.5	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
0.12(0.06)	0.50(0.19)	1.24(0.49)	2.93(1.55)	178.01(271.28)	0.02(0.02)	0.08(0.04)	0.17(0.11)	0.41(0.21)	0.41(0.21)	0.17(0.11)	0.41(0.21)	0.41(0.21)	0.17(0.11)	0.41(0.21)	0.41(0.21)	0.17(0.11)	0.41(0.21)	0.41(0.21)	0.17(0.11)	0.41(0.21)
0.5	1.0	1.5	2.5	3.5	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
0.14(0.06)	0.54(0.18)	1.26(0.36)	2.98(1.21)	153.42(183.51)	0.04(0.04)	0.18(0.10)	0.30(0.08)	0.31(0.10)	0.31(0.10)	0.30(0.08)	0.31(0.10)	0.31(0.10)	0.30(0.08)	0.31(0.10)	0.31(0.10)	0.30(0.08)	0.31(0.10)	0.31(0.10)	0.30(0.08)	0.31(0.10)
0.5	1.0	1.5	2.5	3.5	0.1	0.2	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3	0.3
0.14(0.06)	0.58(0.22)	1.36(0.43)	3.21(1.38)	199.30(325.00)	0.02(0.03)	0.15(0.09)	0.35(0.11)	0.33(0.10)	0.33(0.10)	0.35(0.11)	0.33(0.10)	0.33(0.10)	0.35(0.11)	0.33(0.10)	0.33(0.10)	0.35(0.11)	0.33(0.10)	0.33(0.10)	0.35(0.11)	0.33(0.10)





Tabla 6.17: Estimaciones de los riesgos latentes y probabilidades  $k = 5$ . Incidencia  $30/10^4$  hab. Continuación

		Riesgos					Probabilidades					
0.2	0.7	1.0	1.5	2.5	2.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
0.20(0.04)	0.73(0.16)	1.31(0.35)	2.36(0.72)	171.35(248.28)	0.18(0.07)	0.29(0.13)	0.27(0.07)	0.20(0.06)	0.02(0.12)			
0.2	0.7	1.0	1.5	2.5	0.1	0.2	0.3	0.3	0.1			
0.19(0.05)	0.74(0.21)	1.29(0.39)	2.25(0.73)	138.81(197.22)	0.09(0.05)	0.33(0.19)	0.35(0.06)	0.17(0.14)	0.03(0.07)			
0.2	0.7	1.0	1.5	2.5	0.6	0.1	0.1	0.1	0.1			
0.19(0.05)	0.63(0.29)	1.18(0.35)	2.24(0.73)	143.90(225.38)	0.52(0.23)	0.19(0.16)	0.14(0.05)	0.11(0.03)	0.03(0.06)			
0.2	0.7	1.0	1.5	2.5	0.1	0.1	0.1	0.1	0.1			
0.18(0.07)	0.70(0.26)	1.34(0.43)	2.47(0.45)	216.08(225.76)	0.08(0.06)	0.13(0.08)	0.14(0.06)	0.56(0.29)	0.00(0.26)			
0.5	1.0	2.0	2.5	3.5	0.2	0.2	0.2	0.2	0.2			
0.43(0.21)	0.89(0.31)	1.75(0.67)	2.48(0.46)	3.66(81.42)	0.14(0.12)	0.19(0.06)	0.21(0.10)	0.29(0.10)	0.17(0.10)			
0.5	1.0	2.0	2.5	3.5	0.1	0.2	0.3	0.3	0.1			
0.41(0.20)	0.92(0.28)	1.84(0.57)	2.49(0.31)	3.82(55.98)	0.07(0.06)	0.18(0.08)	0.27(0.12)	0.36(0.19)	0.09(0.07)			
0.5	1.0	2.0	2.5	3.5	0.6	0.1	0.1	0.1	0.1			
0.39(0.24)	0.82(0.36)	1.65(0.76)	2.63(0.84)	67.95(230.26)	0.39(0.37)	0.25(0.29)	0.12(0.06)	0.15(0.06)	0.06(0.09)			
0.5	1.0	2.0	2.5	3.5	0.1	0.1	0.1	0.1	0.1			
0.25(0.16)	0.67(0.21)	1.22(0.37)	2.32(0.30)	3.51(0.11)	0.03(0.04)	0.10(0.03)	0.11(0.04)	0.18(0.08)	0.59(0.11)			
0.5	1.0	1.5	2.5	3.5	0.2	0.2	0.2	0.2	0.2			
0.46(0.11)	0.94(0.14)	1.50(0.15)	2.51(0.14)	3.54(0.13)	0.17(0.08)	0.21(0.07)	0.23(0.08)	0.21(0.06)	0.19(0.08)			
0.5	1.0	1.5	2.5	3.5	0.1	0.2	0.3	0.3	0.1			
0.43(0.12)	0.94(0.16)	1.50(0.12)	2.50(0.09)	3.59(0.29)	0.08(0.06)	0.21(0.09)	0.29(0.10)	0.30(0.08)	0.10(0.06)			



Tabla 6.19: Estimaciones de los riesgos latentes y probabilidades  $k = 5$ . Incidencia  $300/10^4$  hab.

	Riesgos					Probabilidades					
5311	0.2 0.20(0.01)	0.7 0.70(0.02)	1.0 1.00(0.02)	3.0 3.00(0.03)	3.5 3.50(0.04)	0.2 0.20(0.06)	0.2 0.20(0.05)	0.2 0.20(0.06)	0.2 0.20(0.05)	0.2 0.20(0.05)	0.2 0.19(0.06)
5312	0.2 0.20(0.02)	0.7 0.70(0.03)	1.0 1.01(0.03)	3.0 2.98(0.06)	3.5 3.49(0.06)	0.1 0.11(0.04)	0.2 0.20(0.07)	0.3 0.29(0.08)	0.3 0.28(0.06)	0.1 0.11(0.04)	
5313	0.2 0.20(0.00)	0.7 0.70(0.02)	1.0 1.00(0.04)	3.0 2.99(0.06)	3.5 3.50(0.06)	0.6 0.61(0.08)	0.1 0.10(0.05)	0.1 0.10(0.05)	0.1 0.10(0.03)	0.1 0.10(0.05)	
5314	0.2 0.20(0.02)	0.7 0.70(0.02)	1.0 1.01(0.04)	3.0 3.00(0.04)	3.5 3.50(0.02)	0.1 0.10(0.05)	0.1 0.10(0.04)	0.1 0.10(0.04)	0.1 0.10(0.04)	0.1 0.10(0.04)	0.6 0.60(0.07)
5321	0.5 0.56(0.08)	0.7 0.95(0.20)	1.0 1.45(0.36)	1.5 2.16(0.91)	3.5 3.52(116.99)	0.2 0.32(0.11)	0.2 0.22(0.04)	0.2 0.19(0.05)	0.2 0.12(0.13)	0.2 0.18(0.12)	
5322	0.5 0.62(0.05)	0.7 0.99(0.04)	1.0 1.49(0.06)	1.5 2.67(1.07)	3.5 115.15(260.70)	0.1 0.28(0.08)	0.2 0.30(0.07)	0.3 0.28(0.08)	0.3 0.08(0.07)	0.1 0.06(0.08)	
5323	0.5 0.50(0.01)	0.7 0.74(0.12)	1.0 1.10(0.28)	1.5 1.65(0.56)	3.5 3.51(0.08)	0.6 0.62(0.08)	0.1 0.12(0.06)	0.1 0.10(0.04)	0.1 0.08(0.03)	0.1 0.09(0.06)	
5324	0.5 0.53(0.05)	0.7 0.79(0.15)	1.0 1.15(0.28)	1.5 1.71(0.46)	3.5 3.50(0.02)	0.1 0.14(0.06)	0.1 0.11(0.03)	0.1 0.10(0.04)	0.1 0.07(0.05)	0.6 0.58(0.06)	



Tabla 6.21: Estimaciones de los riesgos latentes y probabilidades  $k = 5$ . Incidencia  $300/10^4$  hab. Continuación

Riesgos		Probabilidades									
0.5	1.0	1.5	2.5	3.5	0.6	0.1	0.1	0.1	0.1	0.1	0.1
0.50(0.00)	1.00(0.04)	1.50(0.04)	2.5(0.04)	3.50(0.06)	0.61(0.08)	0.10(0.05)	0.10(0.04)	0.10(0.04)	0.10(0.04)	0.10(0.04)	0.10(0.04)
0.5	1.0	1.5	2.5	3.5	0.1	0.1	0.1	0.1	0.1	0.1	0.6
0.50(0.02)	1.00(0.03)	1.50(0.03)	2.49(0.05)	3.50(0.02)	0.10(0.04)	0.10(0.05)	0.10(0.04)	0.10(0.04)	0.10(0.04)	0.10(0.04)	0.60(0.06)

Tabla 6.22: Estimaciones de los riesgos latentes  $k = 7$ . Incidencia  $3/10^4$ hab.

Riesgos						
0.2 0.05(0.03)	0.5 0.23(0.07)	0.8 0.54(0.13)	1.0 1.08(0.24)	1.5 2.17(0.54)	2.5 6.52(2.68)	3.5 444.28(564.19)
0.2 0.06(0.02)	0.5 0.21(0.06)	0.8 0.51(0.15)	1.0 1.10(0.30)	1.5 2.24(0.64)	2.5 7.29(3.72)	3.5 550.99(448.93)
0.2 0.06(0.02)	0.5 0.22(0.08)	0.8 0.54(0.14)	1.0 1.10(0.33)	1.5 2.24(0.76)	2.5 6.50(4.38)	3.5 460.14(449.33)
0.2 0.06(0.02)	0.5 0.22(0.08)	0.8 0.52(0.14)	1.0 1.10(0.30)	1.5 2.26(0.61)	2.5 7.45(3.62)	3.5 479.77(534.65)
0.2 0.06(0.02)	0.5 0.26(0.08)	0.8 0.61(0.17)	1.0 1.17(0.35)	1.5 2.34(0.71)	2.5 7.67(3.31)	3.5 578.21(648.07)
0.5 0.07(0.02)	0.7 0.27(0.07)	0.9 0.61(0.15)	1.0 1.17(0.27)	1.5 2.30(0.68)	2.0 8.12(2.95)	2.5 523.86(531.53)
0.5 0.07(0.03)	0.7 0.27(0.10)	0.9 0.59(0.19)	1.0 1.19(0.35)	1.5 2.41(0.76)	2.0 8.85(2.86)	2.5 599.11(585.22)
0.5 0.06(0.03)	0.7 0.27(0.10)	0.9 0.63(0.21)	1.0 1.18(0.39)	1.5 2.30(0.85)	2.0 8.25(4.35)	2.5 693.71(852.31)
0.5 0.06(0.02)	0.7 0.25(0.07)	0.9 0.58(0.14)	1.0 1.12(0.24)	1.5 2.27(0.64)	2.0 8.76(3.41)	2.5 704.61(703.47)
0.5 0.07(0.02)	0.7 0.29(0.09)	0.9 0.67(0.15)	1.0 1.25(0.22)	1.5 2.58(0.57)	2.0 9.56(3.79)	2.5 761.12(669.06)

Tabla 6.23: Estimaciones de los riesgos latentes  $k = 7$ . Incidencia  $3/10^4$ hab.  
Continuación.

Riesgos						
0.2 0.05(0.02)	0.4 0.22(0.07)	0.6 0.50(0.16)	1.0 0.99(0.28)	1.3 2.11(0.71)	1.5 8.35(3.95)	2.0 627.78(462.79)
0.2 0.05(0.02)	0.4 0.20(0.07)	0.6 0.49(0.14)	1.0 1.01(0.26)	1.3 2.30(0.73)	1.5 9.14(3.48)	2.0 633.24(741.40)
0.2 0.05(0.02)	0.4 0.22(0.07)	0.6 0.52(0.14)	1.0 1.06(0.28)	1.3 2.17(0.68)	1.5 8.03(3.43)	2.0 489.82(578.89)
0.2 0.05(0.02)	0.4 0.20(0.05)	0.6 0.47(0.12)	1.0 0.94(0.21)	1.3 1.99(0.52)	1.5 8.29(2.67)	2.0 647.43(765.04)
0.2 0.06(0.02)	0.4 0.23(0.07)	0.6 0.55(0.16)	1.0 1.04(0.30)	1.3 2.17(0.73)	1.5 8.86(3.36)	2.0 683.36(605.01)
0.1 0.04(0.02)	0.4 0.19(0.06)	0.6 0.47(0.14)	1.0 0.98(0.35)	2.7 2.07(0.80)	3.0 5.84(3.41)	3.5 465.85(635.07)
0.1 0.04(0.02)	0.4 0.17(0.06)	0.6 0.43(0.15)	1.0 0.98(0.41)	2.7 2.20(0.83)	3.0 6.84(4.24)	3.5 465.00(44.93)
0.1 0.04(0.01)	0.4 0.19(0.06)	0.6 0.45(0.14)	1.0 0.93(0.39)	2.7 1.92(1.03)	3.0 4.96(4.87)	3.5 356.40(453.56)
0.1 0.04(0.02)	0.4 0.17(0.05)	0.6 0.44(0.13)	1.0 0.99(0.35)	2.7 2.20(0.96)	3.0 6.94(4.13)	3.5 439.99(355.82)
0.1 0.05(0.02)	0.4 0.22(0.08)	0.6 0.54(0.19)	1.0 1.07(0.44)	2.7 2.22(0.89)	3.0 7.20(4.17)	3.5 543.92(514.60)
0.5 0.07(0.02)	1.0 0.28(0.07)	1.5 0.66(0.19)	2.0 1.30(0.41)	2.5 2.53(0.73)	3.0 8.04(3.94)	3.5 582.24(494.61)
0.5 0.06(0.02)	1.0 0.25(0.10)	1.5 0.57(0.15)	2.0 1.18(0.33)	2.5 2.33(0.86)	3.0 7.18(4.52)	3.5 521.40(410.14)
0.5 0.06(0.03)	1.0 0.27(0.10)	1.5 0.65(0.19)	2.0 1.27(0.39)	2.5 2.39(0.77)	3.0 7.04(4.60)	3.5 463.52(602.23)
0.5 0.07(0.03)	1.0 0.27(0.09)	1.5 0.62(0.20)	2.0 1.21(0.38)	2.5 2.38(0.78)	3.0 7.25(3.51)	3.5 532.91(414.65)
0.5 0.06(0.02)	1.0 0.27(0.09)	1.5 0.65(0.21)	2.0 1.27(0.46)	2.5 2.37(0.95)	3.0 7.54(4.73)	3.5 513.93(796.55)

Tabla 6.24: Estimaciones de las probabilidades  $k = 7$ . Incidencia  $3/10^4$ hab.

Probabilidades						
0.14 0.04(0.03)	0.14 0.10(0.04)	0.14 0.19(0.09)	0.14 0.24(0.07)	0.14 0.22(0.07)	0.14 0.15(0.06)	0.14 0.03(0.03)
0.4 0.08(0.06)	0.1 0.18(0.09)	0.1 0.22(0.07)	0.1 0.19(0.07)	0.1 0.17(0.05)	0.1 0.11(0.06)	0.1 0.02(0.03)
0.1 0.03(0.03)	0.1 0.07(0.04)	0.1 0.13(0.06)	0.1 0.18(0.08)	0.1 0.24(0.10)	0.1 0.24(0.12)	0.4 0.07(0.09)
0.3 0.06(0.06)	0.2 0.16(0.10)	0.1 0.23(0.09)	0.1 0.21(0.07)	0.1 0.17(0.06)	0.1 0.10(0.05)	0.1 0.02(0.02)
0.1 0.03(0.02)	0.1 0.09(0.06)	0.1 0.23(0.11)	0.4 0.29(0.10)	0.1 0.22(0.10)	0.1 0.10(0.06)	0.1 0.01(0.02)
0.14 0.01(0.01)	0.14 0.07(0.06)	0.14 0.20(0.09)	0.14 0.29(0.10)	0.14 0.26(0.09)	0.14 0.11(0.08)	0.14 0.01(0.03)
0.4 0.02(0.02)	0.1 0.12(0.08)	0.1 0.26(0.09)	0.1 0.28(0.10)	0.1 0.19(0.08)	0.1 0.07(0.07)	0.1 0.00(0.02)
0.1 0.01(0.01)	0.1 0.06(0.05)	0.1 0.16(0.09)	0.1 0.25(0.09)	0.1 0.30(0.09)	0.1 0.16(0.15)	0.4 0.01(0.03)
0.3 0.02(0.02)	0.2 0.09(0.08)	0.1 0.24(0.11)	0.1 0.30(0.08)	0.1 0.21(0.08)	0.1 0.09(0.06)	0.1 0.00(0.02)
0.1 0.01(0.01)	0.1 0.07(0.08)	0.1 0.24(0.15)	0.4 0.31(0.11)	0.1 0.22(0.11)	0.1 0.08(0.07)	0.1 0.01(0.01)

Tabla 6.25: Estimaciones de las probabilidades  $k = 7$ . Incidencia  $3/10^4$ hab. Continuación.

Probabilidades						
0.14 0.04(0.03)	0.14 0.11(0.07)	0.14 0.22(0.09)	0.14 0.27(0.09)	0.14 0.24(0.10)	0.14 0.07(0.08)	0.14 0.00(0.01)
0.4 0.07(0.07)	0.1 0.20(0.10)	0.1 0.23(0.07)	0.1 0.22(0.08)	0.1 0.17(0.07)	0.1 0.06(0.05)	0.1 0.00(0.01)
0.1 0.03(0.02)	0.1 0.08(0.05)	0.1 0.16(0.06)	0.1 0.25(0.12)	0.1 0.29(0.11)	0.1 0.13(0.12)	0.4 0.00(0.02)
0.3 0.07(0.04)	0.2 0.18(0.08)	0.1 0.24(0.09)	0.1 0.23(0.07)	0.1 0.19(0.07)	0.1 0.07(0.06)	0.1 0.00(0.01)
0.1 0.03(0.03)	0.1 0.09(0.06)	0.1 0.23(0.09)	0.4 0.29(0.09)	0.1 0.26(0.12)	0.1 0.06(0.06)	0.1 0.00(0.01)
0.14 0.05(0.04)	0.14 0.11(0.05)	0.14 0.17(0.08)	0.14 0.18(0.06)	0.14 0.19(0.09)	0.14 0.20(0.10)	0.14 0.05(0.09)
0.4 0.14(0.08)	0.1 0.19(0.10)	0.1 0.18(0.07)	0.1 0.14(0.05)	0.1 0.15(0.04)	0.1 0.13(0.08)	0.1 0.03(0.05)
0.1 0.04(0.04)	0.1 0.08(0.05)	0.1 0.11(0.05)	0.1 0.14(0.07)	0.1 0.17(0.14)	0.1 0.28(0.15)	0.4 0.10(0.22)
0.3 0.12(0.07)	0.2 0.18(0.07)	0.1 0.18(0.07)	0.1 0.17(0.06)	0.1 0.15(0.07)	0.1 0.13(0.08)	0.1 0.03(0.05)
0.1 0.04(0.04)	0.1 0.09(0.06)	0.1 0.18(0.11)	0.4 0.24(0.09)	0.1 0.22(0.10)	0.1 0.15(0.11)	0.1 0.02(0.03)
0.14 0.01(0.01)	0.14 0.05(0.03)	0.14 0.13(0.08)	0.14 0.24(0.10)	0.14 0.30(0.09)	0.14 0.20(0.12)	0.14 0.04(0.06)
0.4 0.02(0.02)	0.1 0.09(0.07)	0.1 0.21(0.08)	0.1 0.23(0.07)	0.1 0.23(0.07)	0.1 0.16(0.09)	0.1 0.02(0.04)
0.1 0.01(0.00)	0.1 0.03(0.02)	0.1 0.09(0.05)	0.1 0.18(0.09)	0.1 0.29(0.13)	0.1 0.27(0.19)	0.4 0.05(0.11)
0.3 0.02(0.02)	0.2 0.07(0.06)	0.1 0.19(0.08)	0.1 0.24(0.09)	0.1 0.24(0.08)	0.1 0.16(0.09)	0.1 0.02(0.03)
0.1 0.01(0.01)	0.1 0.03(0.03)	0.1 0.095(0.07)	0.4 0.23(0.12)	0.1 0.34(0.15)	0.1 0.20(0.16)	0.1 0.02(0.04)

Se observó además que en escenarios donde los riesgos latentes están muy cercanos uno del otro (por ejemplo,  $k=2$ ,  $\eta_1 = 0.8$  y  $\eta_2 = 1.2$ ), al modelo propuesto se le dificulta distinguir las clases, es aquí en estos escenarios en donde los valores estimados de los riesgos latentes pueden tomar valores atípicos extremos. Por otro lado, el impacto de la incidencia en cada uno de los diferentes escenarios es evidente, tiende a mejorar la estimación cada vez que tenemos más información, sin embargo en aquellos escenarios en los que resulta difícil distinguir las clases la situación de valores extremos persiste. Cada vez que aumentamos el número de clases en el modelo el problema se vuelve más complejo y resulta más complicado distinguir las clases. En un anexo posterior revisaremos el error de ajuste del modelo.

### A.3 Probabilidades finales de $k$

Este anexo contiene las tablas de las probabilidades finales de  $k$  obtenidas con el modelo propuesto. Para cada  $k$  y cada escenario se observa un valor real y los  $k$  supuestos los cuales permiten estudiar la posible ocupación de cada una de las clases existentes en el modelo. Se muestran los valores medios y las desviaciones típicas correspondientes a los diferentes  $k$ .

Tabla 6.26: Probabilidades finales con  $k$  real  $k = 1$  ( $\eta = 1, p = 1$ )

Inc.	k supuesto						
	1	2	3	4	5	6	7
$3/10^4$ hab	<b>0.94(0.12)</b>	0.05(0.08)	0.01(0.02)	0.00(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
$30/10^4$ hab	<b>0.99(0.02)</b>	0.01(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
$300/10^4$ hab	<b>0.99(0.00)</b>	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.27: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc. =  $3/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.68(0.21)</b>	0.19(0.12)	0.07(0.09)	0.03(0.03)	0.02(0.08)	0.01(0.02)
2	0.00(0.00)	<b>0.82(0.17)</b>	0.13(0.11)	0.04(0.11)	0.01(0.02)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.57(0.24)</b>	0.22(0.12)	0.12(0.15)	0.05(0.07)	0.02(0.04)	0.03(0.08)
4	0.05(0.18)	<b>0.83(0.20)</b>	0.10(0.11)	0.02(0.05)	0.00(0.00)	0.00(0.01)	0.00(0.00)

Tabla 6.28: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc. =  $30/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.94(0.10)</b>	0.06(0.09)	0.00(0.01)	0.00(0.01)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	<b>0.97(0.03)</b>	0.03(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.91(0.11)</b>	0.07(0.10)	0.01(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.97(0.03)</b>	0.02(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Las probabilidades finales de  $k$  obtenidas en este estudio permiten estimar el número desconocido de clases, además usando dichas probabilidades y la estimación de los riesgos latentes podemos obtener los riesgos relativos estimados para cada una de las áreas pequeñas. En la mayoría de escenarios estudiados el valor mayor de la media se corresponde con el  $k$  real correspondiente, además esta situación va siendo más contundente a medida que la incidencia aumenta.

Tabla 6.29: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc. =  $300/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.99(0.04)</b>	0.01(0.04)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	<b>0.99(0.01)</b>	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.98(0.02)</b>	0.01(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.99(0.00)</b>	0.01(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.30: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 2.5)$ , Inc. =  $3/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.01)	<b>0.62(0.22)</b>	0.23(0.13)	0.09(0.12)	0.04(0.04)	0.02(0.02)	0.01(0.01)
2	0.01(0.04)	<b>0.78(0.15)</b>	0.15(0.10)	0.03(0.03)	0.02(0.03)	0.01(0.02)	0.00(0.00)
3	0.00(0.00)	<b>0.57(0.19)</b>	0.23(0.12)	0.10(0.09)	0.06(0.08)	0.03(0.05)	0.02(0.04)
4	0.18(0.31)	<b>0.67(0.29)</b>	0.11(0.11)	0.03(0.08)	0.01(0.03)	0.00(0.00)	0.00(0.00)

Tabla 6.31: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 2.5)$ , Inc. =  $30/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.93(0.12)</b>	0.06(0.10)	0.00(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	<b>0.95(0.12)</b>	0.04(0.10)	0.01(0.03)	0.00(0.02)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.91(0.13)</b>	0.08(0.11)	0.01(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.95(0.11)</b>	0.05(0.11)	0.00(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.32: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 2.5)$ , Inc. =  $300/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.99(0.01)</b>	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	<b>0.99(0.01)</b>	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.98(0.02)</b>	0.01(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.99(0.01)</b>	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.33: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc. =  $3/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.05(0.15)	<b>0.59(0.17)</b>	0.21(0.10)	0.08(0.07)	0.04(0.10)	0.02(0.07)	0.01(0.01)
2	0.36(0.36)	<b>0.48(0.28)</b>	0.11(0.13)	0.03(0.04)	0.02(0.06)	0.00(0.01)	0.00(0.01)
3	0.22(0.32)	<b>0.50(0.22)</b>	0.16(0.10)	0.06(0.06)	0.03(0.05)	0.01(0.03)	0.01(0.01)
4	<b>0.67(0.35)</b>	0.26(0.27)	0.05(0.09)	0.01(0.03)	0.00(0.01)	0.00(0.00)	0.00(0.00)

Tabla 6.34: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc. =  $30/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.90(0.14)</b>	0.08(0.11)	0.01(0.03)	0.00(0.00)	0.00(0.10)	0.00(0.00)
2	0.00(0.00)	<b>0.92(0.10)</b>	0.07(0.09)	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.89(0.14)</b>	0.10(0.11)	0.02(0.03)	0.00(0.01)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.93(0.11)</b>	0.05(0.06)	0.01(0.07)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.35: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc. =  $300/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.97(0.07)</b>	0.02(0.06)	0.00(0.00)	0.00(0.01)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	<b>0.98(0.02)</b>	0.01(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.98(0.03)</b>	0.02(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.98(0.03)</b>	0.02(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.36: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc. =  $3/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	<b>0.87(0.19)</b>	0.10(0.14)	0.02(0.05)	0.00(0.01)	0.00(0.01)	0.00(0.00)	0.00(0.00)
2	<b>0.90(0.16)</b>	0.08(0.12)	0.01(0.03)	0.00(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	<b>0.89(0.19)</b>	0.08(0.13)	0.02(0.06)	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	<b>0.95(0.08)</b>	0.05(0.06)	0.00(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.01)	0.00(0.00)

Tabla 6.37: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc. =  $30/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.02(0.10)	<b>0.87(0.14)</b>	0.10(0.10)	0.01(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
2	0.28(0.39)	<b>0.66(0.37)</b>	0.06(0.10)	0.01(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.15(0.29)	<b>0.75(0.29)</b>	0.09(0.14)	0.01(0.02)	0.00(0.01)	0.00(0.00)	0.00(0.00)
4	<b>0.64(0.38)</b>	0.34(0.369)	0.02(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.38: Prob. finales con  $k = 2$ ,  $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc. =  $300/10^4$  hab

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.97(0.05)</b>	0.03(0.05)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	<b>1.00(0.00)</b>	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	<b>0.97(0.04)</b>	0.03(0.04)	0.00(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	<b>0.98(0.05)</b>	0.02(0.05)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.39: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc. =  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.40(0.25)</b>	0.30(0.15)	0.14(0.10)	0.10(0.12)	0.04(0.04)	0.03(0.10)
2	0.00(0.00)	<b>0.51(0.27)</b>	0.27(0.14)	0.12(0.13)	0.06(0.07)	0.02(0.05)	0.01(0.03)
3	0.06(0.20)	<b>0.48(0.26)</b>	0.23(0.15)	0.11(0.11)	0.05(0.06)	0.05(0.11)	0.20(0.03)
4	0.00(0.00)	<b>0.78(0.19)</b>	0.17(0.15)	0.03(0.05)	0.01(0.03)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	<b>0.44(0.26)</b>	0.27(0.15)	0.14(0.12)	0.08(0.09)	0.05(0.07)	0.03(0.06)

Tabla 6.40: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc. =  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	<b>0.69(0.23)</b>	0.20(0.13)	0.07(0.12)	0.02(0.04)	0.02(0.09)
2	0.00(0.00)	0.00(0.00)	<b>0.80(0.17)</b>	0.16(0.12)	0.04(0.06)	0.01(0.03)	0.00(0.00)
3	0.00(0.00)	0.00(0.01)	<b>0.72(0.19)</b>	0.20(0.14)	0.05(0.09)	0.02(0.07)	0.01(0.02)
4	0.00(0.00)	0.11(0.24)	<b>0.79(0.23)</b>	0.09(0.09)	0.01(0.01)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	0.00(0.05)	<b>0.72(0.19)</b>	0.20(0.14)	0.05(0.06)	0.02(0.08)	0.00(0.01)

Tabla 6.41: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc. =  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	<b>0.93(0.07)</b>	0.06(0.07)	0.00(0.00)	0.00(0.00)	0.00(0.01)
2	0.00(0.00)	0.00(0.00)	<b>0.94(0.08)</b>	0.06(0.08)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	0.00(0.00)	<b>0.92(0.11)</b>	0.07(0.09)	0.01(0.02)	0.00(0.02)	0.00(0.00)
4	0.00(0.00)	0.00(0.00)	<b>0.96(0.07)</b>	0.04(0.07)	0.00(0.01)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	0.00(0.00)	<b>0.94(0.07)</b>	0.06(0.07)	0.00(0.00)	0.00(0.00)	0.00(0.00)

Tabla 6.42: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc. =  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.02(0.10)	<b>0.55(0.22)</b>	0.24(0.14)	0.11(0.10)	0.05(0.06)	0.02(0.05)	0.01(0.02)
2	0.03(0.09)	<b>0.62(0.18)</b>	0.22(0.12)	0.08(0.07)	0.03(0.05)	0.02(0.09)	0.01(0.04)
3	<b>0.46(0.35)</b>	0.34(0.22)	0.13(0.14)	0.04(0.06)	0.01(0.02)	0.02(0.09)	0.00(0.01)
4	0.26(0.35)	<b>0.56(0.28)</b>	0.13(0.10)	0.04(0.07)	0.01(0.03)	0.00(0.01)	0.00(0.01)
5	0.02(0.11)	<b>0.50(0.21)</b>	0.23(0.11)	0.13(0.12)	0.06(0.07)	0.03(0.04)	0.04(0.09)

Tabla 6.43: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc. =  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	<b>0.68(0.22)</b>	0.22(0.17)	0.08(0.13)	0.02(0.03)	0.01(0.01)
2	0.00(0.00)	0.01(0.05)	<b>0.69(0.22)</b>	0.20(0.15)	0.08(0.11)	0.03(0.08)	0.01(0.03)
3	0.00(0.00)	0.00(0.04)	<b>0.68(0.22)</b>	0.21(0.14)	0.08(0.13)	0.02(0.08)	0.01(0.02)
4	0.00(0.00)	0.24(0.34)	<b>0.63(0.31)</b>	0.11(0.11)	0.02(0.04)	0.00(0.01)	0.00(0.00)
5	0.00(0.00)	0.00(0.01)	<b>0.69(0.22)</b>	0.21(0.16)	0.06(0.12)	0.02(0.09)	0.01(0.01)

Tabla 6.44: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc. =  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	<b>0.92(0.12)</b>	0.06(0.08)	0.01(0.08)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	0.00(0.00)	<b>0.92(0.11)</b>	0.07(0.09)	0.01(0.01)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	0.00(0.00)	<b>0.92(0.09)</b>	0.07(0.08)	0.01(0.01)	0.00(0.01)	0.00(0.00)
4	0.00(0.00)	0.00(0.00)	<b>0.93(0.12)</b>	0.06(0.12)	0.00(0.01)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	0.00(0.00)	<b>0.93(0.09)</b>	0.07(0.09)	0.00(0.01)	0.00(0.00)	0.00(0.00)

Tabla 6.45: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc. =  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	<b>0.75(0.30)</b>	0.19(0.22)	0.04(0.06)	0.01(0.03)	0.00(0.02)	0.00(0.00)	0.00(0.01)
2	<b>0.75(0.28)</b>	0.20(0.22)	0.03(0.05)	0.01(0.02)	0.00(0.00)	0.00(0.00)	0.00(0.00)
3	<b>0.89(0.16)</b>	0.08(0.12)	0.02(0.04)	0.00(0.01)	0.00(0.00)	0.00(0.00)	0.00(0.00)
4	<b>0.82(0.27)</b>	0.14(0.20)	0.03(0.05)	0.00(0.01)	0.00(0.01)	0.00(0.00)	0.00(0.00)
5	<b>0.76(0.29)</b>	0.18(0.20)	0.04(0.07)	0.10(0.01)	0.00(0.01)	0.00(0.00)	0.00(0.02)

Tabla 6.46: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc. =  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.69(0.27)</b>	0.23(0.19)	0.05(0.09)	0.02(0.08)	0.01(0.08)	0.00(0.00)
2	0.00(0.00)	<b>0.78(0.25)</b>	0.18(0.21)	0.03(0.07)	0.01(0.02)	0.00(0.02)	0.00(0.00)
3	0.19(0.34)	<b>0.66(0.31)</b>	0.13(0.15)	0.02(0.05)	0.00(0.01)	0.00(0.00)	0.00(0.00)
4	0.00(0.02)	<b>0.88(0.16)</b>	0.10(0.14)	0.01(0.03)	0.00(0.00)	0.00(0.00)	0.00(0.00)
5	0.00(0.04)	<b>0.81(0.19)</b>	0.16(0.15)	0.02(0.05)	0.01(0.03)	0.00(0.03)	0.00(0.00)

Tabla 6.47: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc. =  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	<b>0.86(0.18)</b>	0.13(0.17)	0.01(0.02)	0.00(0.00)	0.00(0.00)
2	0.00(0.00)	0.00(0.00)	<b>0.86(0.17)</b>	0.12(0.16)	0.01(0.03)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	0.01(0.10)	<b>0.85(0.19)</b>	0.12(0.17)	0.02(0.04)	0.00(0.01)	0.00(0.00)
4	0.00(0.00)	0.14(0.29)	<b>0.79(0.28)</b>	0.05(0.06)	0.01(0.08)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	0.00(0.00)	<b>0.86(0.18)</b>	0.13(0.17)	0.01(0.03)	0.00(0.03)	0.00(0.00)

Tabla 6.48: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc. =  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.62(0.23)</b>	0.22(0.13)	0.09(0.09)	0.04(0.08)	0.02(0.06)	0.01(0.02)
2	0.00(0.00)	<b>0.69(0.19)</b>	0.21(0.14)	0.06(0.07)	0.02(0.03)	0.01(0.03)	0.00(0.00)
3	0.23(0.35)	<b>0.53(0.27)</b>	0.16(0.12)	0.05(0.06)	0.02(0.04)	0.01(0.02)	0.00(0.02)
4	0.03(0.12)	<b>0.77(0.20)</b>	0.15(0.14)	0.04(0.08)	0.01(0.02)	0.00(0.00)	0.00(0.00)
5	0.01(0.08)	<b>0.59(0.24)</b>	0.24(0.14)	0.09(0.09)	0.04(0.06)	0.02(0.03)	0.01(0.04)

Tabla 6.49: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc. =  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.01(0.07)	<b>0.77(0.19)</b>	0.17(0.16)	0.04(0.07)	0.01(0.01)	0.00(0.01)
2	0.00(0.00)	0.07(0.20)	<b>0.73(0.23)</b>	0.17(0.18)	0.02(0.05)	0.01(0.03)	0.00(0.00)
3	0.00(0.00)	0.07(0.18)	<b>0.73(0.21)</b>	0.16(0.14)	0.03(0.08)	0.01(0.02)	0.00(0.01)
4	0.00(0.00)	<b>0.51(0.37)</b>	0.45(0.34)	0.04(0.04)	0.00(0.01)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	0.02(0.11)	<b>0.75(0.22)</b>	0.17(0.17)	0.04(0.07)	0.01(0.05)	0.00(0.01)

Tabla 6.50: Prob. finales con  $k = 3$ ,  $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc. =  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	<b>0.92(0.11)</b>	0.07(0.10)	0.01(0.02)	0.00(0.00)	0.00(0.01)
2	0.00(0.00)	0.00(0.00)	<b>0.93(0.09)</b>	0.06(0.07)	0.01(0.02)	0.00(0.00)	0.00(0.00)
3	0.00(0.00)	0.00(0.00)	<b>0.94(0.07)</b>	0.06(0.07)	0.00(0.01)	0.00(0.00)	0.00(0.00)
4	0.00(0.00)	0.00(0.00)	<b>0.95(0.10)</b>	0.05(0.09)	0.00(0.02)	0.00(0.00)	0.00(0.00)
5	0.00(0.00)	0.00(0.00)	<b>0.92(0.13)</b>	0.07(0.12)	0.00(0.01)	0.00(0.00)	0.00(0.00)

Tabla 6.51: Prob. finales con  $k = 5$ ,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc. =  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.18(0.22)	<b>0.32(0.16)</b>	0.22(0.15)	0.12(0.10)	0.11(0.18)	0.05(0.08)
2	0.00(0.00)	0.31(0.27)	<b>0.33(0.15)</b>	0.18(0.11)	0.09(0.09)	0.07(0.10)	0.03(0.04)
3	0.00(0.00)	0.19(0.24)	<b>0.29(0.20)</b>	0.21(0.18)	0.13(0.16)	0.10(0.14)	0.08(0.14)
4	0.00(0.00)	<b>0.56(0.31)</b>	0.28(0.19)	0.10(0.13)	0.05(0.11)	0.01(0.03)	0.00(0.01)

Tabla 6.52: Prob. finales con  $k = 5$ ,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc. =  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.16(0.23)	<b>0.31(0.23)</b>	0.30(0.24)	0.17(0.21)	0.06(0.13)
2	0.00(0.00)	0.00(0.00)	0.17(0.22)	<b>0.41(0.25)</b>	0.24(0.17)	0.12(0.14)	0.07(0.13)
3	0.00(0.00)	0.00(0.00)	<b>0.47(0.30)</b>	0.30(0.20)	0.13(0.13)	0.08(0.14)	0.02(0.05)
4	0.00(0.00)	0.00(0.00)	<b>0.45(0.33)</b>	0.35(0.22)	0.14(0.17)	0.04(0.08)	0.02(0.05)

Tabla 6.53: Prob. finales con  $k = 5$ ,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc. =  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.51(0.35)</b>	0.35(0.32)	0.14(0.23)
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.14)	<b>0.48(0.35)</b>	0.36(0.31)	0.14(0.20)
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.13)	<b>0.57(0.34)</b>	0.31(0.30)	0.09(0.17)
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.14)	<b>0.59(0.33)</b>	0.29(0.29)	0.10(0.18)

Tabla 6.54: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ ,  
Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.27(0.23)	<b>0.30(0.13)</b>	0.19(0.13)	0.13(0.12)	0.07(0.10)	0.06(0.11)
2	0.03(0.12)	<b>0.36(0.26)</b>	0.28(0.15)	0.16(0.14)	0.09(0.12)	0.05(0.09)	0.04(0.10)
3	0.01(0.08)	<b>0.33(0.25)</b>	0.27(0.17)	0.16(0.14)	0.13(0.20)	0.06(0.08)	0.04(0.06)
4	0.00(0.00)	<b>0.57(0.28)</b>	0.27(0.18)	0.09(0.10)	0.04(0.07)	0.01(0.03)	0.00(0.01)

Tabla 6.55: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ ,  
Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.07(0.15)	<b>0.39(0.24)</b>	0.31(0.23)	0.15(0.17)	0.08(0.14)
2	0.00(0.00)	0.00(0.00)	0.05(0.15)	<b>0.38(0.25)</b>	0.31(0.21)	0.17(0.20)	0.09(0.16)
3	0.00(0.00)	0.00(0.00)	0.23(0.29)	<b>0.37(0.20)</b>	0.21(0.18)	0.13(0.16)	0.05(0.10)
4	0.00(0.00)	0.01(0.08)	0.33(0.32)	<b>0.42(0.25)</b>	0.17(0.21)	0.05(0.08)	0.02(0.03)

Tabla 6.56: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ ,  
Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.41(0.30)	<b>0.42(0.30)</b>	0.18(0.21)
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.07)	<b>0.49(0.34)</b>	0.35(0.30)	0.15(0.22)
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.11)	<b>0.53(0.31)</b>	0.32(0.27)	0.12(0.18)
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.04(0.15)	<b>0.58(0.31)</b>	0.28(0.27)	0.10(0.16)

Tabla 6.57: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ ,  
Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.01)	0.22(0.21)	<b>0.31(0.15)</b>	0.22(0.18)	0.13(0.11)	0.08(0.10)	0.04(0.06)
2	0.07(0.20)	<b>0.39(0.23)</b>	0.26(0.16)	0.14(0.11)	0.07(0.09)	0.04(0.10)	0.02(0.06)
3	0.00(0.00)	<b>0.32(0.26)</b>	0.26(0.16)	0.19(0.18)	0.10(0.12)	0.08(0.15)	0.04(0.09)
4	0.00(0.00)	<b>0.47(0.25)</b>	0.32(0.15)	0.12(0.12)	0.06(0.09)	0.02(0.04)	0.01(0.01)

Tabla 6.58: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ ,  
Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.20(0.22)	<b>0.36(0.25)</b>	0.29(0.23)	0.15(0.18)
2	0.00(0.00)	0.00(0.00)	0.01(0.04)	0.15(0.19)	<b>0.36(0.24)</b>	0.29(0.23)	0.19(0.19)
3	0.00(0.00)	0.00(0.00)	0.06(0.15)	<b>0.40(0.25)</b>	0.27(0.18)	0.16(0.16)	0.11(0.15)
4	0.00(0.00)	0.00(0.00)	0.02(0.10)	<b>0.52(0.26)</b>	0.28(0.19)	0.13(0.13)	0.05(0.08)

Tabla 6.59: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ ,  
Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.51(0.31)</b>	0.34(0.28)	0.14(0.23)
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.46(0.32)</b>	0.37(0.28)	0.17(0.22)
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.01)	<b>0.61(0.29)</b>	0.29(0.26)	0.10(0.18)
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.13)	<b>0.66(0.28)</b>	0.26(0.24)	0.06(0.10)

Tabla 6.60: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ ,  
Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.35(0.26)</b>	0.32(0.15)	0.16(0.11)	0.08(0.10)	0.05(0.08)	0.03(0.06)
2	0.04(0.16)	<b>0.50(0.26)</b>	0.26(0.15)	0.11(0.12)	0.06(0.10)	0.02(0.05)	0.01(0.03)
3	0.00(0.00)	<b>0.36(0.27)</b>	0.27(0.15)	0.16(0.13)	0.10(0.12)	0.06(0.08)	0.05(0.10)
4	0.00(0.00)	<b>0.57(0.29)</b>	0.28(0.19)	0.10(0.11)	0.04(0.06)	0.01(0.02)	0.00(0.01)

Tabla 6.61: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ ,  
Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.20(0.26)	<b>0.35(0.24)</b>	0.26(0.24)	0.19(0.22)
2	0.00(0.00)	0.00(0.00)	0.00(0.02)	0.15(0.20)	<b>0.40(0.27)</b>	0.28(0.24)	0.17(0.19)
3	0.00(0.00)	0.00(0.00)	0.05(0.14)	<b>0.37(0.25)</b>	0.31(0.21)	0.17(0.19)	0.10(0.14)
4	0.00(0.00)	0.00(0.00)	0.07(0.17)	<b>0.46(0.28)</b>	0.29(0.22)	0.13(0.17)	0.05(0.09)

Tabla 6.62: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ ,  
Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.43(0.39)</b>	0.34(0.34)	0.23(0.32)
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.45(0.34)</b>	0.36(0.30)	0.19(0.25)
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.55(0.34)</b>	0.33(0.31)	0.11(0.20)
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.53(0.36)</b>	0.31(0.30)	0.16(0.25)

Tabla 6.63: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ ,  
Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	<b>0.31(0.25)</b>	0.30(0.15)	0.19(0.14)	0.10(0.10)	0.06(0.09)	0.04(0.07)
2	0.02(0.09)	<b>0.47(0.25)</b>	0.28(0.17)	0.13(0.12)	0.06(0.07)	0.03(0.04)	0.01(0.03)
3	0.00(0.00)	0.30(0.25)	<b>0.31(0.19)</b>	0.17(0.13)	0.11(0.10)	0.06(0.07)	0.05(0.12)
4	0.00(0.03)	<b>0.58(0.28)</b>	0.27(0.18)	0.10(0.12)	0.03(0.04)	0.01(0.02)	0.00(0.01)

Tabla 6.64: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.05(0.14)	0.32(0.25)	<b>0.33(0.23)</b>	0.30(0.26)
2	0.00(0.00)	0.00(0.00)	0.00(0.03)	0.12(0.22)	<b>0.34(0.24)</b>	0.33(0.24)	0.21(0.22)
3	0.00(0.00)	0.00(0.00)	0.07(0.16)	<b>0.31(0.25)</b>	<b>0.31(0.22)</b>	0.18(0.18)	0.13(0.18)
4	0.00(0.00)	0.00(0.00)	0.05(0.17)	<b>0.41(0.31)</b>	0.29(0.24)	0.18(0.20)	0.07(0.12)

Tabla 6.65: Prob. finales con  $k = 5, (\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.53(0.36)</b>	0.31(0.30)	0.16(0.25)
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.52(0.36)</b>	0.33(0.33)	0.15(0.23)
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	<b>0.60(0.34)</b>	0.28(0.29)	0.12(0.22)
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.06)	<b>0.60(0.34)</b>	0.30(0.30)	0.09(0.19)

Tabla 6.66: Prob. finales con  $k = 7, \eta'_s = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.14 (0.18)	<b>0.31(0.17)</b>	0.23(0.13)	0.16(0.13)	0.11(0.13)	0.05(0.07)
2	0.00(0.00)	0.12(0.18)	<b>0.26(0.16)</b>	0.23(0.16)	0.18(0.15)	0.11(0.11)	0.10(0.14)
3	0.00(0.00)	0.20(0.23)	<b>0.32(0.17)</b>	0.22(0.16)	0.13(0.14)	0.07(0.11)	0.06(0.13)
4	0.00(0.00)	0.09(0.13)	<b>0.29(0.18)</b>	0.24(0.16)	0.17(0.15)	0.12(0.15)	0.09(0.13)
5	0.00(0.00)	0.20(0.23)	<b>0.31(0.18)</b>	0.20(0.14)	0.13(0.11)	0.10(0.13)	0.07(0.08)

Tabla 6.67: Prob. finales con  $k = 7, \eta'_s = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.10)	0.08(0.16)	0.31(0.26)	<b>0.61(0.29)</b>
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.04(0.13)	0.18(0.21)	<b>0.40(0.24)</b>	0.38(0.27)
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.06)	0.21(0.25)	0.38(0.28)	<b>0.39(0.30)</b>
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.10)	0.13(0.21)	0.34(0.26)	<b>0.50(0.29)</b>
5	0.00(0.00)	0.00(0.00)	0.00(0.02)	0.04(0.09)	0.17(0.19)	0.38(0.25)	<b>0.40(0.28)</b>

Tabla 6.68: Prob. finales con  $k = 7$ ,  $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.03(0.14)	<b>0.97(0.14)</b>
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.07(0.19)	<b>0.93(0.19)</b>
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.04(0.14)	<b>0.96(0.14)</b>
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.07(0.20)	<b>0.93(0.20)</b>
5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.13)	<b>0.98(0.13)</b>

Tabla 6.69: Prob. finales con  $k = 7$ ,  $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.03(0.12)	<b>0.52(0.20)</b>	0.24(0.12)	0.10(0.07)	0.06(0.12)	0.03(0.05)	0.02(0.09)
2	0.01(0.04)	<b>0.47(0.22)</b>	0.28(0.14)	0.13(0.10)	0.06(0.06)	0.04(0.04)	0.03(0.08)
3	0.00(0.02)	<b>0.55(0.23)</b>	0.25(0.14)	0.10(0.12)	0.05(0.08)	0.03(0.10)	0.01(0.03)
4	0.02(0.11)	<b>0.48(0.21)</b>	0.25(0.12)	0.14(0.13)	0.06(0.07)	0.02(0.02)	0.02(0.04)
5	0.10(0.20)	<b>0.56(0.20)</b>	0.20(0.12)	0.07(0.07)	0.04(0.07)	0.01(0.01)	0.01(0.02)

Tabla 6.70: Prob. finales con  $k = 7$ ,  $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.04(0.10)	<b>0.35(0.23)</b>	0.30(0.19)	0.21(0.21)	0.10(0.12)
2	0.00(0.00)	0.00(0.00)	0.10(0.20)	<b>0.40(0.25)</b>	0.24(0.20)	0.16(0.18)	0.10(0.15)
3	0.00(0.00)	0.00(0.03)	0.10(0.18)	<b>0.40(0.24)</b>	0.28(0.20)	0.14(0.15)	0.08(0.13)
4	0.00(0.00)	0.00(0.00)	0.08(0.15)	<b>0.38(0.23)</b>	0.27(0.21)	0.19(0.20)	0.07(0.09)
5	0.00(0.00)	0.00(0.01)	0.17(0.24)	<b>0.37(0.22)</b>	0.22(0.16)	0.16(0.21)	0.08(0.14)

Tabla 6.71: Prob. finales con  $k = 7$ ,  $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.24(0.38)	<b>0.76(0.38)</b>
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.13)	0.26(0.36)	<b>0.72(0.38)</b>
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.07)	0.32(0.39)	<b>0.66(0.40)</b>
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.24(0.36)	<b>0.76(0.36)</b>
5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.14)	0.26(0.38)	<b>0.71(0.39)</b>

Tabla 6.72: Prob. finales con  $k = 7$ ,  $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.02)	<b>0.44(0.22)</b>	0.27(0.13)	0.14(0.11)	0.09(0.12)	0.03(0.05)	0.02(0.04)
2	0.00(0.00)	<b>0.43(0.23)</b>	0.27(0.15)	0.13(0.13)	0.08(0.13)	0.07(0.14)	0.02(0.05)
3	0.00(0.03)	<b>0.50(0.22)</b>	0.26(0.13)	0.13(0.13)	0.06(0.09)	0.03(0.04)	0.02(0.07)
4	0.00(0.00)	<b>0.39(0.18)</b>	0.30(0.17)	0.15(0.12)	0.09(0.11)	0.05(0.08)	0.02(0.03)
5	0.04(0.09)	<b>0.49(0.20)</b>	0.23(0.13)	0.13(0.13)	0.06(0.08)	0.02(0.04)	0.02(0.07)

Tabla 6.73: Prob. finales con  $k = 7$ ,  $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.02)	0.17(0.18)	0.30(0.21)	<b>0.33(0.23)</b>	0.19(0.18)
2	0.00(0.00)	0.00(0.00)	0.02(0.07)	0.23(0.20)	<b>0.34(0.22)</b>	0.23(0.17)	0.18(0.20)
3	0.00(0.00)	0.00(0.00)	0.04(0.10)	0.26(0.22)	<b>0.32(0.23)</b>	0.23(0.21)	0.15(0.19)
4	0.00(0.00)	0.00(0.00)	0.01(0.03)	0.21(0.21)	<b>0.30(0.21)</b>	0.29(0.20)	0.20(0.17)
5	0.00(0.00)	0.00(0.00)	0.05(0.12)	0.29(0.25)	<b>0.30(0.20)</b>	0.23(0.20)	0.13(0.16)

Tabla 6.74: Prob. finales con  $k = 7$ ,  $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.02(0.14)	0.26(0.38)	<b>0.71(0.39)</b>
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.01)	0.24(0.36)	<b>0.75(0.36)</b>
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.01)	0.24(0.34)	<b>0.76(0.34)</b>
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.24(0.38)	<b>0.76(0.38)</b>
5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.06)	0.23(0.34)	<b>0.76(0.34)</b>

Tabla 6.75: Prob. finales con  $k = 7$ ,  $\eta's = (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5)$ , Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.17(0.23)	<b>0.30(0.16)</b>	0.21(0.16)	0.17(0.16)	0.09(0.11)	0.05(0.07)
2	0.00(0.00)	0.13(0.19)	<b>0.32(0.23)</b>	0.22(0.21)	0.18(0.21)	0.10(0.16)	0.05(0.06)
3	0.00(0.00)	0.31(0.27)	<b>0.33(0.16)</b>	0.19(0.16)	0.09(0.12)	0.05(0.08)	0.03(0.09)
4	0.00(0.00)	0.13(0.19)	<b>0.30(0.20)</b>	0.23(0.18)	0.19(0.23)	0.09(0.12)	0.07(0.11)
5	0.00(0.00)	0.14(0.19)	<b>0.33(0.16)</b>	0.21(0.13)	0.16(0.17)	0.10(0.11)	0.07(0.12)

Tabla 6.76: Prob. finales con  $k = 7$ ,  $\eta's = (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5)$ , Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.11(0.18)	0.27(0.24)	<b>0.34(0.25)</b>	0.27(0.24)
2	0.00(0.00)	0.00(0.00)	0.01(0.06)	0.18(0.21)	<b>0.30(0.22)</b>	<b>0.30(0.22)</b>	0.20(0.20)
3	0.00(0.00)	0.00(0.00)	0.01(0.05)	0.10(0.18)	<b>0.33(0.23)</b>	0.31(0.20)	0.26(0.23)
4	0.00(0.00)	0.00(0.00)	0.00(0.10)	0.21(0.23)	0.27(0.19)	<b>0.28(0.22)</b>	0.24(0.24)
5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.16(0.20)	<b>0.33(0.24)</b>	0.29(0.23)	0.22(0.23)

Tabla 6.77: Prob. finales con  $k = 7$ ,  $\eta's = (0.1, 0.4, 0.6, 1.0, 2.7, 3.0, 3.5)$ , Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.17(0.31)	<b>0.83(0.31)</b>
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.01)	0.20(0.33)	<b>0.80(0.33)</b>
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.27(0.35)	<b>0.73(0.35)</b>
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.29(0.37)	<b>0.71(0.37)</b>
5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.23(0.33)	<b>0.77(0.33)</b>

Tabla 6.78: Prob. finales con  $k = 7$ ,  $\eta's = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)$ , Inc.=  $3/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.01)	<b>0.36(0.25)</b>	0.33(0.15)	0.16(0.13)	0.08(0.08)	0.04(0.07)	0.02(0.05)
2	0.00(0.00)	<b>0.33(0.24)</b>	0.28(0.12)	0.17(0.11)	0.11(0.11)	0.06(0.08)	0.05(0.12)
3	0.00(0.06)	<b>0.46(0.26)</b>	0.30(0.14)	0.14(0.11)	0.05(0.07)	0.03(0.08)	0.01(0.02)
4	0.00(0.00)	0.29(0.25)	<b>0.30(0.15)</b>	0.18(0.14)	0.11(0.10)	0.07(0.10)	0.04(0.07)
5	0.00(0.01)	<b>0.49(0.26)</b>	0.27(0.12)	0.14(0.13)	0.05(0.07)	0.04(0.11)	0.01(0.04)

Tabla 6.79: Prob. finales con  $k = 7$ ,  $\eta's = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)$ , Inc.=  $30/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.06(0.11)	0.31(0.26)	<b>0.33(0.27)</b>	0.30(0.27)
2	0.00(0.00)	0.00(0.00)	0.01(0.03)	0.20(0.24)	0.27(0.22)	<b>0.32(0.27)</b>	0.21(0.22)
3	0.00(0.00)	0.00(0.00)	0.02(0.08)	0.15(0.21)	<b>0.38(0.23)</b>	0.24(0.19)	0.21(0.21)
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.13(0.20)	<b>0.32(0.25)</b>	0.30(0.23)	0.25(0.24)
5	0.00(0.00)	0.00(0.00)	0.01(0.07)	0.19(0.24)	<b>0.31(0.26)</b>	0.30(0.29)	0.18(0.23)

Tabla 6.80: Prob. finales con  $k = 7$ ,  $\eta's = (0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5)$ , Inc.=  $300/10^4$  hab.

Esc.	k supuesto						
	1	2	3	4	5	6	7
1	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.05(0.19)	<b>0.95(0.19)</b>
2	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.01(0.01)	0.14(0.30)	<b>0.85(0.31)</b>
3	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.12(0.28)	<b>0.88(0.28)</b>
4	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.09(0.26)	<b>0.91(0.26)</b>
5	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.00(0.00)	0.14(0.31)	<b>0.86(0.31)</b>

## A.4 Porcentaje de correspondencia

También se hizo un estudio del porcentaje de correspondencia entre el máximo de la probabilidad final de  $k$  con el  $k$  real correspondiente en cada un de los escenarios. En concordancia con el apartado anterior, encontramos que en muchos escenarios el porcentaje es mayor en la columna del  $k$  real estudiado, mostrando así las bondades del modelo en cuanto a recuperación de la estructura verdadera. Igual que antes estos resultados también están afectados por el cambio que genera el aumento de la incidencia de la enfermedad en una población.

Tabla 6.81: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=1.  $\eta = 1$ ,  $p = 1$

Incidencia/k supuesto	1	2	3	4	5	6	7
$3/10^4$ hab	<b>99</b>	1	0	0	0	0	0
$30/10^4$ hab	<b>100</b>	0	0	0	0	0	0
$300/10^4$ hab	<b>100</b>	0	0	0	0	0	0

Tabla 6.82: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2,  $(\eta_1, \eta_2) = (0.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	<b>88</b>	7	4	0	1	0
2	0	<b>95</b>	3	2	0	0	0
3	0	<b>78</b>	13	5	2	1	1
4	5	<b>94</b>	1	0	0	0	0

Tabla 6.83: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k_{\text{real}}=2$ ,  $(\eta_1, \eta_2) = (0.5, 3.5)$ ,  $\text{Inc.} = 30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	98	2	0	0	0	0
2	0	100	0	0	0	0	0
3	0	99	1	0	0	0	0
4	0	100	0	0	0	0	0

Tabla 6.84: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k_{\text{real}}=2$ ,  $(\eta_1, \eta_2) = (0.5, 3.5)$ ,  $\text{Inc.} = 300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	100	0	0	0	0	0
2	0	100	0	0	0	0	0
3	0	100	0	0	0	0	0
4	0	100	0	0	0	0	0

Tabla 6.85: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k_{\text{real}}=2$ ,  $(\eta_1, \eta_2) = (0.5, 2.5)$ ,  $\text{Inc.} = 3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	85	11	4	0	0	0
2	0	98	2	0	0	0	0
3	0	85	8	4	2	1	0
4	20	77	2	1	0	0	0

Tabla 6.86: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k_{\text{real}}=2$ ,  $(\eta_1, \eta_2) = (0.5, 2.5)$ ,  $\text{Inc.} = 30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	98	2	0	0	0	0
2	0	99	1	0	0	0	0
3	0	98	2	0	0	0	0
4	0	99	1	0	0	0	0

Tabla 6.87: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k_{\text{real}}=2$ ,  $(\eta_1, \eta_2) = (0.5, 2.5)$ ,  $\text{Inc.} = 300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	100	0	0	0	0	0
2	0	100	0	0	0	0	0
3	0	100	0	0	0	0	0
4	0	100	0	0	0	0	0

Tabla 6.88: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2  $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	6	86	5	0	2	1	0
2	37	60	2	0	1	0	0
3	26	74	1	0	0	0	0
4	73	26	2	0	0	0	0

Tabla 6.89: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2  $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	97	3	0	0	0	0
2	0	99	1	0	0	0	0
3	0	98	2	0	0	0	0
4	0	99	0	1	0	0	0

Tabla 6.90: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2  $(\eta_1, \eta_2) = (0.5, 1.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	99	1	0	0	0	0
2	0	100	0	0	0	0	0
3	0	100	0	0	0	0	0
4	0	100	0	0	0	0	0

Tabla 6.91: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2  $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	95	4	1	0	0	0	0
2	97	3	0	0	0	0	0
3	96	3	1	0	0	0	0
4	100	0	0	0	0	0	0

Tabla 6.92: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2  $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	1	97	2	0	0	0	0
2	28	71	1	0	0	0	0
3	15	81	4	0	0	0	0
4	68	32	0	0	0	0	0

Tabla 6.93: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=2  $(\eta_1, \eta_2) = (0.8, 1.2)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	100	0	0	0	0	0
2	0	100	0	0	0	0	0
3	0	100	0	0	0	0	0
4	0	99	1	0	0	0	0

Tabla 6.94: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	55	36	5	3	0	1
2	0	68	25	5	1	1	0
3	6	68	17	5	1	3	0
4	0	92	8	0	0	0	0
5	0	63	22	10	2	2	1

Tabla 6.95: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	85	9	3	1	2
2	0	0	96	3	1	0	0
3	0	0	91	6	2	1	0
4	0	9	90	1	0	0	0
5	0	1	90	6	1	2	0

Tabla 6.96: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	100	0	0	0	0
2	0	0	99	1	0	0	0
3	0	0	98	2	0	0	0
4	0	0	99	1	0	0	0
5	0	0	99	1	0	0	0

Tabla 6.97: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	1	80	12	3	2	2	0
2	1	88	8	1	0	1	1
3	55	39	4	1	0	1	0
4	26	70	3	1	0	0	0
5	2	77	13	4	2	0	2

Tabla 6.98: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	1	80	12	3	2	2	0
2	1	88	8	1	0	1	1
3	55	39	4	1	0	1	0
4	26	70	3	1	0	0	0
5	2	77	13	4	2	0	2

Tabla 6.99: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.5, 1.0, 2.0)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	98	1	1	0	0
2	0	0	98	2	0	0	0
3	0	0	99	1	0	0	0
4	0	0	99	1	0	0	0
5	0	0	99	1	0	0	0

Tabla 6.100: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	85	15	0	0	0	0	0
2	85	15	0	0	0	0	0
3	98	2	0	0	0	0	0
4	88	12	0	0	0	0	0
5	84	14	1	1	0	0	0

Tabla 6.101: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	85	15	0	0	0	0	0
2	85	15	0	0	0	0	0
3	98	2	0	0	0	0	0
4	88	12	0	0	0	0	0
5	84	14	1	1	0	0	0

Tabla 6.102: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (0.8, 1.0, 1.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	94	6	0	0	0
2	0	0	95	5	0	0	0
3	0	1	95	4	0	0	0
4	0	13	86	0	1	0	0
5	0	0	94	6	0	0	0

Tabla 6.103: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	76	17	3	3	1	0
2	0	92	8	0	0	0	0
3	24	71	4	1	0	0	0
4	2	90	7	1	0	0	0
5	1	79	15	3	1	0	1

Tabla 6.104: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	76	17	3	3	1	0
2	0	92	8	0	0	0	0
3	24	71	4	1	0	0	0
4	2	90	7	1	0	0	0
5	1	79	15	3	1	0	1

Tabla 6.105: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=3,  $(\eta_1, \eta_2, \eta_3) = (1.0, 1.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	99	1	0	0	0
2	0	0	99	1	0	0	0
3	0	0	99	1	0	0	0
4	0	0	99	1	0	0	0
5	0	0	97	3	0	0	0

Tabla 6.106: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	20	46	21	5	7	1
2	0	37	46	8	4	5	0
3	0	27	33	19	9	5	7
4	0	62	30	4	3	1	0

Tabla 6.107: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	20	35	28	13	4
2	0	0	19	52	20	4	5
3	0	0	59	28	7	5	1
4	0	0	55	34	9	2	0

Tabla 6.108: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 3.0, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	51	36	13
2	0	0	0	3	54	32	11
3	0	0	0	2	62	31	5
4	0	0	0	2	67	23	8

Tabla 6.109: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	35	41	8	9	3	4
2	2	51	27	11	4	2	3
3	1	46	31	11	9	2	1
4	0	75	21	3	1	0	0

Tabla 6.110: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	5	47	34	10	4
2	0	0	5	46	30	12	7
3	0	0	29	44	12	11	4
4	0	1	33	55	9	2	0

Tabla 6.111: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 0.7, 1.0, 1.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	42	45	13
2	0	0	0	1	51	37	11
3	0	0	0	3	63	27	7
4	0	0	0	3	64	26	7

Tabla 6.112: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	28	42	19	5	5	1
2	8	53	26	8	3	1	2
3	0	39	25	21	6	8	1
4	0	65	30	2	3	0	0

Tabla 6.113: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	16	41	32	11
2	0	0	0	14	45	30	11
3	0	0	6	50	23	13	8
4	0	0	1	67	25	6	1

Tabla 6.114: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.2, 0.7, 1.0, 1.5, 2.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	58	32	10
2	0	0	0	0	54	31	15
3	0	0	0	0	72	23	5
4	0	0	0	2	78	19	1

Tabla 6.115: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	44	38	10	5	2	1
2	4	67	18	7	3	1	0
3	0	47	30	14	5	1	3
4	0	71	25	3	1	0	0

Tabla 6.116: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	22	37	21	20
2	0	0	0	18	42	28	12
3	0	0	6	44	31	15	4
4	0	0	8	55	25	11	1

Tabla 6.117: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 2.0, 2.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	45	35	20
2	0	0	0	0	49	34	17
3	0	0	0	0	61	31	8
4	0	0	0	1	55	37	7

Tabla 6.118: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	45	32	15	5	1	3
2	2	63	26	7	1	1	0
3	0	42	38	9	4	2	5
4	0	72	25	3	0	0	0

Tabla 6.119: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	5	35	32	28
2	0	0	0	13	38	31	18
3	0	0	7	38	31	15	9
4	0	0	6	48	29	15	3

Tabla 6.120: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=5,  $(\eta_1, \eta_2, \eta_3, \eta_4, \eta_5) = (0.5, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	57	30	13
2	0	0	0	0	54	36	10
3	0	0	0	0	64	27	9
4	0	0	0	1	65	28	6

Tabla 6.121: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	17	34	23	15	8	3
2	0	13	35	21	18	6	7
3	0	22	43	22	8	2	3
4	0	6	45	27	12	4	6
5	0	25	43	16	8	5	3

Tabla 6.122: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	5	21	74
2	0	0	0	4	16	41	39
3	0	0	0	1	21	40	38
4	0	0	0	2	9	33	56
5	0	0	0	2	14	38	46

Tabla 6.123: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.2, 0.5, 0.8, 1.0, 1.5, 2.5, 3.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	0	2	98
2	0	0	0	0	0	6	94
3	0	0	0	0	0	2	98
4	0	0	0	0	0	6	94
5	0	0	0	0	0	2	98

Tabla 6.124: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	1	82	12	1	2	1	1
2	0	73	20	4	1	0	2
3	0	80	13	3	3	1	0
4	2	73	14	9	1	0	1
5	9	81	6	1	3	0	0

Tabla 6.125: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	2	44	25	21	8
2	0	0	12	47	21	10	10
3	0	0	10	50	27	9	4
4	0	0	6	51	27	15	1
5	0	0	20	42	15	16	7

Tabla 6.126: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.5, 0.7, 0.9, 1.0, 1.5, 2.0, 2.5)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	0	22	78
2	0	0	0	0	2	24	74
3	0	0	0	0	0	31	69
4	0	0	0	0	0	24	76
5	0	0	0	0	3	26	71

Tabla 6.127: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.=  $3/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	64	22	7	6	1	0
2	0	61	24	7	4	3	1
3	0	73	17	5	4	0	1
4	0	64	24	7	3	3	0
5	0	75	13	8	1	1	2

Tabla 6.128: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.=  $30/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	16	32	36	16
2	0	0	2	26	39	17	17
3	0	0	3	28	35	24	10
4	0	0	1	20	29	33	18
5	0	0	4	35	34	19	8

Tabla 6.129: Porcentaje de correspondencia entre el máximo de la probabilidad final con  $k$  real=7,  $\eta's = (0.2, 0.4, 0.6, 1.0, 1.3, 1.5, 2.0)$ , Inc.=  $300/10^4$  hab.

Escenarios	k supuesto						
	1	2	3	4	5	6	7
1	0	0	0	0	0	9	91
2	0	0	0	0	0	23	77
3	0	0	0	0	0	20	80
4	0	0	0	0	0	23	77
5	0	0	0	0	1	19	80

# Bibliografía

- C. Anderson, D. Lee, and N. Dean. Identifying clusters in bayesian disease mapping. *Biostatistics*, 15:457–469, 2014.
- M. Barceló, M. Saez, G. Cano Serral, M. Martínez-Beneito, J. Martínez, C. Borrrell, R. Ocaña, I. Montoya, M. Calvo, G. López-Abente, M. Rodríguez Sanz, S. Toro, J. Alcalá, C. Saurina, P. Sánchez-Villegas, and A. Figueiras. Métodos para la suavización de indicadores de mortalidad: aplicación al análisis de desigualdades en mortalidad en ciudades del Estado Español (Proyecto MEDEA). *Gaceta Sanitaria*, 22:596–608, 2008.
- J. Bernardo and F. Smith. *Bayesian theory*. Wiley, New York, 1994.
- J. Besag, J. York, and A. Mollié. Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1–59, 1991.
- N. Best, S. Richardson, and A. Thomson. A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14:35–59, 2005.
- P. Botella. *Suavización espacio-temporal en cartografía de enfermedades*. PhD thesis, Universidad de Valencia, Facultad de Matemáticas, Departamento de Estadística e Investigación operativa, 2010.
- P. Botella, A. López-Quílez, and M. Martínez-Beneito. Spatial moving average risk smoothing. *Stat. Med.*, 32:2595–2612, 2013.
- G. Casella. Explaining the gibbs sampler. *The American Statistician*, 46:167–174, 1992.
- D. Denison and C. Holmes. Bayesian partitioning for estimating disease risk. *Biometrics*, 57:143–149, 2001.
- D. Denison, N. Adams, C. Holmes, and D. Hand. Bayesian partition modelling. *Computational statistics & data analysis*, 38:475–485, 2002.
- R. Doll. The epidemiology of cancer. *Cancer*, 45:2475–2485, 1980.

- P. Elliott, J. Wakefield, N. Best, and D. Briggs. *Spatial Epidemiology: Methods and Applications*. Oxford University Press, 2000.
- J. Etxeberria. *Space-Time disease mapping: modelling and forecasting*. PhD thesis, Universidad Pública de Navarra, Facultad de Matemáticas, Departamento de Estadística, 2012.
- C. Fernández and P. Green. Modelling spatially correlated data via mixtures: a bayesian approach. *Journal of the Royal Statistical Society Series B*, 64: 805–826, 2002.
- J. Ferreira, D. Denison, and C. Holmes. *Partition modelling*. Citeseer, 2002.
- J. Ferrándiz, J. Abellán, A. López-Quílez, P. Sanmartín, H. Vanaclocha, O. Zurriaga, M. Martínez-Beneito, I. Melchor, and J. Calabuig. *GIS for Emergency preparedness and health risk reduction*. Springer Netherlands, Dordrecht, 2002.
- J. Ferrándiz, J. Abellán, V. Gómez-Rubio, A. López-Quílez, P. Sanmartín, C. Abellán, M. Martínez-Beneito, I. Melchor, H. Vanaclocha, O. Zurriaga, F. Ballester, J. Gil, S. Pérez-Hoyos, and R. Ocaña. Spatial analysis of the relationship between mortality from cardiovascular and cerebrovascular disease and drinking water hardness. *Environmental Health Perspectives*, 112:1037–1044, 2004.
- J. Fleiss. *The Measurement of Interrater Agreement, Statistical Methods for Rates and Proportions, Second Edition*. John Wiley & Sons, Inc., New York, 1981.
- F. Forbes, M. Charras-Garrido, L. Azizi, S. Doyle, and D. Abrial. Spatial risk mapping for rare disease with hidden markov fields and variational em. *Ann. Appl. Stat.*, 7:1192–1216, 2013.
- R. Gangnon. Impact of prior choice on local Bayes factors for cluster detection. *Statistics in medicine*, 25:883–895, 2006.
- R. Gangnon and M. Clayton. Bayesian detection and modeling of spatial disease clustering. *Biometrics*, 56:922–935, 2000.
- A. Gelman. Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1:515–534, 2006.
- S. Geman and D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741, 1984.
- J. Geweke. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. Technical report, Staff Report 148, Federal Reserve Bank of Minneapolis, Feb 1991.

- P. Green and S. Richardson. Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97(460):1055–1070, 2002.
- W. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.
- M. Hossain and A. Lawson. Local likelihood disease clustering: Development and evaluation. *Environmental and Ecological Statistics*, 12:259–273, 2005.
- M. Hossain and A. Lawson. Cluster detection diagnostics for small area health data: with reference to evaluation of local likelihood models. *Statistics in Medicine*, 25(5):771–786, 2006.
- M. Hossain and A. Lawson. Space-time Bayesian small area disease risk models: development and evaluation with a focus on cluster detection. *Environmental and Ecological Statistics*, 17(1):73–95, 2010.
- A. Iftimi, F. Martínez-Ruíz, A. Santiyán, and F. Montes. Spatio-temporal cluster detection of chickenpox in Valencia, Spain in the period 2008-2012. *Geospatial health*, 10(1), 2015a.
- A. Iftimi, F. Montes, A. Santiyán, and F. Martínez-Ruíz. Space-time airborne disease mapping applied to detect specific behaviour of varicella in Valencia, Spain. *Spatial and spatio-temporal epidemiology*, 14:33–44, 2015b.
- L. Knorr-Held. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine*, 19:2555–2567, 2000.
- L. Knorr-Held and G. Rasser. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56:13–21, 2000.
- M. Kulldorff. A spatial scan statistic. *Commun Stat A Theory Methods*, 26:1481–1496, 1997.
- M. Kulldorff, R. Heffernan, J. Hartman, R. Assuncao, and F. Mostashari. A spacetime permutation scan statistic for disease outbreak detection. *PLoS Med*, 2, 2005. URL <http://dx.doi.org/10.1371/journal.pmed.0020059>.
- A. Lawson. *Statistical Methods in Spatial Epidemiology 2nd Edition*. Wiley, New York, 2006.
- A. Lawson. *Bayesian disease mapping. Hierarchical modeling in spatial epidemiology*. CRC Press. Taylor & Francis Group. A chapman & hall book, New York, 2009.
- A. Lawson. Hotspot detection and clustering: ways and means. *Environmental and Ecological Statistics*, 17(2):231–245, 2010.

- A. Lawson. *Bayesian disease mapping: Hierarchical modeling in spatial epidemiology*. CRC Press, New York, 2013.
- A. Lawson and A. Clark. *Markov Chain Monte Carlo methods for putative sources of hazard and general clustering*. In Lawson, A., Biggeri, A., Bohning, D., Lesafre, E., Viel, J. F., y Bertollini, R., editors, *Disease Mapping and Risk Assessment for Public Health*, Wiley, 119–142, New York, 1999.
- A. Lawson and A. Clark. Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21:359–370, 2002.
- A. Lawson and D. Denisson. Spatial cluster modelling: An overview. *Spatial Cluster Modelling. CRC*, 1:2217–2242, 2002.
- A. Lawson, A. Biggeri, D. Boehning, E. Lesafre, J. Viel, A. Clark, P. Schlattmann, and F. Divino. Disease mapping models: an empirical evaluation. *Statistics in Medicine*, 19(17/18):2217–2241, 2000.
- D. Lee and A. Lawson. Cluster detection and risk estimation for spatio-temporal health data. *Annals of the Institute of Statistical Mathematics*, 43:1–59, 2014.
- B. Leroux, X. Lei, and N. Breslow. Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical models in epidemiology, the environment and clinical trials. Springer, Berlin Heidelberg. Halloran M.E, Berry, D. (Eds)*, 116:179–192, 1999.
- N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44:335–341, 1949.
- A.F. Militino, M.D. Ugarte, and C. Dean. The use of mixture models for identifying high risks in disease mapping. *Statist. Med.*, 20:2035–2049, 2001.
- A. Mollié. *Bayesian mapping of disease*. In Gilks, W. R., Richardson, S., y Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall-CRC, 359–380, London, 1996.
- I. Peña-Rey, M. Martínez de Aragón, A. Villaverde, M. Terres, E. Alcalde, and B. Suárez. Epidemiología de la varicela en los períodos pre y postvacunación. *Rev Esp Salud Pública*, 83:711–724, 2009.
- S. Richardson, A. Thomson, N. Best, and P. Elliot. Interpreting posterior relative risk estimates in disease-mapping studies. *Environmental Health Perspectives*, 112(9):1016–1025, 2004.
- C. Robert and G. Casella. *Monte Carlo Statistical Methods (second edition)*. Springer-Verlag, New York, 2004.
- Ch. Rotejanaprasert. Evaluation of cluster recovery for small area relative risk models. *Statistical Methods in Medical Research*, 23:531–551, 2014.

- H. Rue and L. Held. *Gaussian Markov Random Fields: Theory & Applications*. Chapman & Hall/CRC, London, 2005.
- H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B*, 2, 71:319–392, 2009.
- D. Salmerón, J. A. Cano, and M. D. Chirlaque. Reducing Monte carlo error in the Bayesian estimation of risk ratios using log-binomial regression models. *Statistics in Medicine*, 34:2755–2767, 2015.
- P. Schlattman and D. Böhning. Mixture models and disease mapping. *Statistics in medicine*, 12:1943–1950, 1993.
- L. Silva, A. Benavides, and C. Vidal. Análisis espacial de la mortalidad en áreas geográficas pequeñas: El enfoque bayesiano. *Rev Cubana Salud Pública*, 29: 314–322, 2003.
- M.D. Ugarte, T. Goicoa, A.F. Militino, and M. Durbán. Spline smoothing in small area trend estimation and forecasting. *Computational Statistics and Data Analysis*, 53(10):3616–3629, 2009a. doi: 10.1016/j.csda.2009.02.027.
- M.D. Ugarte, T. Goicoa, and A.F. Militino. Empirical Bayes and Fully Bayes procedures to detect high-risk areas in disease mapping. *Computational Statistics & Data Analysis*, 53:2938–2949, 2009b.
- M.D. Ugarte, T. Goicoa, B. Ibañez, and A.F. Militino. Evaluating the performance of spatio-temporal Bayesian models in disease mapping. *Environmetrics*, 20: 647–665, 2009c.
- M.D. Ugarte, T. Goicoa, and A.F. Militino. Spatiotemporal modeling of mortality risks using penalized splines. *Environmetrics*, 21:270–289, 2010.
- J. von Neumann. Various techniques used in connection with random digits. monte carlo methods. *Nat. Bureau Standards*, 12:36–38, 1951.
- J. Wakefield and A. Kim. A Bayesian model for cluster detection. *Biostatistics*, 14:752–765, 2013.
- J. Wakefield and S. Morris. The Bayesian modeling of disease risk in relation to a point source. *Journal of the American Statistical Association*, 96:77–91, 2001.
- D. Wheeler. A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996 – 2003. *International Journal of Health Geographics*, pages 6–13, 2007.