UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

MODELOS JERÁRQUICOS BAYESIANOS ESPACIALES EN EPIDEMIOLOGÍA AGRÍCOLA

Tesis doctoral

Realizada por:

Nora Coromoto Monsalve Graterol

Valencia, 2013

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Departamento de Estadística
e Investigación Operativa Aplicadas y Calidad

MODELOS JERÁRQUICOS BAYESIANOS ESPACIALES EN EPIDEMIOLOGÍA AGRÍCOLA

Tesis doctoral

Realizada por: Nora Coromoto Monsalve Graterol

Dirigida por: Dr. D. Antonio López Quílez

Valencia, Enero 2013

D. Antonio López Quílez, profesor titular del Departamento de Estadística e Investigación Operativa de la Universitat de València CERTIFICA que la presente memoria de investigación:

"MODELOS JERÁRQUICOS BAYESIANOS ESPACIALES EN EPIDEMIOLOGÍA AGRÍCOLA"

ha sido realizada bajo su dirección por Nora Coromoto Monsalve Graterol, y constituye su tesis para optar al grado de Doctor.

Y para que así conste, en cumplimiento con la normativa vigente, autoriza su presentación ante el Departamento de Estadística e Investigación Operativa Aplicadas y Calidad de la Universidad Politécnica de Valencia para que pueda ser tramitada su lectura y defensa pública.

En Valencia, Enero 2013.

Fdo: Antonio López Quílez

Índice general

Índice de tablas

Índice de figuras

Lista de acrónimos

MCMC Métodos Markov Chain Monte Carlo 1

GLM Modelos Lineales Generalizados 9

GLMM Modelos Lineales Mixtos Generalizados 9

GLSM Modelo Espacial Lineal Generalizado 10

CODA Convergence Diagnostic and Output Analysis 26

CTV *Citrus tristeza virus* 41

CV Comunidad Valenciana 41

ICAR Distribución Autoregresiva Condicional Intrínseca 51

CAR Distribuciones Condicionales Autoregresivas 51

SAR Distribuciones Autoregresivas Simultáneas 53

HDSM Hierarchical Dynamic Spatial Model 64

ARMA Autoregressive Moving Average 81

ARIMA Autoregressive Integrated Moving Average 82

PH Proportional hazard model 92

WDTM Weibull Discrete Time Model 104

CMGPH Cox Model with Gamma process in baseline hazard 108

CMPFH Cox Model with polygonal function in baseline hazard 110

SPDE Stochastic Partial Diferential Equation 130

INLA Integrated Nested Laplace Approximation 130

Resumen

Esta tesis está basada en la modelización jerárquica espacial desde la perspectiva Bayesiana para el estudio de enfermedades en cultivos agrícolas. Esta metodología en Epidemiología agrícola es aún un campo poco desarrollado. La necesidad de controlar la variabilidad espacial presente en la mayoría de los datos observados en Agricultura, exige la búsqueda de nuevas alternativas de modelización capaces de recoger adecuadamente la estructura de interrelaciones entre los individuos estudiados. En este sentido, el objetivo general de la tesis es el aporte de herramientas de modelización generales en el ámbito del análisis espacial, que permitan estudiar la presencia de enfermedades en cultivos agrícolas y describan la distribución de los patrones de contagio cuando se tiene poca información y no se tienen covariables explicativas.

En los Capítulos 2 y 3 se abordan propuestas de modelización basadas en modelos jerárquicos para datos asociados en una red de localizaciones y se considera la componente temporal a través de una covariable que recoge la historia de la enfermedad en el tiempo. En especial, en el Capítulo 2, se construyen modelos con estructura espacial dinámica y en ellos se consideran fuentes de variabilidad no observadas (efecto de heterogeneidad); por otro lado, en el Capítulo 3, se presentan tres modelizaciones en el contexto de datos de supervivencia. Cada una de ellas, estima el tiempo de supervivencia de los individuos afectados por

la evolución de una enfermedad en el tiempo y por la presencia de heterogeneidad no observada. Gracias a la covariable dependiente del tiempo considerada en las tres modelizaciones y a la construcción de una estructura espacial dinámica (*frailty*), se puede relajar el supuesto de proporcionalidad generalmente asumido en el modelo de Cox y enmarcar tales propuestas en el contexto de modelos espacio-temporales.

En el capítulo 2, se demuestra que la dinámica de los riesgos está determinada por información que depende del pasado y por efectos aleatorios. Estos efectos recogen la variabilidad no observada (heterogeneidad) y la variabilidad espacial. Así mismo, en el capítulo 3, se demuestra que partiendo de datos observados en una red de localizaciones es posible construir modelos de supervivencia. Gracias a los tres modelos desarrollados en este capítulo, se puede pensar en modelar la función de riesgo (*hazard*) desde tres perspectivas diferentes. Comenzando por un modelo Weibull con tiempos discretizados sobre períodos de un año y continuamos con dos propuestas basadas en procesos de conteo. Estas dos últimas modelizaciones son distintas, ya que por un lado, se considera un proceso Gamma en la distribución a priori que define a la función de riesgo base y en la segunda se asignan funciones poligonales a este riesgo.

En el capítulo 4, se propone un modelo jerárquico capaz de predecir en cualquier punto de la región, la probabilidad o riesgo de enfermedad de un individuo en el contexto agrícola. Gracias a la metodología INLA-SPDE, es posible proponer un modelo de regresión aditivo con estructura espacial (dentro de la clase de modelos Gaussianos latentes) de variable respuesta Bernoulli controlado por pocos hiperparámetros.

Gracias a la metodología desarrollada en el capítulo 4, es posible hacer predicción (*kriging* Bayesiano) al considerar la ocurrencia del fenómeno en una región continua. Usando el *kriging* Bayesiano es posible incorporar en el modelo fuentes de incertidumbre asociadas a los parámetros de predicción y de esta forma encontrar estimaciones más realistas. Además,

es posible construir mapas de riesgos en los que se estima la incertidumbre tanto en lugares observados como en los no observados. La metodología INLA combinada con el enfoque SPDE, ofrece un marco teórico excelente para fenómenos que necesitan predicción. La ilustración de la metodología con datos reales permite reconocer su utilidad en estudios epidemiológicos no sólo en el contexto agrícola.

En general, las modelizaciones propuestas reconocen la existencia de correlación espacial a pequeña escala. Al ilustrar la metodología con datos reales, se reconoce la importancia de la variabilidad espacial y es gracias a ella que puede llegar a comprenderse la dinámica de contagio y el patrón de movilidad de los agentes causantes de la enfermedad en el contexto agrícola. Los modelos con mejores ajustes contienen en su estructura no sólo el efecto de la covariable con la historia de la enfermedad sino la influencia del efecto aleatorio espacial dinámico.

Para abordar problemas desde el contexto epidemiológico es necesario entender estadísticamente el proceso, para ello se necesitan modelos capaces de capturar heterogeneidad usualmente no observada y que generalmente no es explicada en las covariables disponibles. Pensar que los individuos son extraídos de una población homogénea, no es adecuado, especialmente en fenómenos donde existen factores de riesgo ocultos que gracias a la cercanía entre los individuos son compartidos. De esta forma parece adecuado, diseñar modelos jerárquicos que permitan tratar la heterogeneidad existente en la población en alguna de sus capas o niveles. Por lo tanto, un proceso espacial combinado con modelos jerárquicos y vistos desde el paradigma Bayesiano, permite la construcción de herramientas útiles en estudios epidemiológicos en cualquier contexto, y permiten estudiar la incidencia y extensión de fenómenos asociados a un proceso espacial. En particular, su utilidad queda demostrada en Agricultura.

Resum

Aquesta tesi està basada en la modelització jeràrquica espacial des de la perspectiva Bayesiana per a l'estudi de malalties en cultius agrícoles. Aquesta metodologia en Epidemiologia agrícola és encara un camp poc desenvolupat. La necessitat de controlar la variabilitat espacial present en la majoria de les dades observades en Agricultura, exigeix la recerca de noves alternatives de modelització capaces de recollir adequadament l'estructura d'interrelacions entre els individus estudiats. En aquest sentit, l'objectiu general de la tesi és aportar eines de modelització generals en l'àmbit de l'anàlisi espacial, que permeten estudiar la presència de malalties en cultius agrícoles i descriguen la distribució dels patrons de contagi quan es té poca informació i no es tenen covariables explicatives.

En els Capítols 2 i 3 s'aborden propostes de modelització basades en models jeràrquics per a dades associades en una xarxa de localitzacions i es considera la component temporal a tráves d'una covariable que recull la història de la malaltia en el temps. Especialment, en el Capítol 2, es construïxen models amb estructura espacial dinàmica i en ells es consideren fonts de variabilitat no observades (efecte d'heterogeneïtat); d'altra banda, en el Capítol 3, es presenten tres modelitzacions en el context de dades de supervivència. Cadascuna d'elles, estima el temps de supervivència dels individus afectats per l'evolució d'una malaltia en el temps i per la presència d'heterogeneïtat no observada. Gràcies a la

covariable depenent del temps considerada en les tres modelitzacions i a la construcció d'una estructura espacial dinàmica (*frailty*), es pot relaxar el supòsit de proporcionalitat generalment assumit en el model de Cox i emmarcar tals propostes en el context de models espai-temporals.

En el capítol 2, es demostra que la dinàmica dels riscos està determinada per informació que depén del passat i per efectes aleatoris. Aquests efectes recullen la variabilitat no observada (heterogeneïtat) i la variabilitat espacial. Així mateix, en el capítol 3, es demostra que partint de dades observades en una xarxa de localitzacions és possible construir models de supervivència. Gràcies als tres models desenvolupats en aquest capítol, es pot pensar a modelar la funció de risc (*hazard*) des de tres perspectives diferents. Començant per un model Weibull amb temps discretitzats sobre períodes d'un any i continuant amb dues propostes basades en processos de conteig. Aquestes dues últimes modelitzacions són distintes, ja que d'una banda, es considera un procés Gamma en la distribució a priori que defineix a la funció de risc base i en la segona s'assignen funcions poligonals a aquest risc.

En el capítol 4, es proposa un model jeràrquic capaç de predir en qualsevol punt de la regió, la probabilitat o risc de malaltia d'un individu en el context agrícola. Gràcies a la metodologia INLA-SPDE, és possible proposar un model de regressió additiu amb estructura espacial (dintre de la classe de models Gaussians latents) de variable resposta Bernoulli controlat per pocs hiperparàmetres.

Gràcies a la metodologia desenvolupada en el capítol 4, és possible fer predicció (*kriging* bayesià) en considerar l'ocurrència del fenomen en una regió contínua. Usant el *kriging* bayesià és possible incorporar en el model fonts d'incertesa associades als paràmetres de predicció i d'aquesta forma trobar estimacions més realistes. A més, és possible construir mapes de riscos en els quals s'estima la incertesa tant en llocs observats com en els no observats. La metodologia INLA combinada amb l'enfocament SPDE,

oferix un marc teòric excel·lent per a fenòmens que necessiten predicció. La il·lustració de la metodologia amb dades reals permet reconéixer la seua utilitat en estudis epidemiològics no només en el context agrícola.

En general, les modelitzacions proposades reconeixen l'existència de correlació espacial a petita escala. En il·lustrar la metodologia amb dades reals, es reconeix la importància de la variabilitat espacial i és gràcies a ella que pot arribar a comprendre's la dinàmica de contagi i el patró de mobilitat dels agents causants de la malaltia en el context agrícola. Els models amb millors ajusts contenen en la seua estructura no només l'efecte de la covariable amb la història de la malaltia sinó la influència de l'efecte aleatori espacial dinàmic.

Per a abordar problemes des del context epidemiològic és necessari entendre estadísticament el procés, per a això es necessiten models capaços de capturar heterogeneïtat usualment no observada i que generalment no és explicada en les covariables disponibles. Pensar que els individus són extrets d'una població homogènia, no és adequat, especialment en fenòmens on existeixen factors de risc ocults que gràcies a la proximitat entre els individus són compartits. D'aquesta forma sembla adequat, dissenyar models jeràrquics que permeten tractar l'heterogeneïtat existent en la població en alguna de les seues capes o nivells. Per tant, un procés espacial combinat amb models jeràrquics i vists des del paradigma bayesià, permet la construcció d'eines útils en estudis epidemiològics en qualsevol context, i permeten estudiar la incidència i extensió de fenòmens associats a un procés espacial. En particular, la seua utilitat queda demostrada en Agricultura.

Abstract

This thesis is based on Bayesian hierarchical spatial models for the study of diseases in agricultural groves. This methodology have been little used in agricultural Epidemiology . The need to control the spatial variability present in most of the observed data in agriculture, requires finding new ways of modeling capable to properly collect the structure of relationships between individuals studied. In this sense, the overall aim of the thesis is the contribution of general modeling tools in the field of spatial analysis for the study of the presence of a disease in agricultural groves and that help to describe the distribution patterns of infection when we have few data and in absence of explanatory variables.

In Chapters 2 and 3 we proposed hierarchical models capable to study data associated a lattice of fixed locations and in they are considered a temporal component through a covariate that collects the history of the disease over time. In particular, in Chapter 2, are constructed dynamic models with spatial structure and they are considered unobserved sources of variability (effect of heterogenity) on the other hand, in Chapter 3 we present three modeling in the context of survival data. In each of them, we estimate survival time of individuals affected by the evolution of a disease over time and by the presence of unobserved heterogeneity. Thanks to the time-dependent covariate considered in the three modelings and to building a dynamic spatial structure (frailty) is possible relax the restriction of the

proportional hazards Cox model. These proposals framed in the context of spatial-temporal models.

In Chapter 2, we show that the dynamic of risk is determined by information that depends of past (process history) and by a random effect of present. In these effects be reflect unobserved variability (heterogeneity) and spatial variability. Likewise, in the Chapter 3, we show that starting from observed data in a lattice of fixed locations is possible build survival models. Thanks to the three models developed in this chapter, we can think of modeling the hazard function from three different perspectives. We start with a Weibull model with discretized times over periods of one year and we continue with two proposals based on counting processes. These latter two modeling are distinct because on one hand is considered a Gamma process in the prior distribution that defined to the baseline hazard function and in the second is assigned a polygonal function to this baseline hazard.

In Chapter 4, we propose a hierarchical model capable to predict at any point in the region, the probability or risk of disease by one individual in the agricultural context. Thank the methodology SPDE-INLA, it is possible to propose a Structured Additive Regression model with spatial effect (known as Latent Gaussian model) with random variable Bernoulli controlled by a few hyperparameters.

With the methodology developed in Chapter 4 it is possible to predict (kriging Bayesian) the occurrence of a phenomenon in a continuous region. Using the kriging Bayesian we can incorporate sources of uncertainty associated with the prediction parameters which leads to more realistic and accurate estimates. It is also possible to build risk maps through which we can estimate the uncertainty both in places observed as well as unobserved. The INLA methodology combined with the SPDE approach provides an excellent theoretical framework for predicting phenomena. The illustration of the methodology with real data allows recognize its

usefulness in epidemiological studies not only in the agricultural context. In general the various proposals of modeling recognize the existence of a small-scale spatial correlation. The illustration the methodology with real data allows recognize the importance of spatial variability and it is thanks to her that we may come to understand the dynamics of a disease and the mobility pattern of disease causing agents in groves agricultural. The models with best fit have in their structure the effect of the covariate with the history of the disease and the influence of a dynamic spatial random effect.

Tackle problems from the epidemiological context requires us to understand the process statistically. Therefore, we need to design models capable of capturing unobserved heterogeneity that is not usually explained in the available covariates. To think that individuals are drawn from a homogeneous population is inadequate, especially in phenomena where there are hidden risk factors that are shared due to the proximity between subjects. Thus, design hierarchical models that allow us to represent the heterogeneity of the population in any of their layers or levels seems appropriate. Therefore study a spatial process using the hierarchical models from the Bayesian paradigm allows build useful tools in epidemiological studies in any context. Also allow us to study the incidence and the distribution of a phenomena associated with a spatial process. In particular, usefulness of methodology proposal is demonstrated in agriculture context.

Agradecimientos

Quiero comenzar dedicando este hermoso triunfo a Dios quien me acompaña y guía en todo momento.
A mi esposo Arnaldo, por su amor, paciencia y sabios consejos, gracias. Te amo.
A mis padres, por su amor. A ustedes les debo ser quien soy hoy día. Los amo.
A mi hermanita Luz María por ser un apoyo incondicional. Gracias por tu amor.
A mi suegra Doña Ana quien ha sido un gran apoyo en todos estos años. Gracias por su ejemplo.
A mis cuñados, Eyra, Rodolfo, Wilmer, Gustavo y José por su apoyo.
A mis amigos y colegas Venezolanos, gracias por su amistad.
A mis compañeros de despacho, Adriana, José María y Andrés por su amistad y compañía.
A Antonio López, quien ha sido mi maestro en todo este tiempo. Además de ser mi Director, ahora puedo decir que cuento con un nuevo amigo. Gracias infinitas.
A Ana Aparicio Gaitano por su ayuda, colaboración y palabras de aliento.
A Ana María Debón por su amistad y apertura.
A mis amigos Sudamericanos, Omar, Mónica, Nela y familia por abrirme la puerta de sus hogares. Gracias por su amistad y compañía.

A mi amigo Malon Mendoza, gracias por tu amistad y ayuda incondicional. Om Sai Ram.
A la UCLA, institución que con su apoyo económico ha hecho posible este logro.
A los que no menciono y que han contribuido de alguna manera con este logro. Muchas gracias.

A Dios por ser mi eterno conductor.
A mi esposo y a mi familia por todo su amor.

Capítulo 1

Introducción

1.1. Motivación

Los científicos a través de una amplia gama de disciplinas han reconocido la importancia de la dependencia espacial en los datos y el proceso subyacente de interés. En un principio debido a las limitaciones computacionales, se trataron tales dependencias por aleatorización y por bloqueo en lugar de la caracterización explícita de las dependencias.

Los primeros desarrollos en modelos espaciales comenzaron en los años 1950 y 1960, estuvieron motivados por problemas en ingeniería de minas y meteorología (Cressie, 1993), seguido por la introducción de campos aleatorios de Markov (Besag, 1974). En los últimos años del siglo 20, la aplicación de los modelos jerárquicos espaciales y espacio-temporales se han convertido en herramientas cada vez más populares gracias a los avances de las técnicas computacionales, tales como los Métodos Markov Chain Monte Carlo (MCMC).

Los métodos de modelado espacial y espacio-temporales son cada vez más importantes en las ciencias del medio ambiente y en otras ciencias, donde los datos se derivan de procesos en entornos espaciales. Desafortunada-

mente, la aplicación de los tradicionales modelos estadísticos espaciales basados en covarianza resultan inapropiados o computacionalmente ineficientes en muchos problemas. Por otro lado, los métodos convencionales a menudo son incapaces de permitir al investigador cuantificar la incertidumbre correspondiente a los parámetros del modelo, en especial, en modelos espaciales o espacio-temporales complejos donde el número de parámetros es mayor.

Un objetivo principal en la caracterización rigurosa de ciertos fenómenos es la estimación los parámetros que rigen los procesos y su predicción. Por lo tanto, es necesario contar con herramientas flexibles y capaces de acomodar relaciones complejas entre los datos y al mismo tiempo permitan la incorporación de las diversas fuentes de incertidumbre presentes en los fenómenos estudiados.

Los enfoques tradicionalmente basados en la verosimilitud han permitido modelar y comprender muchas estructuras de datos, sin embargo, en situaciones complicadas con modelos muy parametrizados y en presencia de pocos datos, la estimación por máxima verosimilitud es a menudo problemática o imposible. En los últimos años se han desarrollado métodos de aproximación numérica para afrontar estas limitaciones. Estos métodos han sido utilizados en muchos casos, especialmente en aquellos donde se tiene una alta dimensión en el espacio de parámetros, entre los métodos más conocidos se pueden mencionar el método Newton-Raphson y algoritmo E-M (Givens y Hoeting, 2005). Sin embargo estos métodos, en algunas situaciones pueden ser difíciles de implementar y no tienen lugar para acomodar la incertidumbre en múltiples niveles.

Las limitaciones de los métodos tradicionales pueden ser abordadas si representamos los problemas como modelos jerárquicos, este enfoque, permite descomponer el problema en una serie de niveles unidos por simples reglas de probabilidad. De esta forma se construye un marco de inferencia flexible y capaz de incorporar incertidumbre e información conocida en

forma a priori. Además conserva muchas ventajas del enfoque tradicional de verosimilitud, ya que considera múltiples fuentes de datos y estructuras de datos significativas en el modelo.

El desarrollo de los métodos Monte Carlo con cadenas de Markov y la introducción de modelos jerárquicos desde la perspectiva Bayesiana han generado una explosión de la investigación en diferentes áreas científicas, tanto en el contexto teórico como aplicado. Todo este avance se ha traducido en el desarrollo de complejos modelos jerárquicos Bayesianos. Este progreso ha ocurrido sólo en algunas áreas científicas, entre las que destacan, ciencias medioambientales, Medicina, Minería, Epidemiología en salud pública, restauración de imágenes, Ecología y Veterinaria (Biggeri et al. 2006). En campos como la Agricultura son pocos los trabajos enmarcados en esta metodología. Los métodos Bayesianos se adaptan fácilmente a la estimación de parámetros enlazados en un modelo jerárquico. Aún cuando es posible emplear métodos no Bayesianos para realizar estimaciones en modelos jerárquicos, a menudo requieren de supuestos adicionales y de tiempos de computación exigentes que hacen más difícil su inferencia, como por ejemplo, invertir matrices de covarianza densas.

Hay pocos trabajos dedicados a la Epidemiología en Agricultura que hagan uso de la metodología Bayesiana para representar estructuras de dependencia espacial contenida en las observaciones. En toda la tesis, se presentan estrategias generales de modelización asociadas con procesos espaciales referidos a datos en una red de localizaciones o a un proceso espacial continuo. Estas modelizaciones pueden ser aplicadas en cualquier contexto en donde se tengan datos espaciales de esta naturaleza y no sólo en el ámbito epidemiológico. Aún cuando las propuestas que desarrollamos analizan y modelan el comportamiento de enfermedades en plantas, esta metodología puede ser empleada en principio en cualquier individuo que conserve la disposición espacial tratada en cada una de

ellas. En el caso de individuos agregados o ubicados en puntos fijos, como: barrios, municipios, condados, latitud, longitud, altitud, entonces se puede pensar en estudiar el fenómeno con alguna de las modelizaciones que presentaremos en los capítulos 2 y 3 (proceso espacial en una red de localizaciones). Mientras que si los individuos están muy cerca geográficamente se puede pensar en un proceso de naturaleza continua y emplear la modelización que desarrollamos en el capítulo 4. En todas estas modelizaciones, la autocorrelación espacial se incluye en alguna de las capas del modelo. Trabajos como el presentado por Illian et al. (2009), en donde se emplean modelos jerárquicos Bayesianos para estudiar patrones puntuales multivariantes en comunidades de plantas con alta biodiversidad, demuestran que una aproximación bayesiana proporciona un marco flexible para incorporar información relativa a la interacción entre plantas.

Trabajos recientes como los de Finley, Banerjee y McRoberts (2009) exploran la potencialidad ofrecida por un modelo espacial multinomial de regresión logística para predecir zonas boscosas. En este mismo sentido, estos autores publican en el 2009, un trabajo en el que emplean modelos espaciales de regresión logística multinomial para estudiar y predecir especies de árboles en bosques. Demuestran que usando modelos jerárquicos desde el enfoque Bayesiano es posible combinar plenamente los datos georeferenciados disponibles y obtener buenas predicciones sobre grupos forestales ubicados en grandes paisajes forestales (procesos espaciales multivariados, uno para cada coeficiente de regresión), similar a lo demostrado por Gelfand et al. (2003).

Los autores Majumdar et al. (2008) son los que por primera vez desarrollan un trabajo en Ecología, donde se hace uso de los modelos jerárquicos con estructura espacial y donde se emplea el co-kriging Bayesiano para estudiar los nutrientes y las concentraciones de carbono en el suelo y demuestran que esta metodología puede tener una amplia utilidad en otras áreas.

1.2. El proceso espacial

La variabilidad espacial está omnipresente en cualquier investigación medioambiental y en cualquier ciencia vinculada al ambiente, como la Ecología, Epidemiología, Agricultura, Toxicología, Geología, entre otras. El estudio de la variabilidad es un área relativamente nueva dentro de la Estadística. La Estadística Espacial fue brevemente presentada por Fisher en los años 30 en su investigación estadística aplicada a la Agricultura.

En Estadística Espacial el punto crucial es cómo modelizar la variabilidad espacial. La aleatorización espacial ha tenido un impacto directo sobre el desarrollo de cultivos resistentes, productivos y adaptados al tipo de suelo y a las condiciones climatológicas. La distribución aleatorizada de los tratamientos en las parcelas justifica realizar un análisis de la varianza a fin de contrastar las diferencias entre los tratamientos. Sin embargo, controlar el sesgo de esta forma implica un precio en términos de la eficiencia del análisis.

En estudios relacionados con el medioambiente, no suele ser posible realizar un riguroso diseño del experimento. La situación cambia cuando pasamos de estudiar plantas a analizar organismos y fenómenos móviles. La movilidad de los individuos está en relación con una mayor diversidad genética, de forma que, como unidades experimentales presentan una mayor heterogeneidad. De esta forma, nos enfrentamos a problemas asociados con las observaciones. A menudo una única observación constituye la información disponible.

Los problemas medioambientales vienen relacionados con observaciones espaciales de distinta naturaleza. Los datos pueden ser continuos o discretos, estar agregados espacialmente o ser observaciones individuales en puntos del espacio, sus localizaciones se encuentran dispuestas de forma regular o irregular, e incluso, estas localizaciones provienen de una región espacial continua o de un conjunto discreto.

Los datos espaciales se pueden clasificar en tres grupos fundamentales según el contexto de observación del que provienen: observaciones de un fenómeno continúo en el espacio, datos en una red fija de localizaciones y sucesos que ocurren en el espacio proporcionando un conjunto aleatorio de puntos llamado patrón puntual. Estos tipos de datos diferenciados dan origen a formas distintas de modelización y, por tanto, de análisis estadístico. La proximidad dependerá de la información contenida en el dato espacial.

Los métodos estadísticos pueden intentar salvar las dificultades creadas por la carencia de diseño experimental mediante el estudio de la variabilidad. El investigador puede aventurar cuáles son las posibles causas de esta variabilidad, pero un modelo adecuado debería describir la situación real estudiada. La presencia de la dimensión espacial en un problema exige la creación y el desarrollo de un marco estadístico que permita inferir adecuadamente sobre los procesos y sus parámetros de interés. Los datos espaciales son habitualmente dependientes entre sí y requieren modelos capaces de recoger la estructura de interrelaciones presente.

En general, los métodos estadísticos estándar asumen independencia entre las observaciones. Cuando usamos estos métodos para analizar datos espacialmente correlacionados, el error estándar de los parámetros de covarianza es subestimado y la significación estadística es sobreestimada (Cressie, 1993).

Una consideración adicional sobre el comportamiento de los modelos estadísticos espaciales es el nivel de agregación espacial. Los vecindarios se agrupan en barrios, municipios, comarcas, provincias y estados. Pero los datos pueden ser recogidos a un nivel de agregación y las covariables a otro, e incluso las decisiones políticas pueden tomarse en un tercer nivel distinto. El cambio de nivel de agregación espacial puede conducir a conclusiones completamente diferentes. No es un problema fácil de resolver y requiere un cuidado especial en cualquier fenómeno de estudio.

La modelización espacial de riesgos ha hecho uso repetidamente de distintas herramientas para conferir estructura de dependencia espacial a las observaciones objeto de modelización. Es común encontrar estudios de Disease Mapping aplicados en estudios de diversas áreas, en especial, en fenómenos de la salud. Las técnicas de Dissease Mapping o Cartografía de Enfermedades son adecuadas para realizar estudios con datos agregados. Las iniciativas en cartografía en el contexto agrícola suelen dar lugar a la construcción de modelos estadísticos para estudiar las relaciones entre los atributos de cobertura del suelo y variables como, el suelo, variables climáticas y topográficas provenientes de imágenes espectrales de satélite. Mayormente los trabajos encontrados en Agricultura están dedicados al estudio de sus recursos, es decir, al uso y tenencia de la tierra, la gestión de bosques, humedad del suelo, tipos de suelos, concentración de carbono en el suelo, nivel de producción del suelo, etc.; ilustraciones de esta perspectiva se puede encontrar en Benirschka y Binkley (1994), Bockstael (1996), Garrigues et al. (2006), Nelson y Hellerstein (1997), Bell y Bockstael (2000), Florax et al. (2003), Anselin et al. (2002), Irwin y Bockstael (2002), Kim et al. (2002) entre otros. Estos trabajos incorporan la dependencia espacial en el modelo de regresión siguiendo los principios generales de la geoestadística (Cressie, 1993) o mediante la utilización de un proceso autoregresivo espacial para el término del error (Anselin, 2001b).

Autores como Benedetti et al. (2010) han publicado recientemente un texto dedicado a métodos de investigación agraria y establecen censos y datos administrativos del uso del suelo con fines estadísticos agrarios, además cubren temas relacionados con el diseño de muestras y estimación desde el contexto frecuentista.

En literatura más reciente es posible encontrar trabajos dedicados a la Agricultura de precisión, en los que se combinan datos a muy pequeña escala obtenidos de GPS. En este contexto, los datos contienen información espacial oculta y suelen modelarse a través de técnicas Data Mining.

Desde la perspectiva Bayesiana son innumerables los trabajos que hacen uso de datos espaciales. Diggle y Ribeiro (2007), discuten el uso de modelos basados en datos geoestadísticos. Werner Hartman (2006), propone un modelo jerárquico para datos espaciales usando campos aleatorios de Markov para estudiar la composición elemental del suelo forestal; Kneib y Fahrmeir (2006), proponen modelos mixtos con estructura de regresión aditiva para datos espacio-temporales multi-categóricos para estudiar la salud de los bosques. Estos autores junto con otros investigadores publican en el año 2011 un texto que incorpora nuevos métodos para el estudio espacio-temporal en la salud de los bosques. Estos trabajos representan una muestra de la utilidad de la metodología Bayesiana referida a procesos espaciales y constituyen una muestra de la literatura más reciente relacionada con Agricultura.

1.3. Modelos jerárquicos

Son modelos probabilísticos para colecciones de variables formulados como combinaciones de diversas componentes denominadas niveles, capas o etapas. Esta estrategia es útil especialmente en la construcción de modelos complejos. Este tipo de modelización permite enlazar modelos provenientes de diversas ciencias (ambientales, médicas, sociales, biológicas, educativas y económicas, entre otras), combinando diferentes fuentes de información y empleando relaciones entre las variables estudiadas. Esta capacidad de adaptarse a situaciones complejas y gracias al desarrollo de técnicas inferenciales asequibles mediante simulación (Möller, 2003) han permitido que se conviertan en una herramienta principal en la modelización estadística de problemas epidemiológicos.

La construcción de un modelo jerárquico se hace a través de las distribuciones condicionales, con las cuales, se construye un encadenamiento de dependencias lo que ayuda a flexibilizar y potenciar la conexión entre

modelos complejos. Este tipo de metodología permite introducir no sólo factores de confusión indeterminados sino combinar fuentes de variabilidad y unir modelos parciales. La terminología para designar los elementos de un modelo jerárquico difiere según el enfoque empleado, frecuentista o Bayesiano.

Se puede aplicar el término Geostadística basada en modelos, acuñado por Diggle, Tawn y Moyeed (1998) para enmarcar la aplicación de modelos estocásticos paramétricos explícitos y métodos formales de inferencia en problemas geostadísticos. La complejidad que se deriva de estas estructuras estocásticas dificulta la inferencia en este tipo de modelos. Estas dificultades pueden resolverse planteando el problema bajo el enfoque de modelos jerárquicos espaciales.

La incorporación de asociación espacial en alguna de las capas del modelo conduce a una modelización espacial jerárquico. Dicha asociación espacial puede venir modelizada mediante un proceso espacial continuo, un campo markoviano o un proceso puntual. Por ejemplo, la inclusión de un modelo autonormal genera un *modelo Gaussiano jerárquico espacial*, que está siendo ampliamente utilizado tanto con observaciones continuas como discretas. La inferencia que se hace a partir de los *modelos jerárquicos espaciales* dependerá de la perspectiva usada y del tipo de dato espacial involucrado.

Los Modelos Lineales Generalizados (GLM) constituyen una extensión de los modelos lineales y un caso ilustrativo de los modelos jerárquicos. Los GLM comprenden aquellas distribuciones de familia exponencial uniparamétrica que recogen aditivamente los efectos fijos como una transformación monótona de la media. Esta familia permite modelar una gran variedad de situaciones, con observaciones tanto continuas como discretas. Una importante extensión de este tipo de modelos son los Modelos Lineales Mixtos Generalizados (GLMM) (Breslow y Clayton, 1993), que incorporan en el predictor lineal un conjunto de variables

latentes. Cuando estas variables provienen de un *proceso espacial* se obtiene un Modelo Espacial Lineal Generalizado (GLSM). Lee y Nelder (1996) extienden el concepto de GLMM a modelos jerárquicos lineales generalizados ampliando el uso de distribuciones no Gaussianas para variables latentes.
Los automodelos (Besag, 1974) pueden ser vistos como modelos jerárquicos, en los cuales la dependencia espacial es incorporada en forma indirecta a través de covariables ligadas a las localizaciones y que explican el proceso espacial.

1.3.1. Modelos jerárquicos Bayesianos espaciales

Desde una perspectiva Bayesiana las capas de un modelo son vistas cada una, como un proceso estocástico compuesto de observaciones, factores ocultos y parámetros a estimar. A través del *"Teorema de Bayes"* (Ecuación 1.1 y Ecuación 1.2) es posible que la información de un dato se transfiera a factores asociados a otro dato, para esto se requiere incorporar incertidumbre (estructura probabilística) tanto en las observaciones como en los parámetros de interés. El proceso de aprendizaje a través de la distribución posterior es enorme y constante, es así, que los *modelos jerárquicos Bayesianos* se convierten en una herramienta potencial para el análisis de problemas complejos.
Gracias al análisis Bayesiano es posible transferir la información de los datos a factores asociados a otro conjunto de datos a través del aprendizaje sobre los parámetros. Esta estrategia permite construir modelos jerárquicos con capas complejas que contienen observaciones, factores ocultos y parámetros del modelo. Cuando los datos son recogidos de muchas unidades que son de algún modo similares, como sujetos, animales, ciudades, etcétera, el problema estadístico es combinar la información de varias unidades para entender mejor el fenómeno en estudio. Por lo general, hay variabilidad entre las unidades y un modo natural de acercarse al

problema es construyendo un modelo en etapas *"modelo jerárquico"* y luego usarlo para hacer inferencia, si estas inferencia se hace a través de la distribución posterior entonces se emplea un *"modelo jerárquico Bayesiano"*.

El uso de modelos jerárquicos Bayesianos se ha generalizado en los tres tipos de datos espaciales. Primero se extendió gracias al modelo de Besag et al. (1991) empleado en la suavización de mapas de riesgo de enfermedad o cartografía de enfermedades en áreas pequeñas. Luego se generaliza a los datos geoestadísticos gracias a la publicación de Diggle et al. (1998). Y de forma más reciente los procesos de Cox log-gaussianos introducidos por Möller et al. (1998) para analizar patrones puntuales. Mayormente los trabajos relacionados con procesos espaciales durante los últimos años, han estado dedicados a datos agregados en unidades de área y a datos georeferenciados debido a su flexibilidad aún en el caso de problemas complejos.

Recientemente gracias a los trabajos publicados por Rue et al. (2009) y Lindgren et al. (2011) se ha abierto todo un nuevo mundo para el desarrollo de modelizaciones basadas en el paradigma Bayesiano referidas a cualquiera de los tres tipos de datos espaciales existentes.

1.4. El paradigma Bayesiano

La modelización jerárquica desde el enfoque Bayesiano esta basada en el simple hecho de tratar a la distribución conjunta, como una colección de variables aleatorias que se puede descomponer en una serie de modelos condicionales. La distribución conjunta es difícil de especificar en procesos complejos. En este caso, el producto de la serie de modelos condicionales relativamente simples conduce a una distribución conjunta que generalmente no es conocida.

Cuando se modelan procesos complejos en presencia de datos, es útil

escribir el modelo jerárquico en tres estados:

Estado 1: Modelo para los datos
Estado 2: Modelo para el proceso|parámetros del proceso
Estado 3: Modelo para los parámetros

La idea básica del enfoque Bayesiano, es resolver un problema dividiéndolo en sub-problemas más simples. Cada uno de estos estados pueden a su vez dividirse en muchos sub-estados. Los métodos Bayesianos permiten hacer estimación de forma natural en la modelización jerárquica. Si la estimación con modelos jerárquicos se hace con métodos no Bayesianos, esto llevaría a asumir condiciones adicionales que dificultan la inferencia. Bajo el enfoque Bayesiano, la distribución posterior es obtenida usando el Teorema de Bayes que más adelante será enunciado. El Teorema de Bayes ofrece el mecanismo a través del cual puede accederse a la distribución posterior. Aunque parezca simple en principio, la aplicación del Teorema de Bayes en modelos complejos puede ser un reto. La especificación de las distribuciones a priori a los parámetros involucrados en el modelo constituye un desafío enorme. A pesar de haber sido durante mucho tiempo un tema de discusión en la comunidad estadística, la especificación subjetiva de las previas dependerá del conocimiento científico que se tenga del fenómeno. De hecho, poder incorporar este conocimiento en el modelo hace posible considerar fuentes de incertidumbre adicionales. Debido a la complejidad y alta dimensión natural de los modelos jerárquicos con estructura espacial, a lo largo del trabajo, presentamos dos enfoques distintos para realizar inferencia Bayesiana, uno basado en algoritmos sustentados en los MCMC y el otro, un enfoque determinístico basado en la aproximación de Laplace denominado INLA.

1.4.1. Inferencia

Bajo la perspectiva Bayesiana, la incertidumbre o falta de información sobre el parámetro θ puede ser incorporada a través de distribuciones previas, considerando este parámetro como una variable aleatoria. Sea $\pi(\theta|\lambda)$ la distribución previa, donde λ es un vector de hiperparámetros. Si λ es conocida, la inferencia sobre θ se hace a partir de la distribución posterior $p(\theta|y,\lambda)$, que se obtiene gracias a *"Teorema de Bayes"* que combina la previa y la verosimilitud.

$$p(\theta|y,\lambda) = \frac{p(y,\theta)|\lambda}{p(y|\lambda)} = \frac{p(y,\theta)|\lambda}{\int p(y,\theta|\lambda)d\theta} = \frac{f(y|\theta)\pi(\theta|\lambda)}{\int f(y|\theta)\pi(\theta|\lambda)d\theta} \tag{1.1}$$

En la práctica, λ no es conocido y por tanto, es necesario definir un segundo estado para los hiperparámetros (distribución para $p(\lambda)$), quedando (1.1) como:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{\int f(y|\theta)\pi(\theta|\lambda)h(\lambda)d\lambda}{\int f(y|\theta)\pi(\theta|\lambda)h(\lambda)d\theta d\lambda} \tag{1.2}$$

Alternativamente, se puede reemplazar λ por el estimador $\hat{\lambda}$ obtenido al maximizar la distribución marginal $p(y|\lambda) = \int f(y|\theta)\pi(\theta|\lambda)d\theta$, visto como una función de λ. La inferencia puede estar basada en el estimador de la distribución posterior $p(\theta|y,\hat{\lambda})$ al reemplazar $\hat{\lambda}$ en la ecuación (1.1). Este enfoque es conocido como Análisis Empírico Bayes (Berger, 1985), Maritz y Lwin (1989), Carlin y Louis (2000) para más detalles de esta metodología.

- *Regla de Bayes:* Se debe comenzar con un modelo que provea una distribución conjunta para θ y y. La función de densidad conjunta es escrita como un producto de dos densidades que son referidas frecuentemente como la distribución a priori $p(\theta|\lambda)$ y la verosimilitud $p(y|\theta)$ respectivamente:

$$p(\theta,y) = p(\theta)p(y|\theta) \tag{1.3}$$

Al condicionar en el valor conocido de los datos y, usando la *regla de Bayes* se obtiene la distribución posterior:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{1.4}$$

donde, $p(y) = \sum_\theta p(\theta)p(y|\theta)$, en el caso discreto o $p(y) = \int p(\theta)p(y|\theta)d\theta$ en el caso continuo. Una forma equivalente de (1.4) omite el factor de $p(y)$, el cual no depende de θ y con y fijo puede ser considerado como una constante. El lado derecho de (1.4) puede escribirse como:

$$p(\theta|y) \propto p(\theta)p(y|\theta) \tag{1.5}$$

En el enfoque Bayesiano toda la inferencia se hace a partir de la distribución posterior $p(\theta|y)$.

1.4.2. Métodos Markov chain Monte Carlo (MCMC)

Las técnicas MCMC (generación de cadenas de Markov para usarlas en estudios Monte Carlo) están diseñadas con la finalidad de poder estudiar empíricamente las características de distribuciones complejas. Básicamente consisten en lo siguiente: Dada una distribución $p(\theta)$ completamente conocida, salvo quizás por su constante de proporcionalidad, se trata de generar una o varias realizaciones de una cadena de Markov cuya distribución estacionaria sea $p(\theta)$. Esas realizaciones se utilizarán posteriormente para obtener, por Monte Carlo, aproximaciones a todas las características de $p(\theta)$ que se desean conocer.

Desde la perspectiva de las técnicas MCMC, el estudio de las cadenas de Markov es completamente distinto. Se parte suponiendo que la distribución que deseamos es invariante, $\pi(\theta)$, y se pretende construir un núcleo de transición que produzca una cadena de Markov con distribución

estacionaria $\pi(\theta)$, para la que las medias ergódicas sean estimadores consistentes y podamos aplicar alguna versión del Teorema del Límite.
La primera utilización documentada de estas técnicas es Metropolis et al. (1953), donde se introduce un método que posteriormente es generalizado por Hastings (1970), y que ahora se denomina *algoritmo de Metropolis-Hastings*. Sin embargo, esos trabajos pioneros pasan prácticamente desapercibidos en la literatura estadística durante mucho tiempo. Más recientemente se introduce el *algoritmo de Gibbs* (Geman y Geman, 1984) y el algoritmo de Data Augmentation (Tanner y Wong, 1987), pero es tras la publicación de Gelfand y Smith (1990) cuando este tipo de métodos se convierten en una herramienta indispensable en la aplicación del paradigma Bayesiano.
Una referencia obligada en el estudio de estas técnicas es la monografía editada por Gilks, Richardson y Spiegelhalter (1996). También es de destacar el texto de Gamerman (1997), el artículo de Brooks (1998) y la monografía de Robert y Casella (1999). A continuación se detallan los algoritmos Gibbs sampling y Metropolis-Hastings, mayormente empleados en la inferencia Bayesiana.

Muestreador Gibbs

Entre las técnicas MCMC el algoritmo de Gibbs es uno de los métodos más fáciles de aplicar y, sin duda por ello, el más conocido y utilizado. El artículo de Casella y George (1992) constituye una introducción clara y concisa de este método, y en Gelfand et al. (2003) se presentan aplicaciones del mismo.
El muestreador Gibbs aproxima integrales que no pueden ser calculadas en forma cerrada generando cadenas de Markov Monte Carlo (MCMC), donde la transición del origen de la distribución $\pi(\theta)$ esta formada por las distribuciones condicionales completas $(\pi(\theta_i) = \pi_i(\theta_i|\theta_{-i}))$. Se asume que la distribución de interés es $\pi(\theta)$, donde el vector θ puede descomponerse

en $k \geq 2$ subvectores, $\theta = (\theta_1, \ldots, \theta_k)$. Cada uno de los componentes θ_i de θ puede ser un escalar, un vector o una matriz. Se considera que las distribuciones condicionales completas $\pi_1(\theta_i) = \pi_1(\theta_i|\theta_{-i})$ están disponibles, pudiendo generar valores de las mismas sin excesivo coste computacional, siendo θ_i el vector $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_k)$.
Estas condiciones se presentan habitualmente en el estudio de *modelos jerárquicos* con iniciales conjugadas y posiblemente, datos faltantes o incompletos. En estos casos, los datos aumentados pueden considerarse incluidos en el vector θ, conjuntamente con los parámetros e hiperparámetros del modelo.
El objetivo del algoritmo de Gibbs es obtener una muestra suficientemente grande de la distribución posterior. A partir de ella se podrá hacer inferencias sobre los momentos, las marginales, la distribución predictiva, o cualquier otra característica de la distribución posterior que sea de interés. El problema que se debe resolver es cómo tomar una muestra de la distribución π, cuando los planes para la generación de las muestras son costosos, complicados o simplemente no se conoce el origen de la distribución π, pero es posible generar muestras de las distribuciones $\pi_i(\theta_i)$.
El algoritmo Gibbs puede ser descrito de la manera siguiente:

1. Se inicializa el contador de la iteración de la cadena en $j = 1$ y se asignan valores iniciales para el vector $\theta^{(0)} \leftarrow (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$

2. Repetir hasta convergencia {
   Se obtiene un nuevo valor $\theta^{(j)} = (\theta_i^{(j)}, \ldots, \theta_k^{(j)})'$ a partir de $\theta^{(j-1)}$ por la sucesiva generación de los valores:

$$\theta_1^{(j)} \sim \pi_1(\theta_1|\theta_2^{(j-1)}, \ldots, \theta_k^{(j-1)})$$
$$\theta_2^{(j)} \sim \pi_k(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}, \ldots, \theta_k^{(j-1)})$$
$$\vdots$$

$$\theta_k^{(j)} \sim \pi_k(\theta_k|\theta_1^{(j)}, \ldots, \theta_{k-1}^{(j)})$$

3. Se actualiza el contador de j a $j+1$ y se regresa al paso 2.}

Cuando la cadena converge, los valores resultantes de $\theta^{(j)}$ son una muestra de la distribución π. Se asume la convergencia de la cadena si la cadena se aproxima a una condición de equilibrio cuando el número de iteraciones se incrementa (Gamerman, 1997).

Las condiciones de convergencia para el muestreador Gibbs fueron establecidas por Robert y Smith (1994). Los resultados son presentados en términos de espacios paramétricos, pero pueden extenderse y combinar los parámetros continuos y discretos.

Un valor de la distribución de interés π es obtenido solamente cuando el número de iteraciones de la cadena se aproxima a infinito. En la práctica esto no es posible, y un valor obtenido en una iteración suficientemente grande es tomado como una muestra de la distribución π. La dificultad es la determinación de cuán grande debería ser el número de iteraciones. No hay respuesta simple a esta pregunta y los mayores esfuerzos se han orientado al estudio de las características de la convergencia de cadenas (Gelfand y Smith, 1990).

La muestra obtenida de θ (con $j \to \infty$) es una muestra correlacionada de la distribución posterior de la cual, se puede obtener cualquier cantidad de interés. Usando la estimación Monte Carlo es posible encontrar

$$\hat{E}(\theta_i|y) = \frac{1}{K - j_0} \sum_{j=j_0+1}^{K} \theta_i^{(j)}$$

La iteración de $j = 0$ a $j = j_0$ se conoce como período de inicialización (burn-in). En la práctica, se pueden simular paralelamente m cadenas, en este caso, el estimador posterior de la media sería

$$\hat{E}(\theta_i|y) = m\frac{1}{K - j_0} \sum_{i=1}^{m} \sum_{j=j_0+1}^{K} \theta_{i,j}^{(j)}$$

Muestreador Metropolis-Hastings

Este algoritmo consiste en generar valores de una cadena de Markov cuya distribución estacionaria, distribución marginal de la cadena de Markov, sea la distribución objetivo $\pi(\theta)$. El algoritmo de Metropolis-Hastings proporciona un método sencillo para construir innumerables cadenas de Markov con esa propiedad, lo que nos permitirá buscar entre ellas una que además posea otras propiedades complementarias: rapidez de convergencia a estacionariedad y no demasiada autocorrelación.
Básicamente se trata de utilizar una cadena de Markov auxiliar, para la que disponemos de un generador eficiente de su núcleo de transición $Q(\theta, A)$ (que representa la distribución de probabilidades de pasar en una etapa del punto θ a la región A), y añadirle en cada etapa un mecanismo de aceptación-rechazo con probabilidad de aceptación dada por:

$$\alpha(\theta, \phi) = min\{\frac{q(\phi, \theta)\pi(\phi)}{q(\theta, \phi)\pi(\theta)}, 1\}$$

de manera que si en la etapa i el valor obtenido es $\theta^{(i)}$, en la etapa siguiente se genera ϕ a partir de $q(\theta, \phi)$, que es la función de probabilidad (o densidad en su caso) asociada al núcleo $Q(\theta, A)$. Entonces $\theta^{(i+1)} = \phi$ con probabilidad $\alpha(\theta, \phi)$, o $\theta^{(i+1)} = \theta^{(i)}$ en otro caso.
La mecánica del algoritmo es la siguiente:

1. Se inicializa el contador de la iteración de la cadena en $i = 0$ y se asignan valores iniciales para el vector $\theta^{(0)} \leftarrow (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$

2. Repetir

$$\{\phi \leftarrow\sim q(\theta^{(i)}, \phi)$$

$$r \leftarrow \frac{q(\phi, \theta^{(i)})\pi(\phi)}{q(\theta^{(i)}, \phi)\pi(\theta^{(i)})}$$

$$u \leftarrow\sim Un(0, 1)$$

$$si\ u \leq r\ entonces\ \phi \leftarrow \theta^{(i)}$$

$$\theta^{(i+1)} \leftarrow \phi$$

$$i \leftarrow i + 1\}$$

A la familia de distribuciones $q(\phi, \theta)$ se le conoce como *distribución instrumental*. Un caso particular es cuando el núcleo de transición de probabilidades instrumental es simétrico, $q(\theta, \phi) = q(\phi, \theta)$, entonces se obtiene el *algoritmo de Metropolis*.
Como la distribución objetivo $\pi(\theta)$ sólo aparece en el algoritmo a través de un cociente, no es necesario conocer su constante de integración. Por ello, el algoritmo de Metropolis-Hastings puede ser una herramienta muy útil en la aplicación de las técnicas Bayesianas, donde la constante de integración de la distribución posterior no suele ser conocida.

Cómo llevar a la práctica las técnicas MCMC

Son muchas las preguntas que se plantean en el estudio sobre las características de la convergencia de las cadenas. Aunque todas ellas están bastante relacionadas, se comentarán por separado.

- *Cuántas cadenas generar*: En el artículo de Gelfand y Smith (1990) se propone utilizar solamente el último valor observado de la cadena, generando tantas cadenas como tamaño deseemos que tenga la muestra de la distribución objetivo. De esa forma se consigue una muestra aleatoria cuyo análisis estadístico es muy sencillo.

  Alternativamente, si se utiliza sólo una cadena habrá que descartar las primeras etapas hasta conseguir equilibrio (distribución estacionaria), pero cuando está se haya alcanzado todas las nuevas etapas tienen como distribución marginal la distribución objetivo. La independencia en la muestra obtenida se pierde, por lo tanto, la muestra ya no es aleatoria con lo que se complica ligeramente el análisis estadístico, pero se gana enormemente en eficiencia. Además,

los resultados sobre invariabilidad son asintóticos, luego cuanto más avancemos en la cadena debemos esperar resultados más fiables. Sin embargo, posiciones intermedias entre estos dos extremos pueden ser mucho más interesantes. Así, Gelman y Rubin (1992) propusieron la utilización de pocas cadenas, cada una de ellas empezadas en puntos muy distantes del espacio paramétrico, pero utilizando muchas etapas de cada cadena. Esta estrategia permite realizar un diagnóstico de la convergencia, comparando los resultados obtenidos con las distintas cadenas. En efecto, cualquier cadena puede quedarse atrapada en una moda de la distribución objetivo, dando la sensación de una buena convergencia que, sin embargo, no existe; así si fuese la única cadena generada, las conclusiones del estudio serían erróneas.

- *De qué punto partir*: En teoría, los resultados de un algoritmo MCMC son independientes del valor inicial del algoritmo, pero en la práctica, la elección del valor inicial va a influir en la rapidez con la que se alcance el equilibrio. Lo aconsejable es partir de valores iniciales que estén en zonas de alta probabilidad con respecto a la distribución objetivo: partir del estimador máximo verosímil, por ejemplo. Si se utilizan varias cadenas, es aconsejable que los valores iniciales estén bastante repartidos por el espacio paramétrico, para evitar que todas las cadenas se queden atrapadas en un mismo máximo local. Pueden utilizarse métodos de optimización para localizar la moda, o modas de la distribución objetivo. Gelman y Rubin (1992) propusieron localizar las regiones de alta densidad con respecto a la función objetivo, construir una mixtura de distribuciones t-Student cada una de ellas centrada en una de las regiones localizadas, y utilizar esa mixtura para simular los valores iniciales de las distintas cadenas.

- *Número de iteraciones hasta convergencia:* Posiblemente esta es

la cuestión más importante, y más difícil de resolver, de todas las planteadas. Las primeras etapas del algoritmo todavía estarán influenciadas por el punto inicial, por lo que su uso introduciría un sesgo en los resultados. Esas primeras etapas, hasta que el algoritmo alcanza equilibrio, se conocen como inicialización (burn-in) y deben ser desechadas. Pero, ¿cómo saber que el algoritmo ya ha convergido y está en equilibrio? ¿cuántas etapas debemos desechar?. Para responder a estas interrogantes, se mencionan brevemente algunos métodos de diagnóstico.

Gelfand y Smith (1990) sugirieron observar las gráficas de cuantiles y de autocorrelaciones. Las propias trazas de las series univariantes pueden ser de utilidad, una vez alcanzada la convergencia las trazas de los diversos parámetros deben estabilizarse alrededor de algún valor, sin mostrar ninguna tendencia definida.

Geweke (1992) propone comparar, en cada serie univariante, la media del primer tramo con la del último, utilizando estimadores espectrales de la varianza. Si se descubren diferencias significativas el primer tramo es descartado, en otro caso se considera que toda la serie ya está en equilibrio. Algo parecido proponen Heidelberger y Welch (1983), pero utilizan el test no paramétrico de Cramer von Mises.

Gelman y Rubin (1992) propusieron un test basado en el análisis de varianza, comparando varias cadenas.

Raftery y Lewis (1992) propusieron un método basado en los resultados teóricos sobre convergencia en las cadenas de Markov de dos estados. Para ello sugieren fijar algún cuantil, habitualmente de orden 0.025 ó 0.975 y sustituir la cadena por otra formada por ceros y unos, que ya será una cadena de Markov de dos estados.

- *Cuántas iteraciones después de la convergencia:* Una vez alcanzado el equilibrio, el número de etapas a utilizar es un problema de tamaño

muestral, que puede resolverse de la forma habitual. Así, debemos establecer una cota sobre el error típico de los estimadores de los parámetros que se consideren más relevantes y, a partir de ella, obtendremos el tamaño muestral.

Si la muestra es aleatoria y se estima una característica univariante de la distribución objetivo mediante su media muestral, $\bar{x}$, el error típico del estimador es $s/\sqrt{n}$, siendo n el tamaño de la muestra y s su desviación típica muestral. Alternativamente, si la serie puede aproximarse por un proceso autoregresivo de primer orden, el error típico de la media ergódica es:

$$\frac{s}{\sqrt{n}}\sqrt{\frac{1+r}{1-r}}$$

siendo r un estimador de la autocorrelación de la serie. De esta forma, también resulta sencillo trabajar con muestras no independientes. A $\sqrt{(1+r)/(1-r)}$ se le conoce como *factor de inflación* y puede ser inferior a uno si la autocorrelación es negativa, lo que es muy difícil que ocurra en un algoritmo MCMC.

Acotando el error típico por una cantidad dada $\epsilon > 0$, el tamaño muestral debe ser:

$$n = \text{Parte entera}(\frac{s^2}{\epsilon^2}\frac{1+r}{1-r})$$

- *Cómo saber si la cadena se esta mezclando adecuadamente:* La cadena no sólo debe recorrer todo el soporte de la distribución objetivo, debe hacerlo con rapidez. En otro caso se dice que la cadena no se está *mezclando* bien, y se necesitaría un enorme número de etapas para poder extraer resultados fiables.

  Para comprobar si la cadena se esta mezclando bien resultan muy útiles las gráficas de las trazas univariantes. Tendencias cíclicas en esas trazas indican que la cadena no se está mezclando bien. Otra

herramienta diagnóstica interesante la constituyen los *factores de inflación*, pues si las autocorrelaciones son muy grandes la cadena tardará mucho en poder recorrer todo el soporte de la distribución objetivo.

La inclusión de parámetros de sintonización en el método MCMC utilizado permite cambiar de núcleo sin realizar cambios en el código programado. Así, durante las primeras etapas se debe probar con diversos valores de los parámetros de sintonización, hasta encontrar un valor para el que la cadena converja y se mezcle bien. De no conseguir resultados satisfactorios debe cambiarse el método MCMC, o buscar alguna reparametrización.

Aunque no existen resultados teóricos convincentes, suelen conseguirse mejores resultados cuando las correlaciones entre los parámetros son pequeñas.

- *Se utilizan todas las etapas o se adelgaza la salida:* Al utilizar una única cadena muy larga o varias cadenas no tan largas, la muestra obtenida no es independiente, lo que dificulta su análisis estadístico. Por ello se ha propuesto *adelgazar* la salida utilizando tan sólo una de cada k etapas, siendo k un natural no demasiado grande, de manera que las etapas usadas sean *aproximadamente independientes*. Pero al desechar etapas se está perdiendo información y no se gana gran cosa, pues la estimación utilizando medias ergódicas no es complicada.

  Existe, sin embargo, otra razón que justifica en ocasiones el adelgazamiento de la salida: restricciones en la capacidad de memoria para el almacenamiento de la salida. Teniendo en cuenta que habitualmente se necesitan cientos de miles de etapas, la memoria necesaria para el almacenamiento de la salida puede constituir un gran problema que hay que considerar. Si tenemos una restricción acotando el tamaño máximo de la salida, como es mucho más

informativa una muestra independiente que una muestra relacionada, será conveniente que la adelgacemos. También puede ser útil truncar los valores generados antes de almacenarlos, para utilizar menos cifras decimales.

1.4.3. Inferencia Bayesiana con métodos MCMC

La inferencia en modelos jerárquicos puede realizarse mediante máxima verosimilitud, pero con frecuencia dicha verosimilitud no es totalmente conocida. La metodología Bayesiana, ha extendido el uso de los *modelos jerárquicos* gracias a que la distribución posterior puede ser muestreada por métodos MCMC. No obstante, diversas dificultades prácticas deben tenerse en cuenta para llegar a conclusiones adecuadas.

La complejidad de las estructuras estocásticas que se derivan a partir de la formulación de un *modelo jerárquico Bayesiano* dificulta su inferencia. Esta dificultad se debe a la variedad de posibilidades para la especificación de la distribución previa y a la dificultad de resumir la distribución posterior resultante. Sin embargo, en la actualidad están disponibles algoritmos y algunas herramientas informáticas que permiten realizar tal inferencia a pesar de su complejidad.

Esta inferencia supone un desafío computacional ya que en problemas reales, las integrales requeridas para hacer las estimaciones son generalmente intratables al no tener una forma analítica cerrada. Este obstáculo numérico se resuelve usando métodos de integración MCMC, algunos de los más usados se presentaron en la sección (1.4.2) y empleando herramientas informáticas.

Una herramienta computacional que permite el desarrollo de inferencia Bayesiana usando Muestreo Gibbs es el BUGS (Spiegelhalter et al. 2007), este software tiene dos versiones, el WinBUGS y el OpenBUGS. El OpenBUGS (Lunn et al. 2009a) representa la versión abierta del proyecto BUGS. Mientras que el WinBUGS es una versión estable que se encuentra

disponible, pero no en desarrollo. Las últimas versiones del OpenBUGS se han diseñado para ser al menos tan eficaces y fiables como las del WinBUGS.

El WinBUGS (Bayesian inference Using Gibbs Sampling for Windows, Spiegelhalter et al. 2003) es un sistema capaz de especificar una variedad de distribuciones previas para muchos modelos dados y de muestrear las condicionales completas resultantes. Este sistema consiste en un conjunto de funciones que permiten la especificación de modelos y las distribuciones de probabilidad para todas sus componentes aleatorias tanto para las observaciones como para los parámetros. La especificación de los modelos es sorprendentemente simple dada la complejidad de estos modelos.

Para cada combinación de datos y modelos, WinBUGS genera muestras de los parámetros de modelo para cada iteración $k \geq 1$ después de m iteraciones. Los valores de k y m, así como los parámetros muestreados para ser almacenados, son escogidos por el usuario. Además, el programa provee los estimados basados en muestras de la media posterior y el intervalo de credibilidad para los parámetros. Este sistema emplea para su entrada y salida la sintaxis del lenguaje S desarrollado en Bell Laboratories (AT & T) a finales de los 70 y principios de los 80 por Richard Becker, Jhon Chambers y Allan Willks. Este lenguaje inicialmente fue diseñado para análisis exploratorio de datos y la mayor parte de la funcionalidad estadística fue agregada posteriormente.

El OpenBUGS es un software para el análisis Bayesiano de modelos complejos utilizando los métodos MCMC. Es la variante de código abierto del WinBUGS. Una diferencia fundamental entre ambos es la forma en que el sistema experto selecciona el algoritmo de actualización a utilizar para cada clase de distribución condicional completa en cada nodo. Mientras WinBUGS define un algoritmo para cada clase posible, no hay límites en el número de algoritmos que OpenBUGS puede utilizar, lo que permite una mayor flexibilidad y extensibilidad. En OpenBUGS el usuario puede

seleccionar el programa de actualización que se utilizará para cada nodo después de la compilación.

El resultado de toda técnica MCMC es una realización finita de una cadena de Markov multivariante, por lo que se resume en una matriz de datos. La aplicación informática Convergence Diagnostic and Output Analysis (CODA) (Plummer et al. 2006), está especialmente diseñada para el análisis de esa matriz y puede obtenerse desde la página Web de BUGS, junto con su manual. También existe una versión que funciona como un módulo de R.

CODA utiliza todo el resultado de una técnica MCMC, tanto si se ha obtenido sólo una cadena o varias cadenas en paralelo, para construir un objeto del tipo MCMC que es sobre el que trabaja: es el *input* de la aplicación.

Dentro de CODA, ese objeto MCMC puede manipularse con facilidad, incluyendo un adelgazamiento de la salida, puede resumirse mediante diversos estadísticos descriptivos y gráficas, y pueden realizarse varios diagnósticos de convergencia.

Entre los estadísticos descriptivos proporciona las medias, desviaciones típicas y cuantiles de la distribución empírica. Los errores típicos de la media los calcula suponiendo independencia y mediante métodos de series temporales, para incorporar las correlaciones dentro de las cadenas. Entre las gráficas destacan las trazas de cada serie, y la estimación de la densidad de cada variable. También obtiene las gráficas de autocorrelación y de correlaciones cruzadas entre variables.

Los diagnósticos de convergencia de los cuales dispone, son los de Gelman y Rubin (1992) y Geweke (1992), ambos con sus gráficas asociadas, y los de Heidelberger y Welch (1983) y Raftery y Lewis (1992).

Cuando el WinBUGS no permite trabajar con ciertos tipos de modelos complejos, es necesario elaborar un código de programación específico y, habitualmente, ir modificándolo ligeramente (mediante la utilización de

parámetros de sintonización o buscando reparametrizaciones adecuadas) hasta conseguir que funcione adecuadamente. La elaboración de estos códigos en R suele ser relativamente fácil y cómoda, pero su ejecución puede ser excesivamente lenta para el volumen de operaciones a realizar y pueden presentarse problemas de memoria.

R es un "entorno", es decir, un sistema completamente diseñado y coherente, y no una agrupación incremental de herramientas muy específicas e inflexibles. Posee un lenguaje que fue implementado en base al lenguaje S por Ross Ihaka y Robert Gentleman (University of Auckland, Nueva Zelanda). A partir de 1995 comienza a ser distribuido gratuitamente bajo los términos de la licencia de GNU (Free Software Foundation) y desde entonces, el desarrollo de R ha sido un esfuerzo de colaboración internacional, con trabajo aportado por voluntarios. Desde 1997 la coordinación del desarrollo de R está a cargo de un "Core Team" compuesto por miembros de todas partes del mundo.

R puede ser extendido por medio de programas escritos por el usuario o mediante "bibliotecas" (packages) que pueden ser obtenidos vía Internet en forma gratuita. En R pueden incluirse paquetes como CODA (diagnóstico de convergencia), BRugs (Gelman 2003; Banerjee 2007), spBayes (análisis de modelos espaciales jerárquicos para datos geostadísticos), glmmBUGS, GLMMGibbs, R2WinBUGS, entre muchos otros.

Tanto el WinBUGS como el OpenBUGS pueden ser ejecutados desde R usando los paquetes BRugs o R2WinBUGS, respectivamente. En especial para este trabajo se usará la librería R2WinBUGS y el software OpenBUGS.

1.4.4. Criterios para selección de modelos

La comparación de modelos es requerida en muchas áreas, incluyendo la selección de variables en modelos de regresión, la determinación del número de componentes en un modelo mixto o en la selección de familias

paramétricas. Igual que ocurre en el enfoque frecuentista, la comparación de modelos desde la perspectiva Bayesiana no nos dirá cuál es el modelo verdadero, pero nos acercará al mejor a la luz de los datos y de otras informaciones. En esta sección se han recogido algunos de los métodos más empleados desde el enfoque Bayesiano para comparar modelos, ingrediente indispensable y necesario en la inferencia. En este trabajo se usará sólo el criterio DIC para comparar modelos formulados desde la perspectiva Bayesiana.

Test de Hipótesis

La aproximación bayesiana a las pruebas de hipótesis está basada en el cálculo de la probabilidad condicional de una hipótesis H_0 dada la información disponible, digamos I_0, esto es, $p(H|I_0)$. Cuando la hipótesis nula es $H_0 : \theta \in \Theta_0$ y la alternativa $H_1 : \theta \in \Theta_1$, con $\Theta_0 \bigcap \Theta_1 = \emptyset$, son formuladas, hay creencias a priori sobre ambas, digamos $\xi(H_0|I_0) + \xi(H_1|I_0) = 1$. Por el teorema de la probabilidad total, la distribución a priori de θ es:

$$\xi(\theta|I_0) = \xi(\theta|H_0, I_0)\xi(H_0|I_0) + \xi(\theta|H_1, I_0)\xi(H_1|I_0)$$

donde $\xi(\theta|H_i, I_0)$, son las densidades a priori de θ, condicionadas en cada hipótesis.
La información muestral es usada para calcular los odds a priori:

$$\frac{\xi(H_0|I_0)}{\xi(H_1|I_0)}$$

los odds posteriores en favor de H_0:

$$\frac{\xi(H_0|I_0)}{\xi(H_1|I_0)} = \frac{p(y|H_0)}{p(y|H_1)} \frac{\xi(H_0|I_0)}{\xi(H_1|I_0)}$$

de la cual se deriva la siguiente regla de decisión:
Si $\xi(H_0|I_0) < \xi(H_1|I_0)$ se rechaza H_0
Si $\xi(H_0|I_0) > \xi(H_1|I_0)$ se acepta H_0
Si $\xi(H_0|I_0) = \xi(H_1|I_0)$ indecisión acerca de H_0

Factor de Bayes

A la razón $\frac{p(y|H_0)}{p(y|H_1)}$ se le conoce como *factor de Bayes*, denotado por **BF** o $B_{01}(y)$. Si se quiere probar:

$$H_0 : \theta \in \Theta_0 \; vs. \; H_1 : \theta \in \Theta_1$$

Sea $f(x|\theta)$ la verosimilitud de x dado θ. Tenemos las siguientes formas del factor de Bayes:

1. $B_{01}(x) = \frac{f(x|\theta_0)}{f(x|\theta_1)}$ (Prueba simple vs. simple)
2. $B_{01}(x) = \frac{f(x|\theta_0)}{\int_{\Theta_1} f(x|\theta)\xi_1(\theta)d\theta}$ (Prueba simple vs. compuesta)
3. $B_{01}(x) = \frac{\int_{\Theta_0} f(x|\theta_0)\xi_0(\theta)d\theta}{\int_{\Theta_1} f(x|\theta)\xi_1(\theta)d\theta}$ (Prueba compuesta vs. compuesta)

Jeffreys presenta los siguientes criterios sobre el Factor Bayes (BF) para decidir cuándo optar por H_0:

Factor Bayes (BF)	Decisión
$1 < BF$	Hipótesis Nula se sostiene
$10^{-\frac{1}{2}} < BF < 1$	Evidencia contra H_0, apenas para mencionar
$10^{-1} < BF < 10^{-\frac{1}{2}}$	Evidencia sustancial contra H_0
$10^{-\frac{3}{2}} < BF < 10^{-1}$	Evidencia fuerte contra H_0
$10^{-2} < BF < 10^{-\frac{3}{2}}$	Evidencia muy fuerte contra H_0
$BF < 10^{-2}$	Evidencia decisiva contra H_0

Tabla 1.1: Criterios de Jeffreys sobre el BF para decidir sobre H_0

Cuando las probabilidades a priori son iguales, el factor de Bayes determina la regla de decisión. La evaluación del factor de Bayes involucra el cálculo de

$$p(y|H_0) = \int p(y|H_0, \theta)\xi(\theta|H_0, I_0)d\theta$$

$$p(y|H_1) = \int p(y|H_1, \theta)\xi(\theta|H_1, I_0)d\theta$$

El Factor Bayes proporciona una indicación de cuánto cambian nuestras razones de probabilidad de una situación sin datos, a la luz de los datos, para favorecer un modelo. Puede verse como una medida de la evidencia proporcionada por los datos en favor de un modelo comparado con un competidor. El logaritmo del Factor Bayes ha sido llamado *el peso de la evidencia* proporcionada por los datos (De Santis y Spezzaferri, 1999).
El Factor Bayes puede verse como la versión bayesiana de la prueba clásica de la razón de verosimilitudes (De Santis y Spezzaferri, 1999). Si se asumen dos hipótesis simples, digamos θ_1 y θ_2, el Factor Bayes se reduce a la razón de verosimilitud $f(y|\theta_1)/f(y|\theta_2)$.
Por muchos años el *Factor Bayes* fue considerado apropiado para comparar modelos, pero sólo es posible usarlo con previas propias y para modelos de baja dimensión. Por lo tanto, cuando los M_i modelos son complejos (*modelos jerárquicos*) y en alguno de sus niveles existen previas impropias, esta metodología no puede utilizarse, ya que si $\pi_i(\theta_i)$ es impropia, entonces $p(y|M_i) = \int f(y|\theta_i, M_i)\pi_i(\theta_i)d\theta_i$ también lo será y BF no estará definido.

Los criterios BIC y AIC

Como el Factor Bayes es a menudo difícil o imposible de calcular, sobre todo en modelos con muchos parámetros o efectos arbitrarios o previas impropias, una aproximación al factor de Bayes es el BIC (Criterio de Información Bayesiana). El BIC es un método "de acceso rápido" muy popular, también conocido como el Criterio de Schwarz, permite conocer el cambio entre dos modelos que se comparan a partir de:

$$\triangle BIC = W - (p_2 - p_1)\log \text{ n},$$

donde p_i es el número de parámetros en el modelo M_i, $i = 1, 2$ y

$$W = -2 \log [\frac{\sup_{M_1} f(y|\theta)}{\sup_{M_2} f(y|\theta)}]$$

es el usual radio de verosimilitudes. Schwarz (1978) demuestra que para *modelos no-jerárquicos* (con dos estados) y tamaños de muestra n grandes, el *BIC* se aproxima a $-2\ log\ BF$. Una alternativa al *BIC* es el *criterio de Información Akaike*(AIC), cuya expresión es

$$\triangle AIC = W - 2(p_2 - p_1)$$

Tanto el BIC como el AIC son criterios de bondad de ajuste basados en el radio de la verosimilitud. En ambos, el segundo término representa una penalización corregida por la diferencia entre el número de parámetros de los modelos comparados (se piensa en M_2 como el modelo "saturado" y en M_1 como el modelo "reducido").

Comparación Múltiple de Modelos

De Santis y Spezzaferri (1999) piensan en términos de modelos, digamos $M_1, \ldots, M_s$, donde se asume que M_i está parametrizado por $\theta_i \in \Theta_i$, de dimensión d_i, y con función de densidad de probabilidad de los datos $f_i(y|\theta_i)$ y distribución a priori $\xi(\theta_i)$. Si se tienen las probabilidades a priori para los modelos $p_1, \ldots, p_s$, por el *Teorema de Bayes* se tiene:

$$Pr(M_i|y) = \frac{p_i m_i(y)}{\sum_{j=1}^{s} p_j m_j(y)}$$

donde

$$m_i(y) = \int_{\Theta_i} f_i(y|\theta_i)\xi(\theta_i)d\theta_i,\ i = 1, \ldots, s$$

es la distribución marginal de los datos bajo el modelo M_i. La razón de las probabilidades posteriores nos permite hacer una comparación entre modelos. Para los modelos M_j y M_k se tiene:

$$\frac{Pr(M_j|y)}{Pr(M_k|y)} = \frac{p_j}{p_k} B_{jk}(y),$$

donde

$$B_{jk}(y) = \frac{m_j(y)}{m_k(y)}$$

es el Factor Bayes para el modelo M_j contra el modelo M_k a partir de los datos y.

El Criterio de Información de Deviance (DIC)

Spiegelhalter et al. (2002) propone una generalización del criterio AIC, ya que este último no es apropiado para *modelos jerárquicos* de 3 o más niveles. Esta generalización esta basada en la distribución posterior de la *deviance*,

$$D(\theta) = -2log\ f(y|\theta) + 2log\ h(y) \tag{1.6}$$

donde $f(y|\theta)$ es la función de verosimilitud y $h(y)$ es una función estandarizada de los datos. Este autor sugiere resumir la bondad de ajuste del modelo por la esperanza posterior de la deviance

$$\bar{D} = E_{\theta|y}[D]$$

Por tanto, el *Criterio de Información de Deviance* (DIC) se define como:

$$DIC = \bar{D} + p_D = 2\bar{D} - D(\bar{\theta}) \tag{1.7}$$

donde

$$p_D = E_{\theta|y}[D] - D(E_{\theta|y}[\theta]) = \bar{D} - D(\bar{\theta}) \tag{1.8}$$

El primer término de la definición (1.7), esperanza posterior de la desviación, es una medida Bayesiana de la bondad de ajuste, mientrás que el término p_D es una medida de la complejidad, la cual es razonable que dependa de la información a priori acerca de los parámetros de interés y de las observaciones. Su justificación está basada en la teoría de información. El DIC puede verse en forma similar al AIC o el BIC, sin embargo, este criterio resulta más satisfactorio que los anteriores ya que tiene en cuenta la información a priori. Además, pueden utilizarse distribuciones a priori impropias, pues cada modelo es considerado por separado.

El DIC puede ser calculado durante una ejecución de MCMC monitoreando θ y $D(\theta)$, y al final de la ejecución simplemente se toma la media muestral de los valores simulados de D, menos la estimación de la deviance usando las medias muestrales de los valores simulados de θ. Valores pequeños de DIC indican un modelo mejor ajustado. El DIC consta de dos términos, uno que representa la bondad de ajuste y la otra una penalidad por incrementar la complejidad del modelo.

Pese a la facilidad en la implementación del cálculo del DIC en cada simulación MCMC, presenta varios inconvenientes que obligan a tener cuidado en su aplicación. La comparación de modelos usando DIC no es invariante a la parametrización, por tanto, debe ser escogida de antemano. Cualquier parámetro de escala desconocido que sea reparametrizado puede conducir a cambios sutiles en el valor del DIC calculado.

El DIC dependerá de qué parte de la especificación del modelo, sea considerada verosimilitud y cuál no. Esta consideración es el centro del asunto, ya que es necesario determinar cuáles parámetros son de interés y cuáles serán contados en el cálculo de p_D. El empleo del DIC es aún un asunto en discusión. Su formulación original (1.7) es apropiada en problemas en los cuales se usan modelos lineales generalizados, pero falla en otros contextos.

1.5. Justificación e importancia de la metodología

Previamente se ha señalado que en Agricultura y específicamente en el contexto epidemiológico son pocos los trabajos, en los cuales se emplea la modelización jerárquica bajo la perspectiva Bayesiana. Esta metodología permite en forma natural incorporar fuentes de variabilidad y de incertidumbre no observadas.

En las últimas décadas, los modelos jerárquicos han llamado la atención de los científicos en muchos campos, y son especialmente adecuados para

estudiar el proceso espacial. Los recientes avances computacionales y el desarrollo de algoritmos eficientes han proporcionado las herramientas necesarias para realizar los complicados cálculos que participan en el modelado jerárquico. Los avances en la modelización jerárquica han permitido manejar enormes bases de datos espaciales georeferenciados.

Aunque la literatura sobre modelización espacial jerárquica es rica, la alta dimensionalidad de estos modelos complica su inferencia. Esta complicación obliga el desarrollo de algoritmos computacionales eficientes por un lado, y la implementación de métodos de reducción de la dimensionalidad por el otro lado.

Las propuestas que desarrollamos en los capítulos 2, 3 y 4 creemos pueden ser vistas como estrategias generales de modelización siempre que los individuos que formen parte del fenómeno estudiado estén asociados a un proceso espacial en una red de localizaciones ó a un espacio continuo. Demostramos su utilidad en el contexto de estudios epidemiológicos de enfermedades en plantas y consideramos que como metodología puede ser extendida a campos distintos al epidemiológico.

Los modelos Bayesianos espaciales aplicados en estudios epidemiológicos específicos (Richardson, 2003), pueden ser de enorme ayuda en la vigilancia de enfermedades ya que permiten tener un conocimiento de la variabilidad del riesgo espacial y de aquella variabilidad no espacial en ausencia "de un punto caliente" ó área de alto riesgo.

La modelización de la dependencia espacial no es sencilla, ya que la posición espacial actúa como un sustituto de covariables no observadas. Por lo tanto, es necesario escoger un modelo apropiado para la dependencia espacial, que permita incorporar indirectamente los efectos espaciales como covariables. Varios autores han propuesto diferentes formas de definir la estructura de vecindad, por ejemplo, Cressie y Chan (1989) asumen la estructura de vecindad como una función de la distancia entre centroides de área. Besag et. al (1991) sugiere un modelo que incluye efectos aleatorios espaciales y

no espaciales y asigna a los efectos aleatorios espaciales una distribución condicional autoregresiva normal (CAR).
En general, las modelizaciones que proponemos a diferencia de lo que suele presentarse en la mayoría de los estudios epidemiológicos, no hará uso de las técnicas Dissease Mapping, ya que la formulación de los modelos no esta basada en datos agregados. Creemos que la incorporación de la dependencia espacial como efectos aleatorios en la jerarquía de los modelos, proporciona una mejor explicación de la variabilidad no observada en el fenómeno. La metodología desarrollada esperamos, se convierta en un mecanismo descriptivo que ayude a visualizar y entender la distribución espacial del riesgo. Esta visualización permitirá al menos los siguientes objetivos:

- La formulación de hipótesis etiológicas que permitan conocer las causas subyacentes en la aparición de los riesgos.
- Realizar vigilancia que garantice el control de las causas y factores de riesgo.
- Reconocer patrones espaciales y temporales subyacentes en el fenómeno estudiado.
- Aportar información real y oportuna a los responsables de la toma de decisiones para disminuir el impacto económico causado por el fenómeno estudiado.

1.6. Objetivos y estructura de la tesis

Esta tesis esta basada en la modelización jerárquica espacial desde la perspectiva Bayesiana para el estudio de enfermedades en cultivos agrícolas. Esta metodología en Epidemiología agrícola es aún un campo poco desarrollado. La necesidad de controlar la variabilidad espacial

presente en la mayoría de los datos observados en Agricultura, exige la búsqueda de nuevas alternativas de modelización capaces de recoger adecuadamente la estructura de interrelaciones entre los individuos estudiados. En este sentido, el objetivo general de la tesis es el aporte de herramientas de modelización generales en el ámbito del análisis espacial, que permitan estudiar la presencia de enfermedades en cultivos agrícolas y describan la distribución de los patrones de contagio cuando se tiene poca información y no se tienen covariables explicativas.

Por un lado, en el capítulo 2, planteamos la modelización espacial basada en datos localizados en una red fija de localizaciones y se considera la componente temporal a través de una covariable que recoge la historia de la enfermedad en el tiempo. Para esta propuesta se ha usado como referencia el modelo de Besag, York y Mollié (1991). A diferencia de estos autores, la variable respuesta la definimos como Bernoulli y no como Poisson. Se demuestra que la dinámica de los riesgos está determinada por los efectos aleatorios (espacial y no espacial) y por la covariable con la información del pasado. Con esta propuesta se demuestra que los fenómenos epidemiológicos en Agricultura se explican mejor al considerar en conjunto la dependencia espacial y temporal.

Como segundo aporte, en el capítulo tres, se presentan tres modelizaciones en el contexto de datos de supervivencia. Cada una de ellas estima el tiempo de supervivencia de los individuos afectados por la evolución de una enfermedad en el tiempo y por la presencia de heterogeneidad no observada. Gracias a la covariable dependiente del tiempo considerada en las tres modelizaciones y a la construcción de una estructura espacial dinámica (*frailty*), se puede relajar el supuesto de proporcionalidad generalmente asumido en el modelo de Cox y enmarcar tales propuestas en el contexto de modelos espacio-temporales.

En el capítulo 3, la primera modelización, define la función de riesgo (*hazard*) a partir de la distribución Weibull con discretización del tiempo

sobre períodos de un año. Se selecciona en especial esta forma paramétrica por su capacidad de predicción de tiempos futuros y por su flexibilidad al poseer dos parámetros, de forma y de escala. Las otras dos modelizaciones enmarcadas en el contexto de riesgos proporcionales, están basadas en procesos de conteo y asignan por un lado a la función de riesgo base un proceso Gamma (Spiegelhalter et al. 1996) y por el otro, una función poligonal (Beamonte y Bermúdez, 2003). Estas modelizaciones con cambios en la función de riesgo base se plantean porque algunos autores señalan que la asignación de procesos Gamma puede conducir a estimaciones insesgadas y engañosas (Mostafa y Ghorbal, 2011).

En los capítulos 2 y 3, la inferencia se hace usando métodos MCMC. Sin embargo, en el capítulo 4, la inferencia Bayesiana no se aborda con métodos MCMC. En este capítulo, se propone un modelo jerárquico capaz de predecir en cualquier punto de la región, la probabilidad o riesgo de enfermedad de un individuo en el contexto agrícola. Gracias a los autores Rue et al. (2009) y Lindgren et al. (2011), es posible proponer un modelo de regresión aditivo con estructura espacial (dentro de la clase de modelos Gaussianos latentes) de variable respuesta Bernoulli controlado por pocos hiperparámetros. El principal beneficio de esta propuesta es computacional, ya que cuando se tratan modelos jerárquicos para datos geoestadísticos, los algoritmos basados en métodos MCMC necesitan de muchas horas y días para las estimaciones. Sin embargo, las aproximaciones obtenidas con el enfoque INLA son más rápidas. Con esta propuesta es posible visualizar con mapas, la distribución de los riesgos de la enfermedad en toda la región estudiada. Hasta el momento, no hemos encontrado trabajos dedicados a la Agricultura que hagan uso de esta metodología. Por tanto se convierte, en una herramienta novedosa para abordar diversos fenómenos en este campo científico.

Las propuestas de modelización se exponen detalladamente a lo largo de los capítulos 2, 3 y 4, respectivamente. En el capítulo 1, se presentan las

razones que sirven de motivación a la metodología desarrollada. Se hace una breve introducción a los procesos espaciales y se presenta una revisión de los trabajos mayormente encontrados en Agricultura. Se explica además, en que consiste el paradigma Bayesiano bajo la modelización jerárquica y se presentan los métodos comúnmente empleados en la inferencia Bayesiana. Se justifica y se resalta la importancia de la metodología desarrollada. Además se presentan los objetivos y estructura general de la tesis. Finalmente, el capítulo 5 contiene las conclusiones generales y las líneas futuras de investigación. El cuerpo de la tesis queda completo con la sección dedicada a las referencias bibliográficas utilizadas.

El código R implementado para la estimación de los modelos propuestos se presenta en los apéndices junto con la sintaxis en OpenBUGS para los modelos desarrollados en los capítulos 2 y 3. Además en los apéndices se presentan los funciones programadas en R para el análisis de los datos y las rutinas diseñadas para la obtención de los resultados presentados a lo largo de la tesis. Así mismo, se presenta la programación desarrollada para realizar el kriging Bayesiano desde la perspectiva determinista presentada.

Capítulo 2

Proceso espacial en una red de localizaciones

Este capítulo está dedicado al estudio de situaciones donde las observaciones provienen de un conjunto fijo de localizaciones. En estos casos, la predicción en otros puntos del espacio no tiene sentido ya que el fenómeno observado únicamente ha ocurrido en localizaciones fijas o ha sido observado agregadamente.

Los modelos estadísticos para este tipo de datos tienen que expresar el hecho de que observaciones próximas tienden a ser parecidas (Kensall y Wakefield, 2002). Por tanto, deben incorporar la relación existente entre las observaciones de localizaciones vecinas. La especificación de estas relaciones a partir de las distribuciones condicionales origina los campos aleatorios markovianos. En este capítulo, se presenta la teoría general relacionada con los campos aleatorios markovianos y los automodelos en el caso discreto y continuo mayormente utilizados en este contexto.

Los modelos jerárquicos Bayesianos y el uso de las técnicas MCMC han favorecido el auge de las técnicas Disease Mapping (Cartografía de Enfermedades) en el contexto de Epidemiología. Se han publicado

numerosas monografías en los últimos años dedicadas al estudio de la distribución geográfica de riesgos (Lawson et al. 1999, Banerjee et al. 2003a, Lawson 2006, 2008). Sin embargo, son pocos los trabajos encontrados en Agricultura que combinen la modelización Bayesiana con variables respuesta no Gaussianas en estudios epidemiológicos.

La mayoría de los modelos de suavización suponen que las observaciones en unidades geográficas cercanas se parecen más que las observaciones de unidades geográficas más distantes. Partiendo de las ideas anteriores, en este capítulo, después de presentar los modelos generalmente empleados en Epidemiología, proponemos una modelización general que puede ser aplicada en estudios epidemiológicos y demostramos con un ejemplo su aplicabilidad en el contexto agrícola para el estudio de enfermedades.

Se proponen modelos con estructuras espaciales dinámicas, es decir, la información espacial para cada individuo i vendrá determinada por el número de vecinos enfermos en periodos anteriores $t-1$ conjugando de esta forma la correlación espacial con la información temporal. Además, la modelización propuesta será capaz de capturar fuentes de variabilidad generadas por factores ocultos de riesgo que esperamos contribuya a la obtención de estimaciones más reales de los riegos.

2.1. Datos agrícolas en una red de localizaciones

Con frecuencia encontramos fenómenos espaciales asociados a localizaciones fijas. La fijación de estas localizaciones puede deberse a las condiciones y diseño del muestreo, que nos suelen conducir a una distribución regular de las mismas y si el fenómeno observado esta condicionado a una situación previa, entonces se tendrá una distribución irregular de las localizaciones.

Los datos espaciales localizados en una red, se obtienen frecuentemente de parcelas cultivadas con cítricos o en otros tipos de cultivos. En esta

sección se presenta un conjunto de datos que se usará como ejemplo para ilustrar las metodologías propuestas en los capítulos 2 y 3. Son datos referidos a árboles de naranjo plantados en cultivos ubicados en la Comunidad Valenciana de España. La industria de cítricos en España tiene un importante valor económico y su producción anual está cerca de 6 millones de toneladas métricas, sobre aproximadamente 285000 hectáreas. España es el cuarto productor más importante del mundo y principal exportador de cítricos frescos. La Comunidad Valenciana (CV), formada por las provincias Alicante, Castellón y Valencia, es una de las principales productoras de cítricos. En esta región al este de España a lo largo de la costa mediterránea, más de 90 millones de árboles son cultivados sobre aproximadamente 185000 hectáreas.

Debido al incremento de árboles enfermos con el virus de la tristeza en la Comunidad Valenciana, la industria cítrica Española ha tenido que invertir muchos recursos y esfuerzos para controlar la aparición de esta enfermedad en los cultivos. Un Programa de mejora de la variedad en cítricos se comenzó en 1975 con los objetivos de recuperar "plantas sin virus" por injerto de brotes in vitro y establecer un banco de germplasm de plantas cítricas sanas. En 1983 estos objetivos fueron expandidos con la importación de citrus budwood de otros países (Navarro et al. 1984). Otras estrategias de control fueron incluidas como: el desarrollo de métodos sensitivos de diagnósticos confiables y la especificación de reactivos para la detección de *Citrus tristeza virus* (CTV) en material vegetal. Además han sido numerosos los estudios epidemiológicos dedicados a seguir la incidencia y extensión de la enfermedad.

Sin embargo, pese a todos los estudios y esfuerzos realizados, siguen presentándose nuevos casos de la enfermedad. La tabla 2.1 muestra el porcentaje de infección en las áreas de cítricos de la CV. El promedio aproximado de incidencia de CTV en las provincias de CV fue 9 % en 1989, 35 % en 1995 y 42 % en 1998. El incremento en la provincia de

Alicante fue menos rápido que en las otras dos provincias, probablemente debido al mayor número de árboles de limón cultivados en esta provincia. Solamente 0.2 % de árboles de limón estaban infectados en 1989 (Cambra et al. 2000a). En la provincia de Valencia la extensión de la enfermedad por tristeza fue más rápida. En algunos municipios entre ellos Moncada, la incidencia total aumentó dramáticamente del 4.8 % en 1989 a casi el 100 % en 1998. La extensión de CTV en la provincia de Castellón entre 1989 y 1998, alarmó a los cultivadores de esta zona, donde muchas variedades de Clementina son cultivadas. El promedio de incidencia de árboles infectados en esta provincia fue 6 % en 1989 y 31 % en 1998. En Villarreal (Castellón), se encontró material vegetal infectado y el porcentaje de árboles infectados era muy alto comparado con otros municipios de la provincia donde la extensión de la enfermedad era principalmente debido a vectores áfidos (Cambra et al. 2000a).

Provincia	**1989**	**1995**	**1998**
Municipios Alicante ($n = 10$)			
Incidencia mínima	0.1	1.3	3.0
Incidencia máxima	14.0	39.5	43.5
Promedio de incidencia	3.0	14.0	20.0
Municipios Castellón ($n = 10$)			
Incidencia mínima	1.2	11.0	17.0
Incidencia máxima	20.4	75.0	83.0
Promedio de incidencia	6.0	28.0	31.0
Municipios Valencia ($n = 16$)			
Incidencia mínima	1.1	13.5	17.5
Incidencia máxima	70.4	98.0	99.9
Promedio de incidencia	17.0	63.0	75.0
Total árboles analizados	**66000**[a.]	**23000**[a.]	**7000**[a.]

[a.] aproximadamente

Tabla 2.1: Incidencia de CTV (%) en Municipios de CV en 1989, 1995 y 1998

Para tratar de entender la dinámica de la enfermedad, se han cultivado diferentes variedades citrícolas en parcelas controladas en la Comunidad Valenciana. Estas parcelas han sido examinadas durante los años transcurridos desde su plantación. Los datos usados en los capítulos 2 y 3, provienen de la parcela de Burriana ubicada en Castellón. Durante el tiempo de estudio no se realizó ningún cambio, arranque o reinjerto en esta parcela. El material vegetal inicial usado se comprobó estuvo libre del virus. La parcela está constituida por 300 árboles de naranjo, plantados en 1993 y que permanecieron libres del virus de la tristeza hasta 1994. Los resultados de la prospección se muestra en la Figura 2.1 y la evolución del virus se muestra en la Figura 2.2. En la Figura 2.1 se observa como aumenta

la proporción de árboles infectados en los años analizados, con porcentajes de infección iguales a 0,33 %, 2 %, 10 %, 17 % y 38 % respectivamente.

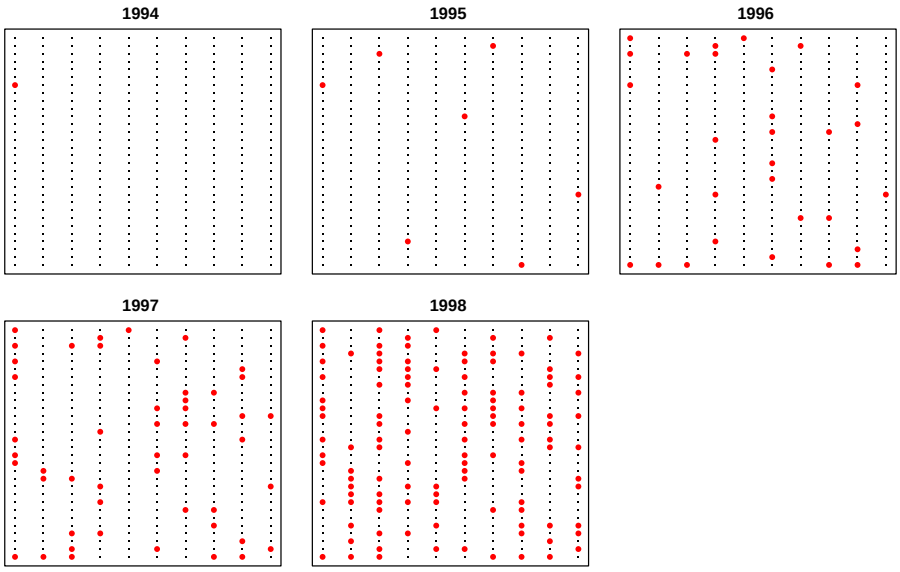


Figura 2.1: Prospección del virus CTV en la parcela Burriana; en rojo (presencia de CTV) y en negro (ausencia de CTV)

La incidencia global del virus en la parcela, expresada como proporción de árboles infectados, permite visualizar el crecimiento de la enfermedad en los últimos años. El virus de la tristeza de los cítricos (CTV; Familia:*Closteroviridae*; Género: *Closterovirus*) (Karasev et al. 1995) es el causante de una de las enfermedades más dañinas y destructivas de los agrios (Bar-Joseph et al. 1989); las pérdidas ocasionadas por ella se estiman en más de 100 millones de árboles injertados sobre naranjo amargo (Román et al. 2004) (unos 38 millones de árboles en América, más de 55 millones en la cuenca mediterránea, especialmente en España, y unos 5 millones en otras zonas), a lo que hay que sumar la baja calidad del fruto y pérdida de producción que se da en varios millones de árboles injertados sobre

patrones tolerantes a la tristeza.

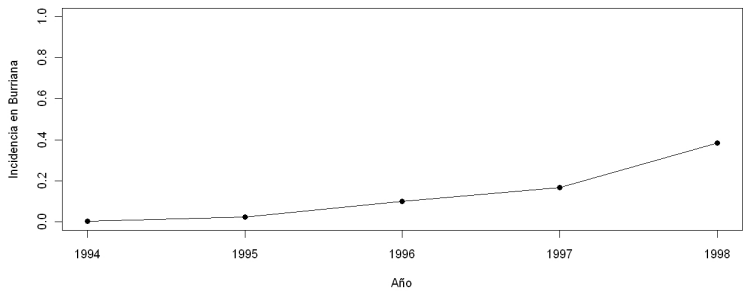


Figura 2.2: Evolución del virus de la tristeza en la parcela Burriana

El virus de la tristeza se disemina principalmente mediante material vegetal infectado o a través de insectos vectores, entre los cuales se encuentran los áfidos como el pulgón café de los cítricos (Toxoptera citricida) considerado como el vector más eficiente (Yokomi et al. 1994). Sin embargo en España, el vector responsable de la enfermedad ha sido el *Aphis gossypii Glover* (Garnsey, 1999). Los datos que usaremos como ejemplo en los tres capítulos, estarán referidos a la presencia del virus de la tristeza en cítricos.

2.2. Campos aleatorios markovianos

Una red de localizaciones o retículo es una colección finita de localizaciones espaciales, que pueden estar distribuidas espacialmente de forma regular o irregular. La relación de vecindad induce un grafo no dirigido con las localizaciones como vértices, de forma que hay una arista entre dos localizaciones cuando son vecinas. Recíprocamente, un grafo de independencia no dirigido entre los puntos del retículo induce un sistema de vecindad donde dos localizaciones son vecinas si existe una arista entre ellas, es decir si son vértices adyacentes en el grafo. Identificaremos una

red de localizaciones con el conjunto de coordenadas de puntos:

$$D \equiv \{(x_i, y_i) : i = 1, \ldots, n\}$$

Para trabajar con procesos de esta naturaleza, es necesario definir un campo aleatorio Markoviano. Sea $\{s_i : (x_i, y_i) \in D \subseteq R^2\}$ para $i = 1, \ldots, n$ el conjunto de localizaciones y sea $Y(s_i)$ la respuesta binaria en el punto s_i, entonces $Y = (Y_1, \ldots, Y_n) = (Y(s_i), \ldots, Y(s_n))'$. Desde esta perspectiva se sabe que las distribuciones condicionales completas para Y_i dependen solamente de los vecinos $j \neq i$. Esta estructura condicional nos permite redefinir la relación de vecindad, a la vez que introducimos algunos conceptos necesarios en la especificación de los modelos. Una localización j es vecina de la localización i si la distribución condicional de Y_i, dados los valores en las demás localizaciones, depende funcionalmente de Y_j, para $j \neq i$.

Una clique es un conjunto de localizaciones en el que cada una de ellas es vecina de todas las demás. Un *campo aleatorio markoviano* es una medida de probabilidad cuyas distribuciones condicionales determinan una estructura de vecindad $\{V_i \subseteq D : i = 1, \ldots, n\}$, es decir que para cada localización i se verifica que

$$P(y_i|y_{-i}) = P(y_i|y_{-i}v_i) \tag{2.1}$$

La estructura de probabilidad de un *campo aleatorio markoviano* viene caracterizada por su función potencial negativa, también denominada función negpotencial. Supongamos que $0 \in \zeta$. Definimos la función potencial negativa $Q(.)$ como

$$Q(y) \equiv \log\{\frac{P(y)}{P(0)}\}, \;\; \mathbf{y} \in \zeta \tag{2.2}$$

Conocer la función $Q(.)$ equivale a conocer $P(.)$, ya que

$$P(y) = \frac{\exp(Q(y))}{\sum_{z\in\zeta} \exp(Q(z))} \tag{2.3}$$

El teorema de Hammersley-Clifford (Clifford, 1990) establece que la función potencial negativa se descompone en sumas de términos correspondientes a las cliques definidas por la estructura de vecindad en la red de localizaciones. Sea k una clique y definimos:

$$y_k \equiv (y_i : i \in k)$$

$$U_k(y_k) \equiv G_k(y_k) \prod_{i \in k} y_i$$

Entonces,

$$Q(y) = \sum_{M_k} U_k(y_k) \tag{2.4}$$

Su importancia en la modelización espacial se debe a la especificación condicional, ya que esta involucra una cantidad pequeña de funciones no nulas. El resultado en sentido opuesto también es importante, y mantiene que la función potencial negativa conduce a una única función de probabilidad conjunta bien definida, siempre que se cumpla la condición de sumabilidad,

$$\sum_{y \in \zeta} \exp(Q(y)) < \infty$$

La construcción de los campos aleatorios markovianos permite establecer una gran variedad de *automodelos* para datos discretos y continuos.
La formulación de los automodelos propuesta por Besag (1974) permite modelizar campos aleatorios markovianos, tomando a las distribuciones condicionales en una familia exponencial lineal y limitando la interacción espacial a cliques de tamaño 2. En el caso discreto,

$$P(y_i|y_{-i}) = \exp[A_i(y_{-i})B_i(y_i) + C_i(y_i) + D_i(y_{-i})],\ i = 1, \ldots, n \tag{2.5}$$

donde $\{B_i(.)\}$ y $\{C_i(.)\}$ tienen formas específicas, siendo $\{A_i(.)\}$ y $\{D_i(.)\}$ funciones de los valores observados en las localizaciones vecinas de i. La forma en que se concreta la interacción entre las localizaciones vecinas es una implicación directa del *teorema Hammersley-Clifford*:

$$A_i(y_{-i}) = \alpha_i + \sum_{j=1}^{n} \beta_{ij} B_j(y_j),\ i = 1, \ldots, n, \tag{2.6}$$

donde $\beta_{ji} = \beta_{ij}$, $\beta_{ii} = 0$ y $\beta_{ik} = 0$ para todo $k \notin V_i$. Para los automodelos, la función Q se puede simplificar en

$$Q(y) = \sum_{i=1}^{n} \{\alpha_i y_i + C_i(y_i) + \sum_{1 \leq i < j \leq n} \beta_{ij} y_i y_j\} \tag{2.7}$$

Los automodelos ofrecen la posibilidad de incluir la influencia de covariables relacionadas a las localizaciones como variables explicativas del proceso espacial. La formulación para el caso continúo es similar a (2.5), sólo que $P(y_i|y_{-i})$ es cambiada por la función de densidad $f(y_i|y_{-i})$.
Los campos aleatorios de Markov incluyen una amplia clase de modelos espaciales, entre estos están:

2.2.1. Caso discreto

Se tienen los siguientes automodelos:

- **Autologístico:** En presencia de datos binarios la distribución condicional es necesariamente de forma logística. El modelo autologístico generaliza la regresión logística introduciendo la dependencia espacial entre las localizaciones. Esta definición puede verse como un modelo jerárquico espacial.

$$\begin{aligned} y_i|y_{-i} \sim Binomial(1, p_i) &\longmapsto \text{1er. nivel} \\ logit(p_i) = \alpha + \textstyle\sum_{j \sim i} \beta y_j &\longmapsto \text{2do. nivel} \end{aligned}$$

  Esto implica que la función potencial es

$$Q(y) = \alpha \sum_i y_i + \beta \sum_{j \sim i} y_i y_j$$

- **Autopoisson:** Cuando los datos espaciales surgen como conteos, la forma natural de modelizar el problema es empleando la distribución de Poisson. Esta definición también puede verse como un *modelo jerárquico espacial*.

$$y_i|y_{-i} \sim Poisson(\lambda_i) \longmapsto \text{1er. nivel}$$
$$log(\lambda_i) = \alpha + \textstyle\sum_{j\sim i} \beta y_j \longmapsto \text{2do. nivel}$$

  Conduce a una función potencial de la forma:

$$Q(y) = \alpha \sum_{i=1} y_i + \beta \sum_{j\sim i} y_i y_j - \sum_i log(y_i!) \tag{2.8}$$

  Una importante aplicación del *autopoisson* es la modelización de la incidencia regional de una determinada enfermedad. A menudo la distribución Poisson es una aproximación de la binomial, que puede ser empleada dando lugar al modelo *autobinomial*. Visto como un *modelo jerárquico* es de la forma:

$$y_i|y_{-i} \sim Binomial(n_i\phi_i) \longmapsto \text{1er. nivel}$$
$$logit(\phi_i) = \alpha + \textstyle\sum_{j\sim i} \beta y_j \longmapsto \text{2do. nivel}$$

  cuya función potencial es:

$$Q(y) = \alpha \sum_i y_i + \beta \sum_{i,j} y_i y_j + \sum_i log\binom{n_i}{y_i}$$

Estos modelos han encontrado aplicación en el análisis de imágenes y la detección con datos tomados desde satélite (Gelfand, et al. 2005).

2.2.2. Caso continuo

El modelo de campo aleatorio markoviano más empleado para datos continuos es el autonormal o autogaussiano,

$$y_i|y_{-i} \sim N(\mu_i + \sum_{j=1}^{n} c_{ij}(y_j - \mu_j), \alpha_i^2) \tag{2.9}$$

En este caso la constante de normalización puede ser evaluada, alcanzando un conocimiento exacto de la distribución de probabilidad conjunta,

$$Y \sim (\mu, (I - C)^{-1}M)$$

donde C es una matriz nxn con elementos c_{ij} tal que $c_{ij}\alpha_j^2 = c_{ji}\alpha_i^2$, y $c_{ii} = 0$, mientras que $M = diag(\alpha_1^2, \ldots, \alpha_n^2)$.
Cuando la distribución condicional no es normal, la constante de normalización es habitualmente intratable, al no tener una expresión analítica. Una desviación del caso Gaussiano puede conducir al planteamiento de otros modelos para datos continuos, como por ejemplo, el autogamma, el autoexponencial y el autobeta.

2.3. Modelización espacial de riesgos

El modelo autogaussiano presenta mayores posibilidades al facilitar una forma cerrada de la distribución de probabilidad conjunta. La construcción del automodelo se realiza mediante probabilidades conjuntas o probabilidades condicionales. Pero existen diferencias en ambas formulaciones, para ilustrar estas diferencias es conveniente la comparación con un proceso temporal.
Si la autoregresión espacial, se hace a través de la expresión condicional de la probabilidad del proceso en cada localización (estructura de vecindad), entonces se construye el modelo condicional autoregresivo espacial (CAR). Mientras que si la autoregresión se incorpora mediante una matriz de dependencia espacial (análogo al término empleado en los modelos de series temporales, expresando la interrelación mutua entre localizaciones

vecinas), entonces se tiene un modelo simultáneo autoregresivo espacial (SAR).

2.3.1. Distribuciones condicionales auto-regresivas

La Distribución Auto-regresiva Condicional Intrínseca (ICAR) propuesta por Besag et al. (1991) pertenece a la familia de Distribuciones Condicionales Auto-regresivas propuestas por Besag (1974). Las CAR son distribuciones multivariantes definidas de forma condicional para cada una de sus componentes. En concreto, diremos que el vector ϕ sigue dicha distribución si para cada una de sus componentes se cumple:

$$\phi_i|\phi_{-i} \sim N\left(\sum_{j:j\neq i} b_{ij}\phi_j, \tau_i^{-1}\right) \tag{2.10}$$

No todos los valores de la matriz B y el vector de precisiones τ conllevan a que la distribución conjunta sea válida, por tanto dichos valores habrán de cumplir ciertas condiciones particulares (Rue y Held, 2005). En concreto, dichas condiciones se cumplen en el caso de la distribución ICAR. Diremos que el vector ϕ sigue una Distribución Autoregresiva Condicional Intrínseca (ICAR) si para cada una de sus componentes se cumple

$$\phi_i|\phi_{-i} \sim N\left(n_i^{-1}\sum_{j:j\sim i} \phi_j, \frac{\sigma^2}{n_i}\right) \tag{2.11}$$

en la expresión (2.11), la relación $j \sim i$ se cumple si ϕ_i y ϕ_j se corresponden con regiones vecinas en cierto sentido y n_i será el número de vecinos de la región i. De esta forma, la distribución ICAR considera que el valor esperado de cada elemento de ϕ coincide con la media de sus valores en las regiones vecinas.

La distribución ICAR es un caso particular de las Distribuciones Condicionales Autoregresivas (CAR) en la que b_{ij} valdrá $1/n_i$ si las

regiones i y j son vecinas y 0 en caso contrario. La precisión τ_i será proporcional al número de regiones que tenga como vecinas (n_i/σ^2). De esta forma la matriz B induce la estructura espacial de los datos en la distribución conjunta del vector ϕ. Además, también se puede demostrar (Banerjee et al. 2003a) que las distribuciones condicionales anteriores, definen una distribución conjunta normal multivariante de la siguiente forma

$$\phi \sim N(\mu \cdot 1, \sigma^2(D - W)^{-1}) \tag{2.12}$$

donde D es una matriz diagonal con elementos $D_{ii} = n_i$ y cada elemento w_{ij} de la matriz W valdrá 1 si y sólo si $i \sim j$, y en otro caso valdrá 0. Sin embargo, ésta no es la única forma de definir W, ya que puede ponderarse por otros criterios, como por ejemplo, proximidad, similitud entre poblaciones, etc. (Ferrándiz et al. 1995; Earnest et al. 2007).
La matriz de precisión de la distribución ICAR presenta una característica que la hace un tanto particular. Todas las filas de la matriz, tal y como la hemos definido, tienen como suma 0. En consecuencia, dicha matriz no es de rango completo y la distribución resultante es impropia. Por esto suele ser habitual considerar que el vector ϕ no puede tomar cualquier valor, sino que sus elementos han de sumar necesariamente 0. De esta forma si la longitud del vector ϕ es n, dicho vector tomará $n-1$ valores independientes (ya que la restricción impuesta disminuye un grado de libertad) y dicho valor coincide exactamente con el rango de la matriz de precisión del vector ϕ. De esta forma, al imponer la restricción mencionada, la distribución ICAR definida sobre el hiperplano del espacio n-dimensional será una distribución propia. Por lo tanto, la condición $\sum_i \phi_i = 0$ permitirá que todas las condicionales sean propias aún cuando la conjunta sea impropia. Existen otras formas de remediar el carácter impropio de la distribución ICAR (Carlin y Banerjee, 2002). La implementación del modelo ICAR es conveniente en la configuración jerárquica Bayesiana debido a su estructura

condicional explícita y la restricción que asegura tener condicionales propias puede ser fácilmente incorporada en el método Gibbs.

La distribución ICAR, tal y como se ha definido, depende de un único parámetro de escala, su precisión (o alternativamente su varianza o desviación típica), de esta forma la estructura de correlación dependerá estrictamente de la configuración geográfica de la región de estudio. Por este motivo, suele ser habitual emplear esta distribución conjuntamente con otro efecto aleatorio Gaussiano independiente para describir la variabilidad que no puede ser explicada por la distribución ICAR.

Los modelos CAR pueden ser extendidos al caso multivariante, así, ϕ_i será un vector de efectos aleatorios asociado con una unidad de área. Si por el contrario, los ϕ_i son vistos como medidas asociadas con una unidad de área i en el tiempo t, entonces se tendrá un modelo condicional autoregresivo multivariante espacio-temporal.

2.3.2. Distribuciones auto-regresivas simultáneas

A diferencia de lo que sucede en las distribuciones CAR, las Distribuciones Autoregresivas Simultáneas (SAR), no consideran las distribuciones condicionales de cada una de las componentes del vector ϕ, sino que realizan una autoregresión de dicho vector en sí mismo para inducir dependencia entre las observaciones. Es decir, si ϕ sigue una distribución SAR, entonces

$$\phi = B\phi + \epsilon, \quad \epsilon \sim N(0, D) \tag{2.13}$$

donde D es una matriz diagonal y B es una matriz estructurada que recoge la estructura espacial de la región de estudio. Suele ser habitual definir la matriz B proporcional a la matriz W definida en la distribución ICAR. De la expresión anterior se deduce que:

$$\phi = B\phi + \epsilon \rightarrow (I - B)\phi = \epsilon \rightarrow \phi = (I - B)^{-1}\epsilon \tag{2.14}$$

Por tanto, en caso de que ϕ siga una distribución SAR resulta

$$\phi \sim N(0, (I - B)^{-1}D((I - B)^{-1})') \tag{2.15}$$

Es decir, la distribución SAR también se reduce a una distribución Normal multivariante con estructura de covarianza dependiente también de la matriz B, pero en forma distinta a los procesos CAR. Las distribuciones CAR y SAR son equivalentes si y sólo si sus matrices de covarianzas son iguales. Cualquier distribución SAR puede ser representada como una distribución CAR, sin embargo, lo contrario no necesariamente es cierto. Una diferencia principal entre las distribuciones SAR y CAR es que la matriz de dependencia espacial para la distribución CAR es simétrica, mientras que esta misma matriz en una distribución SAR no necesita serlo. Aunque esto último puede ser visto como una ventaja en situaciones donde la dependencia espacial de las regiones vecinas sea definida en forma asimétrica, puede conducir al problema de la no identificabilidad en la estimación de los parámetros. Por está razón se prefiere la distribución CAR.

Las distribuciones CAR y SAR son herramientas que permiten inducir distintas formas de dependencia espacial en los modelos de suavización. Sin embargo, las distribuciones CAR se hacen computacionalmente mas indicadas para la inferencia en especial bajo el paradigma Bayesiano.

2.4. Modelización en Cartografía de Enfermedades

Una área común y de interés en estudios bioestadísticos y epidemiológicos es el mapeo de enfermedades (Disease Mapping). En estos campos,

típicamente se suelen manejar datos de conteo de la siguiente clase

$$Y_i = \text{número de casos de enfermos observados en el área } i$$
$$E_i = \text{número de casos de enfermos esperados en el área } i$$

A Y_i se consideran variables aleatorias, mientras que E_i se consideran funciones fijas y conocidas de n_i. n_i, es el número de personas en riesgo de contraer la enfermedad en la región i. Como un punto de partida, podríamos asumir que

$$E_i = n_i \bar{r} \equiv n_i \left(\frac{\sum_i y_i}{\sum_i n_i} \right)$$

$\bar{r}$ es la tasa global de la enfermedad en la región de estudio. E_i por tanto corresponde a una especie de "hipótesis nula". Con esta hipótesis se espera una tasa de enfermedad constante en toda la región. Este proceso se conoce como estandarización interna y con este proceso, se pierden grados de libertad al estimar la tasa global r de nuestros datos actuales.
Un mejor enfoque sería hacer un proceso de estandarización externa. Esto implica crear tasas para la enfermedad por grupos de edades y estratificar la población de acuerdo al grupo, así $E_i = \sum_j n_{ij} r_j$, donde n_{ij} representa el número de años de la persona en situación de riesgo en el área i por grupo de edad j, r_j es la tasa de enfermedad en el grupo de edad j (tomada de la tabla creada). En cualquier caso, en su forma más simple, un mapa de enfermedad es sólo una muestra de las tasas de enfermedad primarias superpuestas sobre las unidades de área.
Los modelos de cartografía de enfermedades tratan de solventar fundamentalmente un problema de estimación en áreas pequeñas sobre la región de estudio. El pequeño tamaño de las unidades geográficas consideradas en multitud de estudios geográficos conlleva a dificultades de estimación que los modelos de cartografía tratan de minimizar. La idea fundamental en todos es compartir la información entre unidades de estudio, ya que

después de todo dicha información es la única herramienta conocida para mejorar las estimaciones que queremos realizar.
En los métodos tradicionales aplicados para representar datos en áreas, si no se comparte información entre las unidades de estudio y E_i no es muy grande, es decir, la enfermedad es rara o la región i es suficientemente pequeña, se suele establecer al número de casos observados en la $i-$ésima región una distribución

$$Y_i|\eta_i \sim Po(E_i\eta_i) \tag{2.16}$$

donde E_i es el número de muertes esperadas, η_i es el riesgo verdadero en la región i. La estimación Máximo Verosímil del riesgo η_i sería simplemente

$$\hat{\eta}_i = RME_i = \frac{Y_i}{E_i} \tag{2.17}$$

RME es la Razón de Mortalidad Estandarizada en la región i. Esta estimación de los riesgos se corresponde al modelo Bayesiano en el que la distribución previa de los riegos η_i es una distribución Uniforme impropia para toda la recta real. La expresión (2.16) suele ser, salvo raras excepciones en la que la hipótesis de Poisson se cambia por Binomial, el punto de partida de la mayoría de modelos de cartografía de enfermedades. Nótese que $Var(RME_i) = Var(Y_i)/E_i^2 = \eta_i/E_i$, de esta forma $\widehat{Var(RME_i)} = \hat{\eta}_i/E_i = Y_i/E_i^2$. Esto a su vez permite el cálculo de los intervalos de confianza tradicionales de η_i (aunque resulte poco manejable ya que los datos son discretos), así como las pruebas de hipótesis.

2.4.1. Modelo Poisson-Gamma

La forma de modelizar los riesgos hasta el momento presentada, detecta la sobredispersión, pero no permite estimar la superficie de riesgo subyacente. Una primera forma de modelar los riesgos, es transfiriendo información

entre las unidades de estudio con lo cual podrán obtenerse estimaciones más realistas que las resultantes de (2.16).
Una de las propuestas más sencillas de transferencia de información entre regiones es considerar que todas ellas siguen una distribución común. La elección por defecto propuesta por Clayton y Kaldor (1987) fue considerar como distribución previa para los riesgos

$$\eta_i \sim Ga(a, b) \tag{2.18}$$

ya que esta distribución es la conjugada de la distribución Poisson asumida en (2.16). Esta propuesta se conoce habitualmente como modelo Poisson-Gamma o modelo de Clayton y Kaldor. La media posterior del riesgo en la región i viene dado por

$$\hat{\eta}_i = \frac{Y_i + a}{E_i + b} = \frac{y_i + \frac{\mu^2}{\sigma^2}}{E_i + \frac{\mu}{\sigma^2}} \tag{2.19}$$

$$= \frac{E_i \left(\frac{y_i}{E_i}\right)}{E_i + \frac{\mu}{\sigma^2}} + \frac{\left(\frac{\mu}{\sigma^2}\right)\mu}{E_i + \frac{\mu}{\sigma^2}} \tag{2.20}$$

$$= w_i RME_i + (1 - w_i)\mu \tag{2.21}$$

donde, $w_i = E_i/[E_i + (\frac{\mu}{\sigma^2})]$, de modo que $0 \leq w_i \leq 1$. De esta forma la estimación puntual (2.19) combina la información de la mortalidad observada en la propia región (Y_i, E_i) y la información propia de la distribución de los riesgos en todas las regiones estudiadas (a,b). Esta estimación es aproximadamente igual a (2.17) cuando w_i es cercano a 1, es decir, cuando E_i es grande y los datos son muy informativos, o cuando σ^2 es grande y la previa es débilmente informativa. Por otro lado, (2.19) será aproximadamente igual a μ cuando w_i este cerca de 0, es decir, cuando E_i sea pequeño y los datos sean escasos, o cuando σ^2 sea pequeño y la previa sea muy informativa.

La estimación de los parámetros a y b se obtiene mediante métodos Bayesianos empíricos, es decir, se obtendrá una estimación puntual de dichos valores en lugar de fijarnos a algún valor arbitrario u obtener su distribución a posteriori.
La modelización Poisson-Gamma ha supuesto un primer paso de indudable relevancia en la mejora de las Estimaciones Máximo Verosímiles de los riesgos. Sin embargo, una crítica habitual que se hace del modelo es que no considera la estructura espacial de los datos, por tanto, en la estimación del riesgo influyen de la misma forma tanto regiones alejadas como cercanas.

2.4.2. Modelo Poisson-Lognormal

La hipótesis de estimar el riesgo sin considerar la estructura espacial no parece del todo razonable desde el punto de vista epidemiológico, ya que localizaciones próximas deberían compartir factores de riesgo similares y por tanto sus riesgos deberían ser también similares.
El modelo Poisson-Gamma aunque suele ser conveniente, computacionalmente falla, al no tener en cuenta la correlación espacial entre los riesgos (η_i). La limitación del modelo Poisson-Gamma se resuelve usando las distribuciones normales descritas en el apartado (2.3). Bajo este enfoque y siguiendo la propuesta de Besag, York y Mollié (1991), en adelante BYM, suele ser habitual asignar una distribución previa a los riesgos de la siguiente forma

$$\log(\eta_i) = \mu + \theta_i + \phi_i \tag{2.22}$$

donde μ es el valor promedio del logaritmo de los riesgos, ϕ_i es un efecto aleatorio con estructura espacial y los θ_i recogen la variabilidad ajena a la componente espacial. Usando la propuesta BYM, Banerjee et al. (2004) propone modelar el logaritmo de los riesgos con la influencia de covariables explicativas, es decir, $\log(\eta_i) = x_i\beta+\theta_i+\phi_i$. Estas covariables son ecológicas

y agregadas a nivel de región y no a nivel individual. Esta forma de modelar los riegos es muy usada en estudios epidemiológicos, pero debe usarse con cuidado, ya que el nivel de agregación puede conducir a problemas de sesgo ecológico. Sin embargo, estos autores, esperan que las covariables sean capaces de explicar alguno o quizás todos los patrones espaciales de Y_i.

Siguiendo a Banerjee et al. (2004) se tendrá entonces que

$$\theta_i \sim N(0, \tau_h)$$
$$\phi \sim ICAR(\tau_c)$$

donde los θ_i capturan la heterogeneidad entre las regiones. Los θ_i son efectos que capturan la variabilidad *Extra-Poisson* del logaritmo de los riesgos relativos sobre la región de estudio completa.

Los ϕ_i son los parámetros que hacen de esta formulación un modelo realmente espacial, ya que capturan las similitudes entre las regiones (agrupamientos). Esta componente modeliza la variabilidad *Extra-Poisson* del logaritmo de los riesgos relativos que varían localmente haciendo que regiones cercanas tengan tasas más similares. τ_h y τ_c son los parámetros de precisión (recíproco de la varianza) y controlan la magnitud de cada efecto aleatorio.

En cuanto a los parámetros de precisión (alternativamente varianza o desviación típica), Gelman (2006) demuestra que la asignación de sus previas no puede elegirse en forma arbitraria. La discusión y el por qué se elige cierto tipo de previas para los parámetros de precisión se presenta en el apéndice 1.

La idea fundamental de esta modelización es que el efecto aleatorio espacial modele los factores de riesgo que abarcan más de una unidad de estudio y en consecuencia hace sus riesgos espacialmente dependientes. Mientras, el efecto heterogéneo resulta conveniente para describir aquellos factores de

riesgo que pudieran tener un efecto interno en las unidades de estudio y que provoca que el riesgo en cualquiera de estás pueda ser muy diferente al de sus vecinas. La importancia, en términos relativos, que tendrán estas componentes aleatorias dependerá de sus desviaciones típicas, parámetros que deberán estimarse en el modelo. Este modelo no sólo se propuso en su día de forma teórica sino que desde entonces ha sido utilizado repetidamente en la literatura epidemiológica (Ferrándiz et al. 2002, 2004; Lope et al. 2006; Barceló et al. 2008).

2.4.3. Otras modelizaciones para riesgos

A diferencia del modelo BYM existen propuestas en la literatura que también combinan la distribución ICAR con otro tipo de procesos, así en Lee y Durban (2009) se propone, desde un enfoque frecuentista, la utilización de splines para modelizar la tendencia espacial de largo rango, junto a efectos aleatorios ICAR para modelizar la dependencia espacial de rango corto.

Desde el punto de vista frecuentista también se han hecho otras propuestas. Leroux (2000) propone modelizar los riesgos con un único efecto aleatorio cuya matriz de precisión es proporcional a

$$(1-\lambda)\cdot I+\lambda\cdot(D-W)$$

donde I es la matriz identidad, W la matriz de estructura espacial de la región de estudio y D una matriz diagonal con el número de vecinos de cada región. λ será un parámetro a estimar por el modelo restringido al intervalo $[0,1]$. De esta manera, para $\lambda=0$ la estructura de covarianza del efecto aleatorio se corresponderá con la de un proceso de independencia completa, mientras que por el contrario para $\lambda=1$ dicha matriz reflejará exclusivamente la estructura espacial de la región de estudio.

Otra estructura espacial alternativa a la propuesta de BYM consiste en el uso de distribuciones CAR propias. En caso de incluir un efecto aleatorio con esta distribución deja de resultar necesario incluir otro efecto aleatorio heterogéneo, ya que el efecto CAR será capaz de modular la dependencia espacial de los riesgos pudiendo reproducir entonces estructuras de dependencia espacial o procesos espacialmente independientes. Aunque esta modelización supone una alternativa a la propuesta BYM, su uso ha sido mucho menos extendido.

La modelización mediante mixturas puede ser otra forma de modelizar la dependencia espacial en modelos de suavización de riesgos. Un ejemplo de ello serían los modelos de Poisson Zero-Inflamados (ZIP) (Lambert, 1992; Ugarte et al. 2004). Estos modelos consideran que los propios datos observados siguen una mixtura de distribuciones de Poisson de dos componentes, la primera de ellas de valor esperado igual a 0 y la otra de valor esperado positivo. Los modelos ZIP suponen

$$Y_i \sim p \cdot Po(0) + (1-p) \cdot Po(E_i\eta_i)$$

donde $p \in [0,1]$ y η_i (o su algoritmo) puede modelizarse de distintas formas, o bien como un único valor común para todos los riesgos (Ugarte et al. 2006), o como un efecto aleatorio posiblemente con estructura espacial (Ramis Prieto et al. 2007). La idea subyacente en esta modelización es que para enfermedades raras los casos observados en las distintas unidades de estudio pueden presentar gran número de ceros, esto puede conducir a sobredispersión, si se asume en la primera capa del modelo una distribución Poisson. Por tanto, en casos donde existe gran número de ceros, asumir en la primera capa una distribución Poisson no resulta adecuado y los modelos ZIP pueden ser una buena opción.

Otra posibilidad que ofrecen los modelos de mixturas es la flexibilización de la estructura espacial definida por la distribución ICAR en el modelo BYM. Lawson y Clark (2002) proponen modelizar el riesgo como

$$log(\eta_i) = \mu + (p_i \cdot \phi_i + (1 - p_i) \cdot \psi_i) + \theta_i \tag{2.23}$$

$p_i \in [0, 1]$ al igual que en el modelo ZIP. θ_i es un efecto aleatorio con distribución Normal y ϕ_i sigue una distribución ICAR, mientras que ψ_i representa una estructura ICAR pero basada en una distribución doble-exponencial o Laplace en lugar de una distribución Normal. La distribución de ψ_i tiene colas mucho más pesadas que la distribución Normal, por lo que es particularmente adecuada en la modelización de estructuras espaciales con discontinuidades o saltos. La mixtura de los efectos aleatorios ϕ y ψ hace que esta propuesta sea capaz de reproducir discontinuidades donde sea necesario y estructuras espaciales más suaves donde haga falta. En Congdon (2007) se proponen otras modelizaciones de mixturas para los riegos similares a las de (2.23).

En los modelos de mixtura presentados hasta el momento, la probabilidad de pertenecer a una componente de la mixtura es independiente para cada una de las unidades de estudio. Sin embargo en la literatura es posible encontrar otros tipos de modelizaciones, ejemplos de ello, son los trabajos de Knorr-Held y Rasser (2000), Denison y Holmes (2001), Gangnon y Clayton (2000), Green y Richardson (2002). En estos trabajos se toma en cuenta la estructura espacial de la región a la hora de asignar las unidades de estudio a las componentes de la mixtura. Además de los modelos de dependencia ya mencionados, se ha propuesto algún modelo de suavización de riesgos basado en medias móviles (Best et al. 2000a, Botella 2010).

Otra propuesta interesante en la modelización espacial de los riesgos que incluye la estructura de vecindad de la región de estudio es la conocida como Wombling (Lu y Carlin, 2005; Lu et al. 2007). En esta modelización se pretende tener en cuenta las posibles discontinuidades en la distribución del riesgo entre regiones vecinas. Es decir, se desea eliminar del modelo aquellas relaciones de vecindad correspondientes a regiones vecinas cuyos riesgos sean impares.

2.5. Modelización de riesgos con estructura espacial dinámica

En las secciones anteriores se ha presentado la metodología utilizada en los últimos años para la modelización espacial de los riesgos en estudios de Disease Mapping. Hemos resaltado las ventajas y desventajas aportadas por cada propuesta. En estudios de Disease Mapping suele ser habitual considerar conteos de datos agregados en un conjunto de localizaciones geográficas. Esta red o grid de localizaciones, por lo general tiene estructura irregular.

El objetivo de este apartado, es desarrollar una modelización general enmarcada en la filosofía de las técnicas Disease Mapping, sin llegar a considerar datos agregados en áreas. La metodología propuesta será capaz de presentar estimaciones de las probabilidades posteriores de enfermar de cada individuo ubicado en una red fija de localizaciones. En este caso, partimos de la rara excepción de no considerar a la variable respuesta como Poisson sino partir de una distribución Bernoulli.

El modelo BYM permite modelizar el riesgo considerando la transferencia de información espacial entre las unidades de estudio. Partiendo de esta idea, planteamos otra forma de modelizar los riesgos de enfermar para ciertos individuos ubicados en una red fija de localizaciones. En este caso, el intercambio de información espacial entre individuos no ocurre en forma agregada sino a través de distancias, luego la noción de vecindad no estará basada en regiones contiguas adyacentes como en la propuesta de BYM.

La propuesta de Banerjee et al. (2004) supone que la inclusión de covariables en el modelo BYM mejora el conocimiento de los parámetros que se desean estimar. Estos autores comentan, que cuando se trabajan unidades de área en forma de: barrios, municipios, provincias, etc., el nivel de agregación de las covariables puede afectar significativamente

la estimación. En este sentido, el modelo que proponemos no considera covariables con algún nivel de agregación. Esto supondrá una diferencia importante en nuestra modelización y con esto se evitará el problema de datos desalineados (Mugglin et al. 2000).

La modelización que desarrollamos puede convertirse en una estrategia para modelar fenómenos en donde los individuos estén asociados a un proceso espacial determinado por una red de localizaciones fijas. Por tanto, es una herramienta con aplicabilidad en diversos contextos científicos que puede sin duda alguna aprovecharse en el estudio de enfermedades en cultivos agrícolas.

La configuración de una grid lineal hará posible incluir en el modelo propuesto la noción de espacio y tiempo en forma conjunta. Para esta formulación conjunta, no se considera la teoría propiamente de series temporales, sin embargo, la información espacial se incluye en forma dinámica, así la estructura espacial considerada en cada instante de tiempo t cambia en función al número de vecinos enfermos que tenga cada individuo en el instante de tiempo $t-1$. Esta configuración confiere a nuestra modelización un carácter espacio-temporal, mediante el cual, se podrá describir el riesgo de enfermar como una colección de procesos (espacio y tiempo) que ayudarán a comprender la evolución de los riegos reales.

Después de revisar la literatura existente en modelización de riesgos, es posible entonces proponer un modelo jerárquico de riesgos suavizados con estructura espacial dinámica, capaz de estimar las probabilidades de enfermar para cierto grupo de individuos. A este modelo lo denominaremos Hierarchical Dynamic Spatial Model y se denotará en adelante como HDSM.

Antes de formular el modelo Hierarchical Dynamic Spatial Model (HDSM) y puesto que se trata de estimar la probabilidad de enfermar a partir de la transferencia de información entre individuos cercanos, la estructura

espacial finalmente considerada se modula a través de distancias menores o iguales a los 10 metros. Se considera sólo el radio de influencia espacial en estas distancias porque después de los 10 metros se observó que el efecto espacial dejaba de ser significativo. El criterio de vecindad basado en distancias puede ser de diversa índole y dependerá de la naturaleza de los puntos del retículo y del fenómeno a estudiar.
En la primera capa del modelo HDSM al igual como se hace mayormente en los modelos de suavización de riesgos, definimos la distribución para la variable respuesta. En este caso, se asigna a Y_{it} una variable Bernoulli, que indica la presencia (1) o ausencia (0) de enfermedad en el individuo i en el periodo de tiempo t, es decir asumimos

$$Y_{it} \sim \text{Ber}(\pi_{it}) \tag{2.24}$$

con $i = 1, \ldots, n$ y $t = 1, \ldots, T$. En la segunda capa de la estructura jerárquica modelizamos el logit de la probabilidad asociada con el riesgo de enfermar de cada individuo i, en el instante de tiempo t, es decir, π_{it}.

$$\text{logit}(\pi_{\text{it}}) = \beta_0 + \beta_1 x_{it} + \theta_{it} + \phi_{it} \tag{2.25}$$

o equivalentemente

$$\pi_{it} = \frac{\exp(\beta_0 + \beta_1 x_{it} + \theta_{it} + \phi_{it})}{1 + \exp(\beta_0 + \beta_1 x_{it} + \theta_{it} + \phi_{it})} \tag{2.26}$$

De esta forma, la probabilidad viene dada por la influencia de un intercepto, una covariable y por dos efectos aleatorios. x_{it} representa el número de vecinos enfermos que tiene el individuo i en el tiempo t dados los individuos enfermos en periodos $t-1$. Es decir, recoge la evolución de la enfermedad en el pasado. Con el término θ_{it} se incluye la heterogeneidad individual y no observada de cada individuo i en el tiempo t. El término ϕ_{it}

incorpora la estructura espacial asociada a cada individuo i en el instante actual t pero eliminando el efecto espacial de aquellos individuos que vienen enfermos en instantes anteriores $t-1$. Luego, la asociación espacial cambiará en cada instante t.

La estructura jerárquica para el modelo HDSM queda completa con la estructura probabilística siguiente

$$\theta_{it} \sim \text{Normal}(0, \sigma_\theta^2) \tag{2.27}$$

$$\phi_{it}|\phi_{-i} \sim N\left(n_i^{-1} \sum_{j:j\sim i} \phi_j, \frac{\sigma_\phi^2}{n_i}\right) \tag{2.28}$$

la relación $j \sim i$ en (2.28) se cumple si ϕ_i y ϕ_j se corresponde con individuos ubicados a distancias menores o iguales a los 10 metros y n_i es el número de vecinos del individuo i en el instante de tiempo t.

Con esta modelización es posible considerar a través de θ_{it} fuentes de variabilidad ajenas al proceso espacial que suelen ser desconocidas y que afectan notablemente la estimación de los parámetros. Además se establece que los riesgos de enfermar dependerán en cierto sentido de una componente que describe la evolución en el pasado, x_{it}, y de una estructura espacial que recoge la dinámica de la enfermedad en el presente, ϕ_{it}.

$$\beta_0 \sim N(0, 0.001) \tag{2.29}$$

$$\beta_1 \sim N(0, \sigma_\beta^2) \tag{2.30}$$

Las desviaciones típicas que definen las varianzas (o alternativamente la precisión) para las variables aleatorias β_1, θ, ϕ se definen como

$$\sigma_\beta \sim \text{Unif}(0,1) \tag{2.31}$$

$$\sigma_\theta \sim \text{Unif}(0,1) \tag{2.32}$$

$$\sigma_\phi \sim \text{Unif}(0,1) \tag{2.33}$$

La estructura definida en (2.25) permite modelar el proceso de dependencia espacial en forma distinta a la propuesto por Besag et al. (1991), ya que el orden de dependencia no depende exclusivamente de regiones adyacentes sino que dependerá de las distancias definidas. La distribución asignada a la desviación típica en (2.33) que define la varianza del efecto aleatorio espacial se mantiene para todos los periodos considerados. Igual consideración se mantiene para la varianza del efecto aleatorio de heterogeneidad. El modelo HDSM descrito desde la ecuación (2.24) hasta la ecuación (2.33) será en adelante referido como modelo base. Esta modelización puede ser extendida a otros casos, esto implicará el incremento en el número de parámetros a estimar.

A partir de la definición jerárquica base podemos configurar otras modelizaciones. La primera de ellas, considera en su estructura aditiva, el intercepto (β_0), la covariable (β_1) y el efecto de heterogeneidad (θ_{it}). Mientras que la segunda configuración que establecemos reconoce sólo la influencia del intercepto, la covariable y el efecto aleatorio espacial (ϕ_{it}). A estas dos configuraciones las llamaremos HDSM1 y HDSM2, respectivamente. En estas nuevas modelizaciones se mantienen los supuestos asumidos en la varianza de cada parámetro definido en el modelo base HDSM.

Otra modelización que proponemos mantiene los supuestos para los dos efectos aleatorios definidos en HDSM. Se considera junto a los dos efectos aleatorios (θ_{it}, ϕ_{it}) el intercepto β_0, pero no toma en cuenta la influencia de la covariable. Con esta modificación queremos saber si la inclusión de la covariable es determinante o no en la bondad de ajuste del modelo base

propuesto. A esta modelización la denotamos como HDSM3.

Consideramos una cuarta modelización donde suponemos que las variables β_0 y β_1 para cada individuo i cambian en el tiempo t, es decir, β_{0t}, β_{1t}. En este caso, al asumir que la variabilidad del intercepto cambia con el tiempo, estamos suponiendo que el riesgo base para cada individuo puede ser diferente. Por otra parte, a la desviación típica de β_1 le asignamos una distribución distinta en cada periodo, de esta forma suponemos que la información histórica contenida en la covariable puede influir en forma distinta en cada instante t. En esta modelización, además consideramos los dos efectos aleatorios tal cual como se definen en HDSM. A esta propuesta la denotamos como HDSM4.

A partir de la modelización HDSM4, se definen otros modelos. El primero de ellos, considera a β_{0t}, β_{1t} junto con la influencia del efecto de heterogeneidad (θ_{it}). Este efecto aleatorio se define igual que en HDSM. A esta nueva propuesta la denotamos como HDSM5. Otro modelo a partir de HDSM4 es aquel que considera a β_{0t}, β_{1t} y el efecto aleatorio espacial ϕ_{it}. En cuanto a la varianza asignada a cada ϕ_{it} se mantienen las características probabilísticas definidas en HDSM. A esta modelización la denotamos como HDSM6. Se plantea también a partir de HDSM4, un modelo que sólo considera el efecto de β_{0t} y los dos efectos aleatorios sin considerar la covariable, al que llamaremos HDSM7. En este caso tanto θ_{it} como ϕ_{it} conservan las definiciones dadas en HDSM.

En la modelización que denotamos HDSM8, se considera la estructura jerárquica más compleja, en este caso, se supone que no existe una distribución común para los parámetros considerados. De esta forma, el riesgo de enfermar para el individuo i en el instante de tiempo t, estará determinado por efectos aleatorios distintos. Esto implicará que a los hiperparámetros que definen las varianzas de β_{0t}, β_{1t}, θ_{it} y ϕ_{it} se les asignará distribuciones previas diferentes en cada instante de tiempo t, resultando en la siguiente reformulación

$$\beta_{0t} \sim N(0, \sigma^2_{\beta_{0t}}) \tag{2.34}$$

$$\beta_{1t} \sim N(0, \sigma^2_{\beta_{1t}}) \tag{2.35}$$

$$\sigma_{\beta_{0t}} \sim \text{Unif}(0, 1) \tag{2.36}$$

$$\sigma_{\beta_{1t}} \sim \text{Unif}(0, 1) \tag{2.37}$$

$$\sigma_{\theta_t} \sim \text{Unif}(0, 1) \tag{2.38}$$

$$\sigma_{\phi_t} \sim \text{Unif}(0, 1) \tag{2.39}$$

Finalmente, se propone una última modelización a partir del modelo HDSM8, en la que se mantiene β_{0t}, β_{1t} junto con el efecto aleatorio espacial ϕ_{it} dado por las desviaciones típicas definidas en (2.39) para cada instante t. A esta propuesta la denotamos como HDSM9.

Con las modelizaciones propuestas, se pretende generar una metodología que abarque situaciones desde las más sencillas a las más complejas y ofrecer las mayores posibilidades de aplicación. Además intentamos encontrar el modelo más parsimonioso, con el cual sea posible explicar la evolución espacio-temporal de los riesgos en presencia de datos espaciales asociados a una red de localizaciones. Gracias a la formulación conjunta del proceso espacio-temporal se espera contar con más grados de libertad para la estimación de los parámetros que beneficie el ajuste de los modelos. La concatenación de la información espacial y temporal requiere de mayor esfuerzo computacional.

La estimación posterior de cada parámetro se lleva a cabo mediante el software Bayesiano OpenBUGS (Lunn et al. 2009; Spiegelhalter et al. 2007). Generamos largas cadenas, de las cuales se rechazan las 5000 primeras iteraciones (hasta obtener la convergencia a la distribución posterior) y sólo guardamos 1 de cada 5 iteraciones (para reducir la autocorrelación en las cadenas) hasta obtener una muestra de 10000. En todos los casos se generan dos cadenas paralelas simultáneamente. El

diagnóstico de convergencia se supervisa con el paquete CODA (Plummer et al. 2006).

2.6. Presencia de CTV en una parcela agrícola

En esta sección ilustramos las distintas modelizaciones propuestas con datos provenientes de una parcela cultivada con árboles de naranjo contagiados con el virus de la tristeza. Las características del conjunto de datos y el impacto del virus se describen en la sección (2.1).

Se sabe que la enfermedad de la tristeza se produce por la alimentación en un árbol sano de un pulgón vector virulífero que antes se ha alimentado en un árbol infectado y ha adquirido el virus. El tiempo que tarda en extenderse el virus por todo el árbol es de varios meses y es poco probable que un pulgón que se alimente en un árbol recién infectado adquiera el virus. Por tratarse de individuos adultos alados, la distribución espacial de las nuevas infecciones producidas en un año determinado no es sencilla. Dicha distribución debe considerar para cada árbol la proximidad de los árboles infectados a su alrededor, es decir, considerar árboles ya infectados en periodos anteriores $t-1$ y aquellos infectados en el instante t.

En general, hemos considerado diferentes distancias en los modelos propuestos, distancias que van desde los 10 metros hasta los 40 metros, esto con la finalidad de distinguir el rango de variabilidad espacial presente en el fenómeno. Sin embargo, en las modelizaciones finales sólo se han considerado distancias menores a los 10 metros. El número de árboles en la parcela es 300, por tanto $i = 1, \ldots, 300$. Los tiempos considerados se obtienen al plantear la relación de vecindad entre los árboles del año t con los árboles enfermos en el año $t-1$, es así, que tenemos los periodos de tiempo, dados por las relaciones 95 dado 94, 96 dado 95, 97 dado 96 y 98 dado 97, correspondientes a $t = 1$, $t = 2$, $t = 3$ y $t = 4$ respectivamente.

Las tablas (2.2) y (2.3) muestran como β_0 y β_1 superan el test de Geweke

(1992) y el test de Heidelberger y Welch (1983). Los valores negativos estimados para β_0 disminuyen la probabilidad básica de enfermar, sin embargo la influencia del resto de efectos aleatorios considerados determinan los valores de las probabilidades obtenidas. Esto sugiere que las probabilidades de enfermar vienen determinadas por otras fuentes de variabilidad que el modelo HDSM es capaz de capturar.

Además en la figura (2.3) se observa que β_0 y β_1 alcanzan muy pronto la convergencia en las dos cadenas simuladas. Igualmente sucede con la convergencia de la deviance para el modelo HDSM; la deviance supera el test de Gelman y Rubin (1992), ya que ambas cadenas se estabilizan cerca del valor 1.0 (figura 2.4).

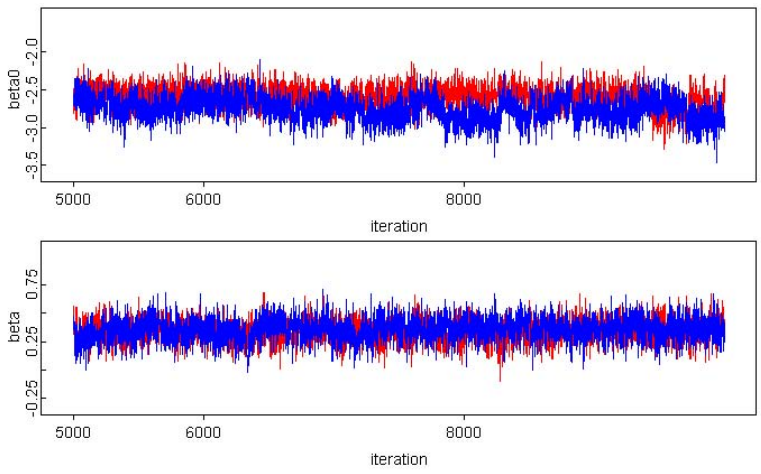


Figura 2.3: Convergencia para β_0 y β_1 bajo el modelo HDSM

Estadístico Z para β_0	**Estadístico Z para β_1**
-0.77	1.31

Tabla 2.2: Diagnóstico Geweke para β_0 y β_1 bajo el modelo HDSM

Variable	**Test de Estacionariedad**	**p-valor**
β_0	pasado	0.37
β_1	pasado	0.14

Tabla 2.3: Diagnóstico Heidelberger para β_0 y β_1 bajo el modelo HDSM

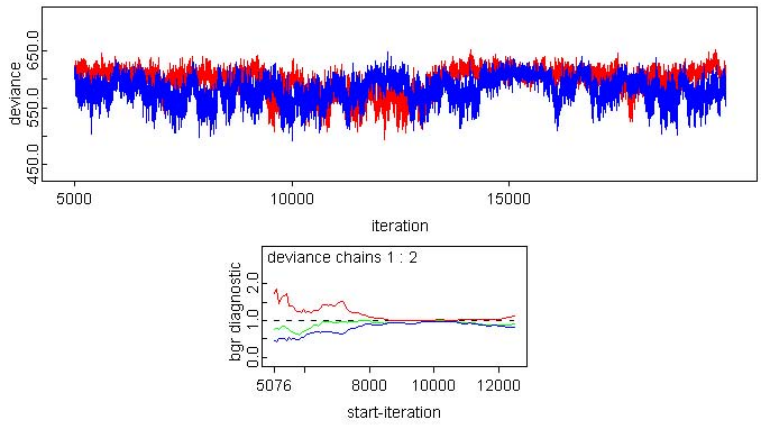


Figura 2.4: Convergencia para la deviance bajo el modelo HDSM

En la tabla (2.4) se muestra el resumen con la descripción de la bondad de ajuste para cada modelo propuesto. Se observa como los modelos HDSM y HDSM2 son los que mejor ajusten presentan. En ambos casos se presenta la estructura espacial ϕ_{it}, lo que sugiere que la probabilidad de un árbol enfermar dependerá notablemente de la existencia de vecinos enfermos ubicados a distancias menores a los 10 metros. Así mismo, en el modelo

HDSM se tiene la influencia del efecto de heterogeneidad θ_{it} que incorpora la variabilidad no observada del individuo i. Con la presencia de este efecto en el modelo se reconocen factores de riesgos ocultos asociados a cada individuo i.

Modelo	**Descripción del modelo**	**DIC**	$\mathbf{p}_D$
HDSM	$\beta_0 + \beta_1 x_{it} + \theta_{it} + \phi_{it}$	622.7	20.71
HDSM1	$\beta_0 + \beta_1 x_{it} + \theta_{it}$	654.7	16.67
HDSM2	$\beta_0 + \beta_1 x_{it} + \phi_{it}$	625.5	11.18
HDSM3	$\beta_0 + \theta_{it} + \phi_{it}$	633.3	30.91
HDSM4	$\beta_{0t} + \beta_{1t} x_{it} + \theta_{it} + \phi_{it}$	627.4	29.26
HDSM5	$\beta_{0t} + \beta_{1t} x_{it} + \theta_{it}$	668.3	28.12
HDSM6	$\beta_{0t} + \beta_{1t} x_{it} + \phi_{it}$	634.5	18.24
HDSM7	$\beta_{0t} + \theta_{it} + \phi_{it}$	640.6	38.67
HDSM8	$\beta_{0t} + \beta_{1t} x_{it} + \theta_{it} + \phi_{it}$	631.5	22.37
HDSM9	$\beta_{0t} + \beta_{1t} x_{it} + \phi_{it}$	642.4	15.5

Tabla 2.4: Resumen con la bondad de ajuste para los distintos modelos propuestos

Los modelos HDSM y HDSM2 son los que mejor explican el fenómeno estudiado, en este sentido, el modelo HDSM esta diciendo que la probabilidad de un árbol enfermar esta determinada por varios efectos, uno, el que recoge el efecto de árboles enfermos en años anteriores (historia del proceso), el otro, por la variabilidad implícita de cada sujeto y por último, de un efecto espacial que recoge la propagación de la enfermedad en el presente. La tabla (2.5) muestra las estimaciones posteriores para los parámetros involucrados en la modelización HDSM. Se observa que el coeficiente que acompaña a la covariable que describe la historia de la enfermedad toma valores positivos, con lo cual es evidente que la dinámica

de la enfermedad esta determinada por un proceso de contagio ocasionado por árboles enfermos en periodos $t-1$.

En la tabla (2.5) también se presenta la estimación de la variabilidad posterior relacionada con cada efecto aleatorio. Para evitar el problema de identificabilidad (Banerjee et al. 2004) se define la proporción de variabilidad explicada por cada efecto aleatorio. De esta forma es posible separar usando la definición empírica de la desviación típica marginal, la cuota de variabilidad explicada por el efecto espacial, de la variabilidad producto del efecto heterogéneo. En este sentido, la proporción de variabilidad espacial se define como $\alpha = \frac{\sigma_\phi}{\sigma_\theta+\sigma_\phi}$. Luego $(1-\alpha)$ será la proporción de variabilidad explicada por el efecto de heterogeneidad. La proporción de variabilidad espacial (α) respecto a la no espacial $(1-\alpha)$ sugiere que existe una fuerte relación espacial en el fenómeno estudiado. Así, es claro que la probabilidad de un árbol enfermar dependerá significativamente de la influencia de árboles enfermos cercanos.

Parámetro	**Media**	**Desv. Típica**	**I. Cred. 95 %**
β_0	-2.61	0.16	[-2.96,-2.32]
β_1	0.34	0.10	[0.13,0.54]
α	0.52	0.14	[0.31,0.79]
Deviance	602.0	19.55	[550.6,630.9]
σ_θ	0.51	0.25	[0.13,0.96]
σ_ϕ	0.49	0.1013	[0.33,0.72]

Tabla 2.5: Parámetros estimados para el modelo HDSM y variabilidad para los efectos aleatorios en el último año, $t=4$

La tabla 2.6 muestra algunas probabilidades estimadas con el modelo HDSM para árboles que aún estaban sanos en el último instante de tiempo considerado. Se tienen árboles con probabilidades que de acuerdo a la

escala de riesgos definida (figura 2.6) se pueden clasificar como individuos con riesgos ponderados entre bajo, moderado y alto de enfermar.

π[árbol,t]	Media	D.Típica	I. Cred. 95 %
π[1,4]	0.054	0.042	[0.008,0.043,0.162]
π[116,4]	0.082	0.051	[0.017,0.072,0.209]
π[125,4]	0.118	0.077	[0.029,0.099,0.324]
π[172,4]	0.255	0.119	[0.069,0.240,0.536]
π[236,4]	0.363	0.148	[0.123,0.347,0.706]

Tabla 2.6: Estimación posterior para π_{i4} bajo el modelo HDSM

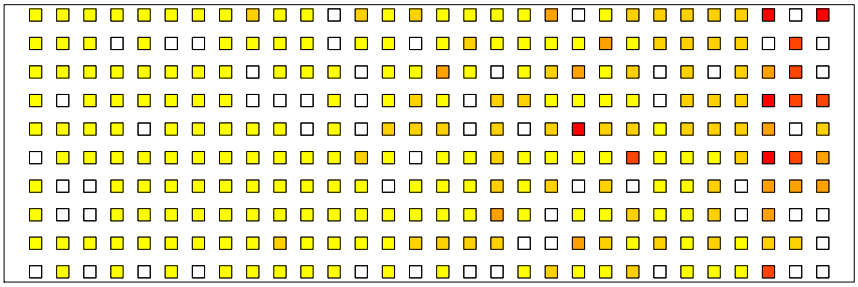


Figura 2.5: Mapa con los riesgos posteriores (π_{i4}) para los árboles analizados en $t = 4$

Por otro lado, en la figura (2.5) se muestra un mapa con la distribución de los riesgos para los 300 árboles analizados. Se observan recuadros con colores que van desde el amarillo claro (riesgo bajo), amarillo mas intenso (riesgo moderado), amarillo oscuro (riesgo medio), naranja (riesgo alto) y color rojo (riesgo máximo) para aquellos árboles con riesgos de enfermar

para el último año considerado. La figura (2.6) al igual que la (2.5), evidencian que la mayoría de los árboles presentan riesgos de enfermar ponderados entre bajo, moderado y medio.

En la figura (2.5), los recuadros en color blanco son árboles que enfermaron en años previos al año 1998. Las escalas que definen a los colores mostrados en la figura (2.5) se establecen en función a la estimación posterior de π_{i4} obtenida bajo el modelo HDSM. Los valores de estas escalas se detallan en la figura (2.6).

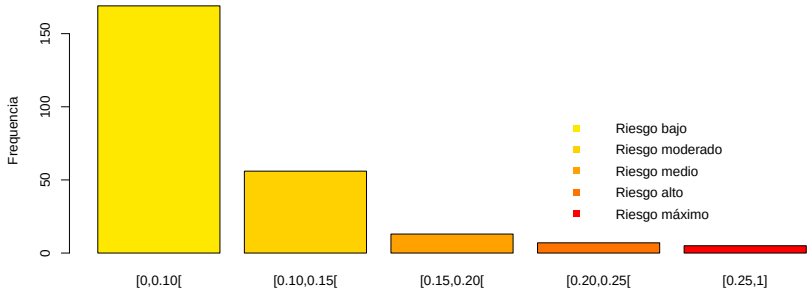


Figura 2.6: Escala de riesgos en función a la estimación posterior de π_{i4}

En el trabajo de Spiegelhalter et al. (2002), los autores sugieren que modelos con valores del DIC que superan en menos de 3 unidades al "mejor" (modelo con menor DIC) son "equivalentes" en cuanto a su capacidad predictiva. Los modelos cuyo DIC supera entre 3 y 7 unidades al modelo con menor DIC se consideran "ligeramente inferiores" y finalmente, los que superan en más de 7 unidades al de menor DIC son considerados sustancialmente inferiores. En cuanto a los resultados presentados en la tabla 2.4 se puede concluir que los modelos HDSM y HDSM2 son equivalentes en cuanto a su capacidad de predicción, mientras que el modelo HDSM4 es ligeramente inferior al mejor modelo encontrado, siendo

el resto de los modelos sustancialmente inferiores al mejor modelo.
En general, al analizar las estimaciones encontradas con el modelo HDSM, es posible reconocer la existencia de un patrón importante de contagio entre los árboles y se demuestra el carácter permanente de la enfermedad en el cultivo, además es evidente que el virus de la tristeza afectará a más árboles a medida que transcurre el tiempo.

2.7. Conclusiones del capítulo

Los métodos de modelización espacial y espacio-temporales son cada vez más importante en las ciencias ambientales y en otras ciencias en donde los datos se derivan de procesos en entornos espaciales. Desafortunadamente, la aplicación de modelos espaciales tradicionales basados en covarianza resultan inapropiados computacionalmente en muchos problemas. Además estos métodos no permiten cuantificar incertidumbres correspondientes a los parámetros del modelo. Sin embargo, el enfoque Bayesiano en el contexto de modelos jerárquicos espaciales dinámicos planteados, permite no sólo cuantificar la incertidumbre de los riesgos de enfermedad en cierto grupo de individuos, sino reconocer la presencia de fuentes de variabilidad no observadas en el fenómeno. Todo esto conduce a estudios más realistas de problemas no sólo en el contexto de Agricultura sino en otras áreas. Consideramos oportuno mencionar, la utilidad de esta metodología para definir medidas preventivas y de control en la reducción de enfermedades que redunden en beneficios económicos.
Combinar los procesos espaciales y temporales supone una mayor complejidad y aunque desde la perspectiva Bayesiana parezca fácil de abordar, la modelización conjunta se convierte en un enorme reto en la práctica. Gracias a la formulación jerárquica y a la descomposición de los procesos en condicionales relativamente simples es posible llegar a modelar estructuras espaciales y temporales complejas.

Los avances en las técnicas computacionales para la modelización jerárquica espacial en las últimas dos décadas, han proporcionado un marco flexible a los investigadores en diferentes campos científicos. En Agricultura y en especial en el contexto epidemiológico, son pocos los trabajos que han aprovechado los beneficios computacionales ofrecidos por el enfoque Bayesiano, por lo tanto, el aprovechamiento que puede hacerse en Agricultura es enorme y con la ilustración de la metodología demostramos que su aplicación en esta área es posible y enriquecedora.
Después de ilustrar la metodología propuesta, es posible agregar las siguientes observaciones finales:

- La estimación positiva de β_1 en las modelizaciones expresa que la covariable con la historia del proceso de contagio es determinante en la dinámica de la enfermedad.

- La poca diferencia en el DIC de los modelos HDSM y HDSM2, sugiere que cualquiera de ellos es capaz de reconocer la importancia del patrón espacial en la dinámica de la enfermedad.

- Los modelos jerárquicos Bayesianos espaciales dinámicos, evidencian que la probabilidad de un árbol enfermar dependerá de un proceso espacial determinado por el pasado y por el presente, y por fuentes de variabilidad ajenas al proceso espacial.

- La metodología empleada sugiere la presencia de patrones de contagio entre árboles ubicados a distancias menores a los 10 metros, así como un patrón de movimiento en el principal vector transmisor *A. gossypii* a estas distancias.

- Los modelos jerárquicos Bayesianos espaciales pueden ser herramientas muy útiles en estudios epidemiológicos y permiten estudiar la incidencia y extensión de una enfermedad en cultivos agrícolas.

El hecho de que los modelos con mejores ajustes contengan en su estructura alguno de los efectos aleatorios o inclusive ambos, coincide con lo afirmado por Leroux et al. (1999). Estos autores demostraron haciendo un estudio de simulación que si los datos son realmente independientes, un modelo con sólo efectos ICAR sin efectos independientes, tendrá una seria sobrestimación en el parámetro de precisión del modelo ICAR. Por tanto, proponer modelos espaciales que incluyan sólo la estructura ICAR sin considerar una estructura con valores independientes (efecto heterogéneo) conducirá a pobres estimaciones en los coeficientes de regresión.
Hacer caso omiso de las correlaciones espaciales en un modelo lineal generalizado puede tener graves efectos en las inferencias, ya que si la correlación es positiva (fenómeno más probable en aplicaciones ecológicas), se pueden obtener errores estándar en los coeficientes de regresión demasiado pequeños, originando que los efectos se juzguen como significativos cuando realmente no lo son (Kneib et al. 2008).

Capítulo 3

Modelos de supervivencia para procesos espaciales en una red de localizaciones

Han sido muchos los enfoques usados para el modelamiento de datos epidemiológicos, incluyendo simples series temporales, modelos puramente espaciales, modelos espacio-temporales y modelos de supervivencia. Típicamente las mediciones en situaciones de enfermedad son tomadas sobre una serie de tiempo discreto. En este caso el orden de las observaciones son de importancia vital, ya que observaciones futuras dependen de periodos anteriores ("lag").

Una forma común de representar este tipo de datos es a través de modelos Autoregressive Moving Average (ARMA) (Box y Jenkins, 1976), los cuales modelan los valores de cada punto en la serie de tiempo por medio de la combinación de dos procesos independientes; el primero (autoregresivo) que trata a los valores observados como una suma ponderada de sus valores en puntos de tiempo anteriores y la segunda (media móvil), corrige los errores en las predicciones pasadas, calculando una suma lineal ponderada de los términos de error. El número de componentes en cada caso es variable y está relacionado con el "lag".

Los modelos ARMA se soportan en la condición de normalidad y en la estacionariedad de los datos. En el caso de que los datos no sean estacionarios, tal condición puede alcanzarse por diferenciación entre tiempos sucesivos, con lo cual se puede remover la tendencia de los datos. El grado de diferenciación requerido actúa como un parámetro adicional en el modelo, obteniendo así, los modelos Autoregressive Integrated Moving Average (ARIMA).

En situaciones epidemiológicas, los modelos ARMA/ARIMA son limitados. Las mediciones en este contexto, corresponden al número de individuos infectados en el tiempo, variable que no se distribuye normalmente. Una forma natural de modelar este tipo de datos es a través de la utilización de modelos lineales generalizados. Alternativamente los datos epidemiológicos pueden ser vistos como procesos puntuales espacio-temporales, donde el fenómeno probabilístico de interés es el tiempo y la ubicación de la infección (Diggle, 2003). Otra forma ampliamente utilizada en datos epidemiológicos es agregar los datos en unidades de área (Lawson, 2008), sin embargo este tipo de metodología, no permite hacer predicciones y su interés se centra en la identificación de tendencias y patrones, así como en las posibles causas de la enfermedad (ver Capítulo 2).

Este capítulo tiene por finalidad emplear el análisis de supervivencia para abordar problemas asociados a procesos espaciales en una red de localizaciones desde una perspectiva Bayesiana. Es común encontrar trabajos en aplicaciones médicas y de epidemiología en otras áreas, en las que se emplea este análisis y de estos trabajos son pocos quienes consideran el efecto espacial en su modelización.

En el capítulo se introducen las técnicas de supervivencia comúnmente empleadas, además se presentan los modelos basados en métodos no paramétricos, semi-paramétricos y paramétricos aplicados en estudios de supervivencia. Se hace especial énfasis en el modelo de riesgos proporcionales propuesto por Cox (1972) por su potencialidad. Estas bases

teóricas nos servirán para proponer una metodología que permita a partir de datos espaciales observados en una red de localizaciones estudiar el fenómeno con modelos de supervivencia.

La dinámica de cualquier enfermedad está determinada por la influencia de efectos aleatorios espaciales y temporales, así como por covariables que pueden o no estar determinadas por el tiempo. Hasta el momento no hemos encontrado trabajos enfocados en supervivencia que combinen modelos jerárquicos Bayesianos con datos espaciales en una red fija de localizaciones. Tampoco hemos encontrado literatura que haga uso de esta metodología que incorpore además un efecto aleatorio espacial junto con covariables dependientes del tiempo. Por tanto, la metodología que proponemos en este capítulo, intenta aprovechar las ventajas del paradigma Bayesiano en la construcción de modelos jerárquicos de supervivencia.

En general, las técnicas de supervivencia se pueden aplicar en una amplia gama de situaciones, siempre que se cumplan las condiciones declaradas por Cox y Oakes (1984): en primer lugar, determinar a partir de donde se inicia el tiempo, es decir debe decidirse la escala para medir el progreso del tiempo y finalmente, la definición exacta de la *falla* o evento a considerar. En nuestra modelización la falla se referirá a un individuo enfermo, un dato censurado corresponderá a un individuo sano y el término *frailty* corresponderá a la presencia del efecto aleatorio espacial.

3.1. Introducción

El análisis estadístico de supervivencia es el conjunto de métodos y técnicas estadísticas diseñadas para modelizar y analizar el tiempo transcurrido entre eventos bien definidos, al que solemos referirnos como *tiempo de supervivencia*. Aunque el análisis de supervivencia toma su nombre de aplicaciones médicas, se utiliza en muchas otras áreas del conocimiento.

La característica principal de las técnicas de supervivencia es su capacidad

para utilizar información de *tiempos censurados*. En la mayoría de los estudios es probable que el tiempo de supervivencia de algunos individuos sólo haya sido observado parcialmente, bien porque permanecen vivos (aún no ha ocurrido el evento final) al terminar el experimento o bien porque hayan abandonado el estudio antes de su finalización.

El objetivo de los estudios de supervivencia es explicar una evolución, por lo que es necesario un seguimiento de los pacientes, son estudios *longitudinales*. El inconveniente de los estudios retrospectivos es la posible modificación en definiciones, incluso la propia definición del diagnóstico y cambios en la propia población estudiada.

Dependiendo del contexto, el *evento de interés* puede ser de diferente índole. Por ejemplo, en epidemiología, el investigador puede considerar como evento, el contagio o muerte del paciente a causa de alguna enfermedad; en aplicaciones de Ingeniería, el evento de interés puede ser la *falla* de un componente físico, mecánico o electrónico de ciertos equipos industriales.

El tiempo (año, meses, semanas o días) de estudio de la ocurrencia de cierto evento de interés comienza en un punto inicial de observación bien definido, hasta establecer un punto final, a este tiempo nos referiremos como *período de seguimiento* o tiempo de estudio. El tiempo de la ocurrencia del evento de interés comúnmente es llamado *tiempo de supervivencia* y representa el período desde el comienzo de observación de un individuo hasta que experimente el evento de interés.

La *variable respuesta*, la cual mide el tiempo de supervivencia de un individuo, es una variable aleatoria T con valores reales positivos y definida sobre un espacio de probabilidad $(\Omega,\mathbb{S},\mathbb{P})$, donde Ω es el espacio muestral, $\mathbb{S}$ es la σ-álgebra de eventos y $\mathbb{P}$ es una medida en la σ-álgebra $\mathbb{S}$ de subconjuntos de Ω, es decir, $\mathbb{P}$ es la medida de probabilidad en $(\Omega,\mathbb{S})$.

Los modelos de supervivencia son generalmente definidos en términos de la función de riesgo (*hazard*). Si los tiempos de supervivencia se consideran

que provienen de una distribución continua, entonces el riesgo representa la tasa instantánea de falla en un punto en el tiempo, dado ese punto de supervivencia. La función de riesgo únicamente determina la distribución de los tiempos de supervivencia (Cox y Oakes 1984, Kalbfleisch y Prentice, 2002) y permite obtener otras cantidades de interés, por ejemplo, la probabilidad de un individuo infectarse la próxima semana o el rango de individuos más probable de infectarse en el futuro. También es sencillo incorporar información de covariables en la función de riesgo.

Un aspecto fundamental que afecta la interpretación de los datos en el análisis de supervivencia es la censura. Las observaciones censuradas son las que contienen información incompleta; típicamente se tienen observaciones censuradas por la derecha, correspondiente a individuos que aún no han experimentado el evento de interés al final del período de estudio. Sin embargo, sigue siendo importante el aporte de información sobre el proceso de supervivencia de fondo que debe ser incorporado en la formulación de la probabilidad. El análisis de supervivencia proporciona un método para esto, ponderando los valores de censura. También hay metodología que incluye la censura por la izquierda, si el modelo así lo requiere.

La especificación de observaciones censuradas adquiere una importancia adicional en los modelos epidemiológicos, donde la exposición a la enfermedad cambia tanto en el espacio como en el tiempo. Al incorporar el aspecto espacial en modelos de supervivencia, se tendrán algunos individuos que por su localización tengan poco riesgo de enfermar, haciendo que exista una mayor variabilidad y sesgo en las estimaciones de los parámetros de interés. Esto se produce por dos razones principales: por la dependencia espacio-temporal que existe en la media de los procesos (efectos de primer orden) y por la rigidez de las localizaciones entre individuos vecinos (efectos de segundo orden). La incorporación en los modelos de supervivencia de covariables espacio-temporales y de efectos

aleatorios que recogen la correlación espacial y temporal permite resolver estos inconvenientes.
Kleinbaum (1995), señala tres objetivos básicos del análisis de supervivencia, estos son

Objetivo 1. Estimar e interpretar la función de supervivencia, o la función de riesgo, a partir de unos datos de supervivencia.

Objetivo 2. Comparar las funciones de supervivencia, o de riesgo, de dos o más grupos de individuos.

Objetivo 3. Establecer la posible relación de algunas covariables con los tiempos de supervivencia.

Para alcanzar estos objetivos es necesario considerar al *tiempo de supervivencia* como una variable aleatoria en la población estudiada. Este tiempo será siempre una *variable aleatoria continua no negativa*. Es no negativa puesto que todo tiempo observado será mayor o igual a cero. Es continua pues dados dos tiempos de supervivencia conocidos cualesquiera, llamémosles t_1 y t_2, es posible que el tiempo de supervivencia del próximo individuo observado esté entre t_1 y t_2, por muy cercanos que estén t_1 y t_2 entre sí.
En el estudio de cualquier variable aleatoria, entre ellas el tiempo de supervivencia T, el interés se centra siempre en las probabilidades asociadas a observaciones de dicha variable. Estas probabilidades constituyen la *Distribución de probabilidades* de la variable aleatoria y pueden obtenerse a partir de la función de distribución de T, $F(t)$.
Consideremos en primer lugar, el caso para datos homogéneos, donde T es una variable aleatoria positiva que representa el tiempo de *falla* o ocurrencia del evento de interés. La función de supervivencia, $S(t)$, es definida tanto para el caso discreto como continuo como la probabilidad de que un individuo sobreviva después del tiempo t, es decir

$$S(t) = P(T \geq t) \qquad 0 < t < \infty \tag{3.1}$$

Aquí $0 <$ S(t) ≤ 1, ya que S(0)= 1 y $\lim_{t \longrightarrow \infty}$ S(t)= 0. La distribución de T puede ser únicamente determinada por la función de supervivencia o como es común por la función de riesgo (*hazard*) o por la función de densidad de probabilidad.

Para el caso continuo, la variable aleatoria T, la función de densidad, *f(t)* viene dada por

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \tag{3.2}$$

donde la función de distribución acumulativa $F(t) = P(T < t) = 1 - S(t)$, así que $S(t) = \int_t^{\infty} f(u)du$. La función de riesgo (*hazard*), *h(t)*, esta definida como el instante potencial de *falla* en el tiempo t, dada la supervivencia en t, es decir

$$h(t) = \lim_{\triangle t \longrightarrow 0} \frac{P(t \leq T < t + \triangle t \mid T \geq t)}{\triangle t}, \quad \triangle t \ll 1 \tag{3.3}$$

Esta es una medida positiva y algunas veces es llamada tasa de falla específica en el tiempo t. Siguiendo el teorema fundamental de cálculo, la ecuación (3.2) puede ser re-escrita como

$$\begin{aligned} h(t) = \frac{dF(t)}{dt} &= \lim_{\triangle t \longrightarrow 0} \frac{F(t + \triangle t) - F(t)}{\triangle t} \\ &= \lim_{\triangle t \longrightarrow 0} \frac{P(T < t + \triangle t) - P(T < t)}{\triangle t} \\ &= \lim_{\triangle t \longrightarrow 0} \frac{P(t \leq T < t + \triangle t)}{\triangle t} \end{aligned} \tag{3.4}$$

Usando (3.4) y la definición de probabilidad condicional, el riesgo puede

ser escrito como

$$\begin{aligned} h(t) &= \lim_{\triangle t \longrightarrow 0} \frac{P(T \leq t + \triangle t)}{\triangle t[P(T \geq t)]} \\ &= \frac{f(t)}{S(t)}, \end{aligned} \tag{3.5}$$

y de (3.2) se tiene que

$$\begin{aligned} h(t) &= -\left[\frac{\frac{dS(t)}{dt}}{S(t)}\right] \\ &\Longrightarrow S(t) = \exp\left(-\int_0^t h(u)du\right) \end{aligned} \tag{3.6}$$

La cantidad $H(t) = \int_0^t h(u)du$ es conocida como función de riesgo acumulativa. Si T es una variable aleatoria discreta entonces la función de probabilidad $f(t) = P(T = t)$ determina la probabilidad exacta de falla en el tiempo *t*. Del mismo modo, la función de riesgo, *h(t)*, puede ser escrita como

$$\begin{aligned} h(t) = P(T = t|T \geq t) &= \frac{P(T = t)}{P(T \geq t)} \\ &= \frac{P(T = t)}{\sum_{j|t_j \geq t} P(T = t_j)} \end{aligned} \tag{3.7}$$

Por lo tanto, es sencillo definir a $P(T = t)$ y $P(T \geq t)$ en términos de la función de riesgo al considerar que $1-h(t)$ es la probabilidad condicional de supervivencia hasta el tiempo *t*. Así al ordenar los tiempos de supervivencia $t_1 < \cdots < t_n$, se tiene que

$$P(T = t_i) = h(t_i) \prod_{j=1}^{i-1} (1 - h(t_j)) \tag{3.8}$$

y

$$P(T \geq t) = \prod_{j|t_j \leq t} (1 - h(t_j)) \tag{3.9}$$

La relación entre las funciones de densidad, distribución y supervivencia es la siguiente

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^{\infty} f(x)dx \tag{3.10}$$

Otra relación importante es

$$S(t) = \exp\{-H(t)\} f(t) h(t) \exp\{-H(t)\} \tag{3.11}$$

donde, *H(t)*, función de riesgo acumulado se define como $\int_0^t h(x)d(x)$.

3.2. Técnicas para datos de supervivencia

Los tiempos de supervivencia se analizan habitualmente mediante técnicas no paramétricas, como las curvas de Kaplan-Meier, puesto que su distribución es difícil de modelizar matemáticamente mediante un modelo de probabilidad paramétrico. Sin embargo, incluir covariables en el análisis de supervivencia es difícil sin incorporar una relación paramétrica entre las covariables y la supervivencia. A continuación se presentan los métodos comúnmente empleados en el análisis de supervivencia.

3.2.1. Métodos no paramétricos

Las técnicas no paramétricas son útiles en particular para explorar datos de supervivencia, ya que no restringen a los datos a que sigan una distribución particular. Las estimaciones de las funciones de supervivencia y de riesgo pueden ser obtenidas fácilmente, así como las medidas estadísticas descriptivas usuales (media, mediana, cuantiles e intervalos de confianza). En el caso de no existir observaciones censuradas, la función de supervivencia empírica puede ser usada para estimar la función de

supervivencia en el tiempo t. Esto indica que la probabilidad de supervivencia más allá del tiempo t es la proporción del número total de pacientes en el estudio que siguen sanos después de t y viene dada por

$$\widetilde{S}(t) = \frac{\texttt{No. de individuos con tiempos de supervivencia} > t}{\texttt{No. total de individuos}}$$

Si los datos contienen observaciones censuradas entonces la función empírica anterior no es válida. En este caso, se puede dividir el período de estudio en un conjunto de intervalos discretos. Las estimaciones de supervivencia se basan entonces en la proporción del número total de individuos considerados en "riesgo" en cada intervalo. El método más conocido en este tipo de casos, es el método de Kaplan-Meier denotado como KM o estimador del producto límite (Kaplan y Meier, 1958). En este capítulo este método se presentará como marco de referencia y será aplicado en el conjunto de datos analizado.

El método desarrollado por Kaplan y Meier (1958), considera una muestra de n individuos y supone que el tiempo de *falla* ocurre al inicio de cada intervalo de tiempo, por lo que cada intervalo contiene un sólo fracaso. Si hay $r \leq n$ fallas en $t_{(j)}$, $j = 1, \ldots, r$ los tiempos se ordenan tal que el primer intervalo $[t_{(0)}, t_{(1)}]$ no contenga falla (i.e. $t_{(0)}$ es el tiempo de origen). En el caso de observaciones empatadas, la censura es tomada después de la falla.

n_j denota el número en riesgo antes de $t_{(j)}$ y sea d_j el número de fallas observadas. Se asume que las fallas son independientes y el estimador de la probabilidad de supervivencia entre $t_{(j)}$ y $t_{(j+1)}$ viene dado por $\frac{n_j - d_j}{n_j}$ con el correspondiente estimador de supervivencia para $t_{(j)} \leq t < t_{(j+1)}$, dado por

$$\widehat{S}(t) = \prod_{k=1}^{j} \left(\frac{n_j - d_j}{n_j} \right) \tag{3.12}$$

es decir, la probabilidad de supervivencia de $t_{(j)}$ a $t_{(j+1)}$ y de todos los intervalos anteriores. A (3.12) se le conoce como estimador de la función de supervivencia Kaplan-Meier. Se puede ver que en (3.12) hay un decrecimiento de la función con $\widehat{S}(0) = 1$ y $\widehat{S}(t)$ es constante en cada intervalo de tiempo $t_{(j)} \leq t < t_{(j+1)}$, con $j = 0, \ldots, r$, y $t_{(r+1)} = \infty$. A partir de este estimador se puede calcular la media, mediana, cuartiles, errores estándar asociados e intervalos de confianza para la estimación de la supervivencia, así como las estimaciones de la función de riesgo y la de riesgo acumulativa.

Los gráficos de estimación de supervivencia y de las curvas de riesgo pueden ser útiles en la inferencia de la verdadera distribución de supervivencia. También se pueden usar otros test no paramétricos como el log-rank y Wilcoxon para comparar grupos y estimaciones.

El método de Kaplan-Meier, es conocido también como *estimador producto* de la función de supervivencia. Cuando no hay datos censurados este método y el de estimación por proporciones darán el mismo resultado. Para incorporar la mayor información posible de los datos censurados, el método de Kaplan-Meier utiliza reiteradamente la regla producto de probabilidades $P(A \bigcap B) = P(A)P(B|A)$, que permite calcular el porcentaje de un porcentaje.

3.2.2. Métodos semi-paramétricos

El tipo de métodos no paramétricos mencionados en la sección (3.2.1) proporcionan formas útiles para las estimaciones de supervivencia y de la función de riesgo asociada, inclusive cuando se incorpora información censurada. Sin embargo, un aspecto clave en el modelado de la supervivencia es investigar el efecto de las covariables en el tiempo de supervivencia, por lo que se hace necesario cambiar el enfoque.

Ya que los métodos no paramétricos únicamente estiman la correspondiente distribución de supervivencia, Cox (1972) propuso especificar un modelo a

través de la función de riesgo, donde cada individuo cuente con un vector de covariables x, así, el riesgo en el tiempo t se compone de dos partes: la primera modela el riesgo en ausencia de covariables (función de riesgo base) y la segunda representa una función paramétrica con el efecto de las covariables en el tiempo de falla por encima del riesgo base.

Cox (1972) primero introdujo el enfoque de riesgos proporcionales, conocido como Proportional hazard model (PH) por su siglas en inglés, como una forma de incorporar información de covariables en un modelo de supervivencia sin tener que asumir ninguna forma en la distribución de los datos. El modelo PH esta definido en términos de la función de riesgo como sigue

$$h(t, x) = h_0(t)\psi(\beta; x) \tag{3.13}$$

donde x es un m-vector de variables explicativas, $\psi(.)$ es una función paramétrica de x y $h_0(t)$ es la función de riesgo base, es decir cuando $x = 0$. Aquí β es un m-vector de parámetros. Una forma común de especificar $\psi(.)$ es usando la función de vínculo *log* en las covariables, es decir, $\psi(\beta; x)$ =exp$(\beta^T x)$. A la expresión (3.13) se conoce como modelo PH *semi-paramétrico*.

Cox y Oakes (1984) ofrecen diversos argumentos a favor del uso de este tipo de modelos. Con respecto a la formulación del modelo, sostienen que el efecto de multiplicar la covariable por un factor constante no es irracional, ya que la evidencia empírica en algunos campos lo respalda. Así mismo, la censura y la aparición de diversos tipos de fallas pueden ser incluidas fácilmente en el modelo y además es posible realizar adaptaciones en estos casos a pesar de no conocer la distribución de supervivencia.

Para adaptarse a los riesgos proporcionales definidos en el modelo PH (3.13), Cox (1972) desarrolló un método de verosimilitud parcial, llamado así porque no hace uso de la censura actual, pero si de los tiempos de supervivencia sin censura. Para ello se considera a n individuos con $r \leq n$

con tiempos de falla ordenados $t_{(j)}$, $j = 1, ..., r$. En una formulación estándar, un individuo i no-censurado con tiempo de falla $t_{(j)}$ y vector x de covariables contribuye $f(t_i, x_i)$ a la verosimilitud; sin embargo, ya que la forma de $f(.)$ es desconocida, una verosimilitud alternativa se consigue usando la probabilidad condicional de falla del individuo i en $t_{(j)}$ dada la supervivencia en $t_{(j)}$ y la noción de intervalos de riesgo.

La técnica funciona en el supuesto de que los intervalos entre tiempos de falla sucesivos no pueden contribuir con información a la verosimilitud ya que conceptualmente h_0 en estos intervalos puede ser cero. La verosimilitud se construye entonces en base a la información dada por los individuos de todo el conjunto de tiempos de falla observados.

Usando la regla de probabilidad condicional y el hecho de que los tiempos de falla se asumen independientes, se establece la siguiente relación al considerar el límite cuando $\triangle t \to 0$

$$\frac{\mathrm{P(falla\ del\ individuo}\ i\ \mathrm{in}\ [\mathrm{t_{(j)}}, \mathrm{t_{(j)}} + \triangle \mathrm{t}))/\triangle \mathrm{t}}{\sum_{\mathrm{k \in R(t_{(j)})}} \mathrm{P(falla\ del\ individuo}\ k\ \mathrm{en}\ [\mathrm{t_{(j)}}, \mathrm{t_{(j)}} + \triangle \mathrm{t}))/\triangle \mathrm{t}}$$

Entonces si el individuo i tiene vector de covariables $x_{(j)}$, la expresión anterior puede ser re-escrita como

$$\frac{h(t_{(j)}|x_{(j)})}{\sum_{k \in R(t_{(j)})} h(t_{(j)}|x_{(j)})} = \frac{exp(\beta^T x_{(j)})}{\sum_{k \in R(t_{(j)})} exp(\beta^T x_{(k)})} \tag{3.14}$$

y usando la definición en (3.3) con $h(t_{(j)}|x_{(j)}) = h_0(t_j)\ \exp(\beta^T x_{(j)})$, se puede obtener la verosimilitud para las fallas en r tiempos de la siguiente manera

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta^T x_{(j)})}{\sum_{k \in R(t_{(j)})} exp(\beta^T x_{(k)})} \tag{3.15}$$

El efecto de las covariables en los tiempos de supervivencia se modela a través de los parámetros β, los cuales son estimados en el modelo de riesgo base (3.13). Existen otros métodos para estimar la forma del riesgo base y la función de riesgo acumulada, ya que el modelo PH sólo está definido

para estimar fallas instantáneas en el tiempo t y no detecta la presencia de datos empatados en la función de verosimilitud parcial (3.15).
Una alternativa al asumir riesgos proporcionales es considerar que los efectos de las covariables aceleran o detienen la falla en el tiempo. Un modelo de vida acelerado representa el logaritmo del tiempo de supervivencia como una combinación lineal de covariables, es decir

$$\log(T) = \beta^T x \tag{3.16}$$

De esta forma las covariables aceleran o desaceleran el tiempo de falla, en contraste al enfoque PH que asume un efecto multiplicativo de las covariables en la función de riesgo base independiente del tiempo.
En los modelos semi-paramétricos se asume que todas las covariables son fijas en el tiempo, por lo tanto, si las covariables son dependientes del tiempo, se tendrá otra interpretación. En un modelo PH, el tiempo sólo aparece en la función de riesgo base y no aparece relacionado con las covariables.

3.2.3. Métodos paramétricos

El modelo de riesgos proporcionales de Cox es una poderosa herramienta en el análisis de datos de supervivencia, pues no requiere asumir ninguna forma paramétrica en el riesgo base para estimar el efecto de las covariables en el tiempo de supervivencia. Puede haber situaciones en las que suponer que la distribución de supervivencia tiene alguna especificación paramétrica no sea razonable. En estos casos hay varias distribuciones que suelen usarse y que en breve presentaremos.
Existen ventajas adicionales al usar modelos paramétricos de supervivencia, sobre todo cuando se tratan de predecir los tiempos futuros de supervivencia. En este caso, el enfoque de riesgos proporcionales de Cox sólo puede estimar la forma del riesgo base hasta el momento de falla más

reciente y por tanto la estimación de predicciones no puede obtenerse.
Estos modelos dependen de un conjunto de parámetros que determinan completamente la forma distributiva que rige el tiempo de supervivencia. Sus estimaciones en cualquier punto del tiempo pueden usarse para predecir el riesgo futuro de falla en otros puntos del tiempo. Otra ventaja de este tipo de modelos, es que conservan la estructura de riesgos proporcionales o de vida acelerada descritos en la sección (3.2.2).
A continuación describiremos algunos de los modelos de supervivencia más comunes para poblaciones continuas y homogéneas.

Modelo Exponencial : Si la función de riesgo $h(t) = \lambda$ donde λ es una constante positiva, entonces los tiempos de supervivencia siguen una distribución exponencial. En este caso, la función de supervivencia esta dada por $\mathrm{S(t)} = \exp(-\lambda \mathrm{t})$ y la función de densidad por $\mathrm{f(t)} = \lambda \exp(-\lambda \mathrm{t})$.

Modelo Weibull : Este modelo tiene una función de riesgo monótona de la forma $h(t) = \alpha \lambda t^{\alpha-1}$ donde λ y α son parámetros positivos. La función de supervivencia es $\mathrm{S(t)}= \exp(-\lambda t^{\alpha})$ y función de densidad $\mathrm{f(t)} = \alpha \lambda \mathrm{t}^{\alpha-1} \exp(-\lambda \mathrm{t}^{\alpha})$. La distribución exponencial es un caso particular del modelo Weibull cuando el parámetro de forma $\alpha = 1$. Como la distribución Weibull depende de dos parámetros es muy flexible y las funciones de riesgo y densidad pueden tomar una variedad de formas diferentes. La inclusión de covariables a través de la función *log* de vinculo en el parámetro de escala λ resulta tanto en un modelo de riesgos proporcionales como en una estructura de vida acelerada. De hecho la distribución Weibull es la única con esta propiedad.

Tanto el modelo Exponencial como el de Weibull tienen formas cerradas de las funciones de riesgo y supervivencia y no son difíciles de trabajar. También hay otras distribuciones que pueden ser usadas, como: Gamma,

log-Normal, log-Logística, Gamma Generalizada, F Generalizada y las distribuciones de valores extremo.
Kalbfleisch y Prentice (2002) establecen que cualquier distribución de supervivencia continua puede ser discretizada al considerar a la variable aleatoria T discreta de la siguiente forma

$$P(T = t) = P(t \leq U < t + 1) \tag{3.17}$$

donde U es una variable aleatoria continua con distribución conocida. Por ejemplo, si U tiene distribución Weibull con parámetro α y λ, entonces

$$\begin{aligned} P(T = t) &= P(t \leq U < t + 1) \\ &= P(U < t + 1) - P(U < t) \\ &= F(t + 1) - F(t) \\ &= S(t) - S(t + 1) \\ &= \exp(-\lambda \mathrm{t}^{\alpha}) - \exp(-\lambda(\mathrm{t} + 1)^{\alpha}) \end{aligned} \tag{3.18}$$

Nótese que en (3.18) se está discretizando sobre períodos de tiempo de longitud 1, pero puede cambiarse de ser necesario.
Asumiendo censura aleatoria para n observaciones individuales, la verosimilitud toma la forma

$$L(\theta) = \prod_{i=1}^{n} [f(t_i|x_i, \theta)]^{\delta_i} [S(t_i|x_i, \theta)]^{1-\delta_i}, \tag{3.19}$$

donde, δ_i, $i = 1, \ldots, n$ es una variable binaria que toma el valor de 1 si el individuo i falla o 0 si hay censura por la derecha (existen formulaciones alternativas para censuras diferentes). De esta forma, las observaciones censuradas por la derecha contribuyen $P(T \geq t)$ a la verosimilitud, es decir, se conocen los tiempos de quienes sobrevivieron hasta el período $[0, t)$.

3.2.4. Modelo de Cox con covariables dependientes del tiempo

Una covariable *dependiente del tiempo* se define como una variable cuyos valores para un individuo dado pueden variar con el tiempo. Esto está en contraposición con el modelo de Cox de riesgos proporcionales, en el que todas las covariables son independientes del tiempo: permanecen constantes para cada individuo a lo largo de todo el estudio e influyen siempre igual en la curva de supervivencia. Existen básicamente tres tipos de variables dependientes del tiempo, todas ellas pueden incorporarse a la regresión de Cox construyendo un modelo que tiene en consideración riesgos no proporcionales. Los tres tipos son: *definidas por el usuario*; *internas* y *externas*.

Las variables definidas por el usuario suelen ser el producto de una función del tiempo por una covariable independiente del tiempo: $g(t) \times C$. Este tipo de variables son las más habituales y se emplean cuando se sospecha que la hipótesis de riesgos proporcionales no se cumple.

Las variables internas son variables dependientes del tiempo cuyo cambio a través del tiempo depende del individuo concreto. Por ejemplo, el tabaquismo en el tiempo t, el índice de obesidad en el tiempo t, o una situación de transplante (ha sido o no transplantado) en el tiempo t. Las variables externas son variables que afectan por igual a todos los individuos del estudio. Por ejemplo, el nivel de polución atmosférica en el tiempo t. Estas variables son las menos utilizadas en análisis de supervivencia. A los modelos presentados hasta ahora se les puede incorporar covariables dependientes del tiempo, aunque se debe tener cuidado en su interpretación.

Consideremos una covariable $x_i(t)$ para el individuo i que varía con el tiempo. Sea $X_i(t) = \{x_i(u); 0 \leq u \leq t\}$ la historia de la covariable hasta el momento t. De esta forma, la función de riesgo para el individuo i en el

tiempo t dependerá de la historia de las covariables en el instante t y el riesgo se puede definir de la siguiente forma

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P\{t \leq T < t + \Delta t | T \geq t, X_i(t)\}}{\Delta t} \tag{3.20}$$

Kalbfleisch y Prentice (2002) también señalan dos tipos de modelos para covariables dependientes del tiempo, modelos para variables internas y externas.

3.3. Modelización basada en un enfoque paramétrico y semiparamétrico

El análisis de supervivencia está ampliamente documentado no sólo en la literatura estadística, sino en campos como: ingeniería, ciencias sociales y epidemiología. Con la intención de encontrar una metodología capaz de abordar problemas relacionados con datos espaciales asociados a una red fija de localizaciones desde el contexto de supervivencia, se propone combinar los trabajos desarrollados por autores como: Cox (1972), Kalbfleisch y Prentice (2002), Bastos y Gamerman (2006), Henderson et al. (2002), Carlin y Banerjee (2002) basados en modelos de supervivencia desde el paradigma Bayesiano. En campos como la Agricultura por ejemplo, hay pocos trabajos que combinen esta metodología y no existen precedentes de modelos que hayan incorporado estructura espacial en forma de *frailty* en el caso de datos espaciales ubicados en un red de localizaciones.

El enfoque Bayesiano tiene un número importante de propiedades, pues permite no sólo la estimación a través de las distribuciones posteriores de los parámetros de interés, sino la estimación de la distribución predictiva posterior para los tiempos de supervivencia. Además, proporciona un método manejable para el ajuste de modelos complejos, en particular, es de especial interés, la incorporación en la estructura jerárquica de

efectos aleatorios (*frailties*) que intentan dar cuenta de la variabilidad no observada en los datos de supervivencia y no explicada por covariables.
El efecto espacial en el enfoque Bayesiano es más sencillo de manejar, ya que bajo este paradigma todos los parámetros se consideran aleatorios (Gilk et. al 1996). Tradicionalmente en análisis de supervivencia los efectos espaciales son conocidos como *frailties* y en nuestra propuesta los *frailties* se incluirán como una capa en el modelo jerárquico. Las distintas modelizaciones que proponemos en esta sección intentan encontrar la mejor forma de representar datos espaciales asociados a una red fija de localizaciones como datos de supervivencia.
En general, el objetivo principal del capítulo, es formular un modelo de supervivencia equipado con efectos aleatorios y covariables dependientes del tiempo que permita entender la evolución espacial y temporal de las probabilidades de supervivencia en cierto grupo de individuos. Otro objetivo también de mucha importancia, es emplear las estimaciones de las distribuciones posteriores para predecir el comportamiento futuro de enfermedades.

3.3.1. Modelo Weibull con tiempos discretos

Si partimos de la propuesta de Henderson y Shimakura (2003) podemos considerar a Z_i como una variable aleatoria para el i-ésimo individuo, donde Z_i tiene distribución positiva con media 1 y varianza τ. Además si consideramos un modelo con riesgos proporcionales como el definido en (3.13) con $\psi(.) = \exp(\beta^T x_i)$, entonces para el i-ésimo individuo con falla observada en t_i o censura en t_i y vector de covariable x_i, un efecto *frailty* z_i puede ser incluido de la siguiente forma

$$h(t_i|x_i) = z_i h_0(t_i)\exp(\beta^T x_i) \tag{3.21}$$

Si hay heterogeneidad individual entonces (3.21) se reduce al modelo

de riesgos proporcionales (3.13), por otro lado, si $z_i > 1$ los individuos experimentan una tasa de fracaso más rápida en t y si $z_i < 1$ se tiene una tasa de riesgo menor en cada t.

La estructura espacial también puede ser incorporada en la especificación del *frailty*, pero se considera el vector *frailty* transformado $\phi = log(Z)$, donde ϕ tiene una distribución normal multivariada y matriz de correlación dada por la estructura espacial. Otra forma de especificar ϕ es usando una distribución condicional autoregresiva (ICAR). En este caso, la media de la respuesta para un individuo esta condicionada a la media de sus vecinos (Besag y Kooperberg, 1995; Besag et al. 1991).

Con las ideas presentadas hasta el momento, formularemos un modelo de supervivencia con estructura jerárquica espacial estimado a partir de métodos MCMC. Para esto consideramos los datos como una discretización de un proceso continuo, aunque la enfermedad pueda ocurrir en cualquier momento. Además, la trayectoria futura de la enfermedad en cada instante dependerá de su historia, y será la covariable la responsable de recoger la información del pasado.

Especificamos un modelo inicial a través de la función de riesgo usando una distribución Weibull. Se selecciona en especial esta forma paramétrica por su capacidad de predecir tiempos futuros y por su flexibilidad al poseer dos parámetros, de forma y de escala. El modelo es análogo al modelo discreto presentado en (3.18) y discutido en la sección (3.2.3), con discretización sobre períodos de un año aunque la formulación utilizada aquí es ligeramente diferente.

Consideremos inicialmente una variable aleatoria continua U representativa del tiempo de supervivencia, donde $U > 0$ y sigue una distribución Weibull con función de riesgo dada por

$$h(u) = \rho\lambda u^{\rho-1} \tag{3.22}$$

y función de supervivencia

$$S(u) = \exp(-\lambda u^{\rho}) \tag{3.23}$$

ρ y λ son parámetros positivos de forma y de escala, respectivamente. Consideremos ahora, una variable aleatoria discreta T, representativa del tiempo en años de supervivencia, donde $T = 1, 2, \ldots$. Si relacionamos a las variables aleatorias U y T, se puede entonces definir el riesgo en el tiempo discreto t como la probabilidad de que U este en el intervalo $(t-1, t]$ dada la supervivencia en $t-1$, es decir,

$$\begin{aligned} h(t) = P(T = t | T > t-1) &= P(t-1 \leq U < t | U \geq t-1) \\ &= \frac{P(t-1 \leq U < t)}{P(U \geq t-1)} \\ &= \frac{P(U < t) - P(U < t-1)}{P(U \geq t-1)} \\ &= \frac{P(U \geq t-1) - P(U \geq t)}{P(U \geq t-1)} \\ &= \frac{S(t-1) - S(t)}{S(t-1)} \\ &= 1 - \frac{S(t)}{S(t-1)} \\ &= 1 - \exp(-\lambda[t^{\rho} - (t-1)^{\rho}]) \end{aligned} \tag{3.24}$$

limitada a la región $(0, 1)$. De las expresiones (3.8) y (3.9) se obtiene la función de supervivencia siguiente

$$\begin{aligned} S(t) = P(T \geq t) &= \prod_{0}^{t} (1 - h(j)) \\ &= \prod_{j=1}^{t} \exp(-\lambda[j^{\rho} - (j-1)^{\rho}]) \\ &= \exp\left(-\sum_{j=1}^{t} \lambda[j^{\rho} - (j-1)^{\rho}] \right), \quad t = 1,2,\ldots, \end{aligned} \tag{3.25}$$

tal que la función de probabilidad viene dada por $P(T = t) = h(t)S(t-1)$, es decir

$$P(T=t) = \begin{cases} 1 - \exp(-\lambda), & t = 1 \\ [1 - \exp(-\lambda[t^\rho - (t-1)^\rho])] \times & \\ \exp\left(-\sum_{j=1}^{t-1} \lambda[j^\rho - (j-1)^\rho]\right), & t = 2,3,\ldots \end{cases} \tag{3.26}$$

Por lo tanto $S(t)$ es una función decreciente acotada por encima de 1 y por debajo de cero para λ y ρ fijos; $S(t) \to 0$ cuando $t \to \infty$. Igualmente es una función de probabilidad limitada en el intervalo [0,1].

Si las covariables son incluidas a través del vinculo *log* en el parámetro de escala λ y se consideran fijas (independientes del tiempo), entonces (3.26) es idéntica al modelo discreto derivado en la sección (3.2.3). Por el contrario, si las covariables dependen de la historia en puntos anteriores del tiempo, la función de riesgo estará condicionada. Así, para una covariable dependiente del tiempo t, el modelo (3.24) se convierte en

$$h(t) = 1 - \exp(-\lambda_{t-1}[t^\rho - (t-1)^\rho]) \tag{3.27}$$

y esto conduce a modificaciones directas en las funciones de supervivencia y probabilidad. Por lo tanto, la probabilidad condicional de falla dada la supervivencia es dependiente del valor asumido por la covariable en el instante de tiempo t; sin embargo, las funciones de supervivencia y probabilidad total contienen información sobre la historia completa de la covariable. Como ya se ha comentado, el conocimiento de las funciones de riesgo, supervivencia y probabilidad se determinan únicamente con la distribución de supervivencia.

Una ventaja importante en el análisis de supervivencia es que las observaciones censuradas pueden ser incorporadas fácilmente. Teniendo en cuenta este hecho, definimos la verosimilitud para n fallas (o censuras) en t_i con $i = 1, \ldots, n$ de la siguiente forma

$$L(.) = \prod_{i=1}^{n} [P(T = t_i)]^{\delta_i} [P(T \geq t_i)]^{1-\delta_i} \tag{3.28}$$

donde δ_i es una variable binaria con valor 1, si ocurre la falla (individuo enfermo) o 0 si la observación es censurada. Las covariables x pueden ser incluidas a través del parámetro de escala λ haciendo $\lambda = \exp(\beta^T x)$.
Como la covariable que consideramos recoge la historia de la enfermedad en los años estudiados, la verosimilitud (3.28) se modifica de la siguiente forma

$$\begin{aligned} L(.) = \prod_{i=1}^{n} & [\exp(-\lambda_i (t_i - 1)^{\rho}) - \exp(-\lambda_i t_i^{\rho})]^{\delta_i} \\ & \times [(-\lambda_i t_i)^{1-\delta_i}] \end{aligned} \tag{3.29}$$

donde, $\lambda_i = \exp(\beta_0 + \beta_1 x_i(t) + \phi_{it})$. $x_i(t)$ representa el número de individuos próximos al i-ésimo individuo en el tiempo t condicionado a aquellos individuos enfermos en $t-1$. Los *frailties* que representan el efecto aleatorio espacial vienen dados por un modelo autoregresivo de la siguiente forma

$$\phi_{it} \sim \text{ICAR}(\sigma_{\phi}^2) \tag{3.30}$$

De acuerdo a lo señalado en el capítulo 2, para la desviación típica que define la varianza de ϕ asignaremos la previa

$$\sigma_{\phi} \sim \text{Unif}(0.5, 1) \tag{3.31}$$

Para completar la formulación del modelo, se asignan las siguientes distribuciones previas al resto de parámetros involucrados en λ_i,

$$\beta_0 \sim \text{N(-1,1)} \tag{3.32}$$

$$\beta_1 \sim \text{N}(0,1) \tag{3.33}$$

$$\rho \sim \text{Gamma}(0.1,10) \tag{3.34}$$

La modelización propuesta en (3.29) basada en la discretización de los tiempos de supervivencia fue propuesta por McKinley (2007). Sin embargo, en su propuesta original no consideró la influencia de efectos aleatorios, es decir, no incluyó el *frailty* espacial, además la forma en que el autor hace la escogencia de las previas es diferente. En adelante a esta propuesta la llamaremos Weibull Discrete Time Model (WDTM).

3.3.2. Modelos de riesgos proporcionales basado en procesos de conteo con cambios en la función de riesgo base

En el análisis de datos como procesos de conteo, incluyendo datos de supervivencia, un enfoque comúnmente utilizado para modelar la función de riesgo base en el contexto de riesgos proporcionales, es asignar un proceso Gamma a la previa del riesgo (Spiegelhalter et al. 1996), a pesar de que algunos autores señalan que puede llevar a estimaciones insesgadas y engañosas (Mostafa y Ghorbal, 2011).

En este apartado se plantean dos modelos jerárquicos Bayesianos basados en procesos de conteo. El primero asigna procesos Gamma a la previa del riesgo base y el segundo introduce funciones poligonales para estimar este riesgo (Beamonte y Bermúdez, 2003). Los procesos de conteo permiten estudiar a los datos de supervivencia a través de la modelización de la intensidad. Diversos autores han contribuido en este campo, entre ellos destacan, Andersen y Gill (1982), Tsiatis (1981) y Naes (1982).

Supongamos que se tienen n individuos de estudio y para el individuo i con $i = 1, 2, \ldots, n$ se tiene que $I_i(t)$ es el proceso de intensidad dado por el

vector de covariables $X_i = (Z_{i1}, \ldots, Z_{in})$. Sea $Y_i(t)$ el indicador del riesgo, es decir, el conjunto de sujetos todavía en riesgo en el tiempo T_i (sujetos sanos y sin censura antes del tiempo t).
Por otro lado, se observa el proceso $N_i(t)$, que cuenta el número de fallas ocurridas en el intervalo [0,t], tal que será constante e igual a cero entre las fallas y será uno cada vez que ocurra una falla. Ya que el proceso estocástico $\{N_i(t), t \geq 0\}$ es de conteo, entonces satisface las siguientes condiciones:

1. $N_i(t) \geq 0$,
2. $N_i(t)$ toma valores enteros,
3. Si $s < t$, $N_i(t) - N_i(s)$ representa el número de fallas que ocurren en el intervalo [s,t].

Por lo tanto, la tasa de una nueva falla es entonces $I_i(t) = Y_i(t)\lambda(t|X_i)$. La intensidad puede ser caracterizada como la probabilidad de que el evento de interés ocurra en el intervalo pequeño [t, $t+dt$], dado que no ha ocurrido antes. Esto es aproximadamente,

$$dN_i(t) \approx \lambda(t|X_i)dt = \lambda_0(t)\exp[\beta^{'} X_i] = I_i(t) \tag{3.35}$$

donde $dN_i(t)$ es el incremento de $N_i(t)$ en el intervalo [t, $t+dt$], es decir, el número de fallas observadas en [t, $t+dt$]. Así, $I_i(t)$ es la intensidad multiplicativa, la cual puede ser representada por

$$I_i(t) = Y_i(t)\lambda(t|X_i) = Y_i(t)\lambda_0(t)\exp[\beta^T X_i], \tag{3.36}$$

donde la intensidad es un producto de un proceso observado y una función no observada. Luego el proceso de intensidad $N_i(t)$ bajo el modelo (3.13) viene dada por

$$I_i(t) = Y_i(t)\exp[\beta^T X_i]\Lambda_0(t), \tag{3.37}$$

donde $\Lambda_0(t)$ representa la probabilidad instantánea de que el sujeto en riesgo en el tiempo t tenga un evento en el próximo intervalo $[t,\ t+dt]$. Supongamos ahora, que los individuos fueron evaluados hasta enfermarse o hasta su censura. De esta forma, tenemos el conjunto de datos observados $D = \{N_i(t), Y_i(t), X_i(t); i = 1, 2, \ldots, n\}$ y además tenemos como parámetros desconocidos $\beta, \Lambda_0(t)$. Asumiendo censura *no-informativa*, la verosimilitud factorizada de los datos tiene la forma

$$L_i(D|\beta, \Lambda_i(t)) = \exp(-\int_{t\geq 0}^{t} I_i(t)dt) \prod_{t\geq 0} [I_i(t)]^{dN_i(t)} \quad i = 1,2,\ldots,n \quad (3.38)$$

La función de verosimilitud para D tiene la distribución conjunta dada por

$$L(D|\beta, \Lambda(t)) = \prod_{i}^{n} L_i(D|\beta, \Lambda_i(t)) \quad (3.39)$$

Lo que hemos presentado hasta el momento en esta sección, nos va a permitir formular dos nuevas modelizaciones, las cuales se detallan a continuación.

Proceso de conteo con función de riesgo base Gamma

Con esta propuesta deseamos incorporar estructura dinámica en el modelo de supervivencia de forma diferente a lo propuesto por Bastos y Gamerman (2006). Estos autores asignan estructura autoregresiva en el vector de coeficientes $\beta(t)$ para considerar cambios en la covariable a través del tiempo, nosotros en cambio, no partiremos de este supuesto. En nuestro caso, será la covariable quien por si misma recoja la evolución de la enfermedad en el pasado. Del mismo modo, consideramos dinamismo en la componente espacial ya que su estructura cambiará en función a los individuos enfermos en periodos anteriores, es decir, t-1. Así, estamos dando un carácter espacio-temporal y dinámico a la modelización propuesta.

Para mejorar la eficiencia en los cálculos de las posteriores, se implementa la modelización basada en datos aumentados y se considera que las variables independientes $dN_i(t)$ tendrán distribución Poisson. Al considerar este cambio en la modelización, se obtiene la siguiente estructura probabilística

$$dN_i(t) \sim \text{Poisson}(I_i(t)dt) \tag{3.40}$$

Así, $I_i(t)dt = Y_i(t)\exp(\beta_0 + \beta_1 x_i(t) + \phi_{it})d\Lambda_0(t)$. Donde $d\Lambda_0(t) = \lambda_0(t)dt$ será el incremento o salto en la integral de la función del riesgo base ocurrida durante el intervalo $[t,t+dt]$.

$$d\Lambda_0(t) \sim \text{Gamma}(Cd\Lambda_0^*(t), C) \tag{3.41}$$

La distribución conjugada para $d\Lambda_0(t)$ fue sugerida por Kalbfleisch (1978). En (3.41) $\Lambda_0^*(t)$ puede ser interpretada como una estimación previa de la función de riesgo desconocida. Mientras que C representa el grado de confianza en esta estimación; valores pequeños de C corresponden a creencias débiles en la asignación. En nuestra modelización suponemos que

$$d\Lambda_0^*(t) = rdt \tag{3.42}$$

donde, r es la tasa de fallo supuesta por unidad de tiempo y dt es el tamaño del intervalo de tiempo. Siguiendo el paradigma Bayesiano, nuestra modelización queda completamente formulada al asignar la distribución previa a ϕ_{it}, β_0 y β_1, de la siguiente forma

$$\begin{aligned} \phi_{it} &\sim \text{ICAR}(\sigma_\phi^2) \\ \sigma_\phi &\sim \text{Unif}(0.5, 3) \end{aligned} \tag{3.43}$$

$$\begin{aligned} \beta_0 &\sim \text{N}(-1, 1) \\ \beta_1 &\sim \text{N}(0, 100) \end{aligned} \tag{3.44}$$

Después de observar D, nuestro interés se centra en la distribución posterior $\mathrm{P}(\beta, \Lambda_0(t)|D)$. Aplicando el Teorema de Bayes $\mathrm{P}(\beta, \Lambda_0(t)|D) \propto \mathrm{P}(D, \beta, \Lambda_0(t))$. Por lo tanto, el modelo de probabilidades se puede expresar como la distribución posterior conjunta dada por

$$\mathrm{P}(\beta, \Lambda_0(t)|D) \propto L(D|\beta, \Lambda_0(t))\mathrm{P}(\beta)\mathrm{P}(\Lambda_0(t)) \tag{3.45}$$

El foco estará centrado en la estimación de la función del riesgo base $\Lambda_0(t)$, vista como un proceso en el tiempo. Como (3.45) no tiene forma cerrada, entonces las estimaciones se harán usando métodos MCMC y escribiendo el modelo en OpenBUGS. Como esta propuesta considera un proceso Gamma con incrementos independientes en el riesgo base, la denominaremos Cox Model with Gamma process in baseline hazard (CMGPH).

Proceso de conteo con función de riesgo base poligonal

Comenzamos considerando el siguiente modelo multiplicativo

$$I_i(t) = Y_i(t)\lambda_0(t)\exp[\beta_0 + \beta_1 x_i(t) + \phi_{it}] \tag{3.46}$$

donde, $Y_i(t)$ sigue siendo el indicador del riesgo, es decir, el conjunto de sujetos todavía en riesgo en el tiempo t (sujetos sanos y sin censura antes del tiempo t.)

Para definir a $\lambda_0(t)$ se parte del modelo aditivo propuesto por Beamonte y Bermúdez (2003). En este caso, $\lambda_0(t)$ se supone una función poligonal no negativa con vértices localizados en los intervalos de tiempo $a_0 = 0 < a_1 < \ldots < a_T < a_{T_{max}+1}$, donde el polígono toma los valores $\tau_0 = 0 < \tau_1 < \ldots < \tau_{T_{max}} < \tau_{T_{max}+1}$ y se hace constante después del tiempo $a_{T_{max}+1}$. De esta forma, $\lambda_0(t)$ puede ser redefinida como

$$\lambda_0(t) = \begin{cases} \tau_j + \frac{(\tau_{j+1}-\tau_j)(t-a_j)}{a_{j+1}-a_j} \ , si\ a_j \leq t \leq a_{j+1}, \ \ j = 1, \ldots, T_{max} \\ \tau_{T_{max}+1} \ , si \ \ t \geq T_{max}+1 \end{cases} \tag{3.47}$$

El proceso Gamma supuesto en la modelización anterior (Kalbfleisch, 1978) asume independencia en los incrementos acumulados, situación que no es realista en la mayoría de los ajustes aplicados y no permite la relación entre intervalos adyacentes. La previa para el vector τ se especifica como un proceso auto-regresivo de primer orden (Gamerman, 1991) quien lo propuso en un contexto similar. Esta idea permite suavizar el efecto de τ_j en el modelo. De esta forma, τ_j se define de la siguiente forma

$$\tau_{j+1} = \tau_j \exp(e_j) \ , \ j = 1, \ldots, T_{max} \tag{3.48}$$

Donde,

$$e_j \sim \mathrm{N}(0, 1) \tag{3.49}$$

Al parámetro e_j se asigna una previa informativa para evitar inconvenientes en las estimaciones de los parámetros de interés. Los valores para el polígono se inicia a partir de τ_1 cuya distribución asignada es la siguiente

$$\tau_1 \sim \mathrm{Gamma}(0.01, 0.01) \tag{3.50}$$

La formulación de esta nueva modelización queda completa, al establecer la siguiente estructura probabilística para el resto de parámetros a estimar. Es decir,

$$\begin{aligned} \phi_{it} &\sim \mathrm{ICAR}(\sigma_\phi^2) \\ \sigma_\phi &\sim \mathrm{Unif}(0, 3) \end{aligned} \tag{3.51}$$

Las variaciones consideradas en el *frailty* espacial en cada instante de tiempo t viene dada por la estructura de vecindad de los vecinos enfermos en tiempos $t-1$.

$$\begin{aligned} \beta_0 &\sim \mathrm{N}(-1, 3) \\ \beta_1 &\sim \mathrm{N}(0, 100) \end{aligned} \tag{3.52}$$

Después de observar a $D = \{N_i(t), Y_i(t), X_i(t); i = 1, 2, \ldots, n\}$, el modelo resultante tendrá entonces la forma definida en la ecuación (3.45). En adelante, a esta modelización la llamaremos, Cox Model with polygonal function in baseline hazard (CMPFH).

3.4. Ilustración con datos de una parcela agrícola

En la mayoría de estudios observacionales se considera siempre un número elevado de covariables, pretendiendo encontrar entre ellas la de mayor valor pronóstico. Sin embargo, la incorporación de muchas variables en un modelo estadístico suele traer inconvenientes y en algunos casos resulta imposible. El conjunto de datos analizado sólo nos da información de sí el árbol está o no enfermo con su localización en la parcela. En este sentido, sólo se define una única covariable $x_i(t)$, construida al calcular el número de vecinos infectados del árbol i en función a los enfermos en el año t-1 localizados a distancias menores o iguales a 10 metros. En este ejemplo, los periodos de tiempo para la covariable $x_i(t)$ están dados por la relación entre los enfermos de los años 94 dado 93, 95 dado 94, 96 dado 95, 97 dado 96 y 98 dado 97, por tanto, $t = 1, \ldots, 5$ y el número de árboles observados es 300, es decir, $i = 1, \ldots, 300$.

Los datos obtenidos de la parcela (sección 2.1) fueron pre-tratados antes de aplicar la metodología de supervivencia propuesta, ya que ahora la variable de interés es el tiempo de supervivencia del árbol i a la infección

con CTV y no la variable respuesta Bernoulli. Como estos individuos son experimentales y controlados, todos los individuos (árboles) se comienzan a observar en el mismo tiempo t_0, es decir, en el año 1993. A partir de este año, se les hace un seguimiento hasta que se produce el evento final, la infección con el virus de la tristeza, en cuyo caso se tiene una observación completa, o termina el seguimiento antes de que el árbol se enferme con el virus, en cuyo caso tendremos una observación censurada.

El período de seguimiento considerado fue de seis años, comenzando desde el año 93 hasta el año 98. La variable aleatoria T, denota el año en que el árbol enferma, siendo el tiempo de censura igual al tiempo máximo considerado dado por $T_{max} = 6$. La censura observada en este tipo de situaciones se conoce como censura aleatoria por la derecha y obedece a un mecanismo que no guarda relación con el estudio.

La tabla (3.1) muestra el conjunto de árboles en riesgo y estimaciones de las probabilidades de supervivencia para los sujetos estudiados usando el método no-paramétrico de Kaplan-Meier. La figura (3.1) muestra las estimaciones de los riegos usando el estimador de Kaplan-Meier y el método de Cox con covariable. En esta figura se observa que ambas estimaciones se comportan igualmente.

Usando la librería survival de R, se ajusta el modelo de riesgo proporcional de Cox, encontrando que no es significativo bajo los test Likelihood ratio, Wald test y Score (logrank). La covariable que recoge el número total de vecinos infectados para el i-árbol ubicados a distancias menores e iguales a 10 metros sólo resulta significativa (p-valor =0.0563) al 1 %. Es importante resaltar, que esta covariable no es la misma que consideramos en las modelizaciones propuestas, ya que aquí la covariable no recoge la historia del contagio solamente resume el número de árboles vecinos enfermos para cada árbol i.

T	En riesgo	Infectados	Prob. Superviv.	I. Conf. 95 %
1	300	1	0.997	(0.99, 1.0)
2	299	6	0.977	(0.96, 0.994)
3	293	23	0.90	(0.868, 0.935)
4	270	20	0.833	(0.794,0.878)
5	250	66	0.617	(0.564, 0.674)

Tabla 3.1: Estimaciones usando método Kaplan-Meier

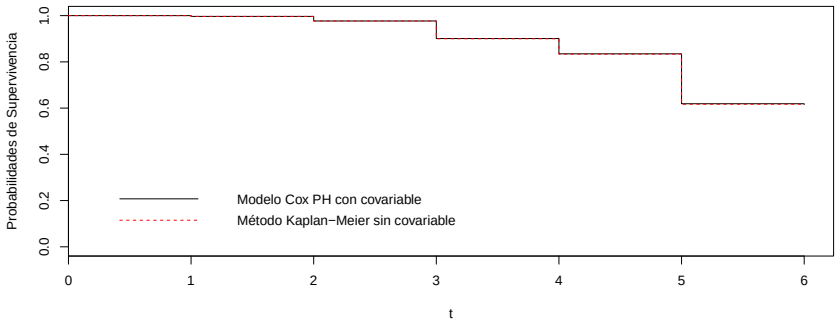


Figura 3.1: Gráfico de supervivencia usando método Kaplan-Meier y Cox PH

Con este tipo de análisis es imposible encontrar información que ayude en la toma de decisiones y permita llevar a cabo medidas preventivas para evitar nuevos casos. Se evidencia una vez más, que en el contexto de estudios epidemiológicos es necesario contar con metodologías que sean capaces de representar problemas complejos en donde existe la interacción de procesos espacio-temporales y que a la vez se conviertan en herramientas computacionalmente factibles de implementar. En este sentido, en las siguientes secciones, nos concentraremos en ilustrar con datos cada una

de las modelizaciones desarrolladas. Además, haremos una comparativa con los beneficios y dificultades de la puesta en marcha de cada una de ellas.

3.4.1. Aplicación del modelo Weibull con tiempos discretos

En la tabla (3.2) se presenta el resumen con la estimación de la deviance obtenida para las distintas modelizaciones desarrolladas en función a la teoría presentada en la sección (3.3.1). Se parte de un modelo básico al que denominamos Base, dado por $\lambda_i = exp(\beta_0)$. A partir de este modelo, se proponen otras modelizaciones producto de la incorporación de nuevas estructuras aleatorias.

Se observa como al incluir a la covariable $x_i(t)$ encargada de recoger la historia del proceso de contagio junto con el *frailty* espacial se consigue mejorar significativamente la bondad de ajuste del modelo Base sin incrementar notablemente el número de parámetros efectivos (p_D). Este resultado sugiere que existe una variabilidad oculta en la dinámica de la enfermedad y reconoce la presencia de factores desconocidos y no incorporados explícitamente en el modelo.

En trabajos anteriores, Li y Ryan (2001) indican que "existe muy poca literatura para modelar datos espacialmente correlacionados en supervivencia". Estos autores en el 2001, proponen una nueva clase de modelos semi-paramétricos de supervivencia extendidos a procesos espaciales correlacionados. Los datos espacialmente correlacionados existen ampliamente en la práctica, a pesar de la falta de métodos estadísticos diseñados específicamente para ellos. En estudios más recientes, Bastos y Gamerman (2006) proponen modelos dinámicos con variación espacial y no limitados a la condición de proporcionalidad ($h(t : X, s) = \exp\{X'\beta(t) + Z + W(s)\}$) y a partir de su propuesta es posible obtener otros modelos como casos especiales, entre ellos: Cox (1972) si $\beta(t) = \beta$, $\forall t$, $Z = 0$ y $W(s) = 0$, $\forall s$; el modelo frailty de Clayton (1978) obtenido si $\beta(t) = \beta$, $\forall t$

y $W(s) = 0$, $\forall s$; modelos de frailty espacial (Henderson et al. 2002; Carlin y Banerjee 2002) son obtenidos si $\beta(t) = \beta$, $\forall t$ y $Z = 0$; el modelo de Gamerman (1991) obtenido si $Z = 0$ y $W(s) = 0$, $\forall s$.

Es importante señalar, que nuestra propuesta de modelización puede verse como un caso particular de la modelización propuesta por Bastos y Gamerman (2006) al asumir que $\beta(t) = \beta$, $\forall t$, pero es diferente en la construcción de la estructura de correlación espacial. Ya que estos autores incorporan la dependencia espacial considerando datos geostadísticos, asumen una función de correlación isotrópica y calculan la distancia Euclidean entre las localizaciones de las observaciones. En cambio nuestra propuesta, esta diseñada para datos acomodados espacialmente en una red de localizaciones (lattice), además la estructura de correlación espacial que construimos considera el número de vecinos de cada árbol a distancias menores e iguales a los 10 metros y no considera el efecto de regiones vecinas (Banerjee y Carlin, 2002). Por tanto, la componente espacial es dinámica, ya que en cada instante de tiempo t, sólo se considera la estructura de vecindad determinada por los árboles sanos y enfermos en el tiempo actual t y no se considera la información espacial de árboles enfermos en los años t-1.

Bajo nuestra modelización, a diferencia de lo propuesto por Bastos y Gamerman (2006), los tiempos de supervivencia, variable aleatoria T, se modela con la distribución Weibull y se asumen tiempos discretos. La definición de *frailty* que hacemos al igual que los autores que han hecho uso de este recurso, se hace a nivel del riesgo (función hazard) y por tanto puede ser fácilmente recuperada después de la exponenciación, si es necesario. Al encontrar que el mejor modelo (menor DIC) es aquel que incorpora la variación espacial, hemos demostrado la importancia de esta componente en la modelización de enfermedades. Esta modelización puede ser una herramienta muy útil para describir la dinámica y evolución de enfermedades en cultivos.

En la tabla (3.3) se tiene que el riesgo base estimado para todos los árboles fue $\exp(-4{,}242) = 0{,}014$ y el riesgo relativo es igual a $\exp(0{,}046) = 1{,}047$. β_0 mantiene el signo negativo al evaluar el intervalo de credibilidad y sus estimaciones indican que existe un riesgo mínimo para cada árbol de aproximadamente 1 %. El coeficiente β_1 alcanza una media positiva y su intervalo de credibilidad aún cuando contiene el cero y asume valores cercanos al cero sugiere que el efecto de la covariable esta presente.

El parámetro ρ alcanza una media de 2.054, sugiriendo así que el riesgo aumentará con el paso del tiempo. La tabla (3.3) también evidencia que existe una variabilidad espacial importante en cada año y las marginales posteriores de σ_ϕ demuestran que la correlación espacial esta presente en el fenómeno estudiado.

	Base	Sólo covariable	Covariable + *Frailty* Espacial
		$\lambda_i = \mathrm{e}^{(\beta_0+\beta_1 x_i(t))}$	$\lambda_i = \mathrm{e}^{(\beta_0+\beta_1 x_i(t)+\phi_{it})}$
$\overline{D}$	772.2	772.7	760.0
p_D	1.882	2.891	6.378
DIC	774.1	775.6	766.4

Tabla 3.2: Resumen de la bondad de ajuste de los modelos bajo la propuesta WDTM

La figura (3.2) muestra el comportamiento general de la supervivencia para el conjunto de árboles analizados. La supervivencia decrece en forma lenta y progresiva con el transcurrir de los años al igual que se observa en la figura (3.1).

Parámetro	Media	D. típica	2.5 %	50 %	97.5 %
β_0	-4.242	0.3021	-4.838	-4.236	-3.684
β_1	0.046	0.1019	-0.16	0.053	0.228
ρ	2.054	0.16	1.74	2.052	2.394
$\sigma_\phi(t=1)$	0.3213	0.0717	0.1969	0.3184	0.4701
$\sigma_\phi(t=2)$	0.3227	0.0613	0.2161	0.317	0.448
$\sigma_\phi(t=3)$	0.3241	0.0598	0.2163	0.3248	0.4546
$\sigma_\phi(t=4)$	0.3172	0.0777	0.1922	0.3134	0.4747
$\sigma_\phi(t=5)$	0.3161	0.07551	0.1919	0.3156	0.469

Tabla 3.3: Resumen de las distribuciones posteriores para la modelización WDTM

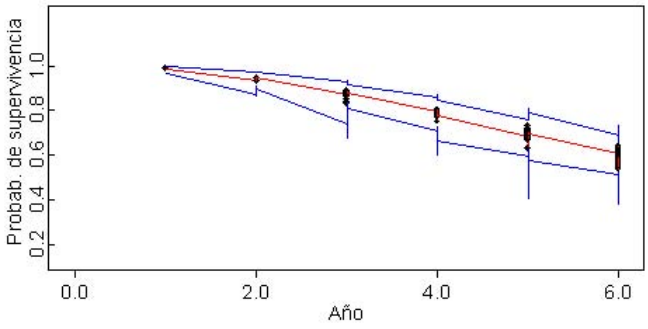


Figura 3.2: Probabilidades de supervivencia bajo la modelización WDTM; intervalo de credibilidad en color azul

3.4.2. Aplicación del modelo basado en procesos Gamma

Las estimaciones obtenidas bajo esta modelización se presentan en la tabla (3.4). También partimos de un modelo Base $(Y_i(t)\exp(\beta_0)d\Lambda_0(t))$, el cual comparamos, con dos modelos que van incorporando nuevas estructuras probabilísticas. Así, tendremos un segundo modelo compuesto sólo por el

efecto de la covariable y un tercero que combina la covariable con el *frailty* espacial.

Es importante recordar que bajo esta propuesta, la condición de proporcionalidad presente en el modelo Cox se *flexibiliza* al considerar que la covariable $x_i(t)$ es dependiente del tiempo. Esta modelización al igual que la anterior, es un caso particular de la propuesta de Bastos y Gamerman (2006) al hacer $\beta(t) = \beta$, $\forall t$, pero diferente en la configuración de la estructura espacial y en el efecto de la covariable. El *frailty* espacial y la covariable mantienen las características que señalamos en la sección anterior.

A diferencia de la propuesta WDTM, en este caso, partimos de un proceso de conteo observado, en donde el riesgo base se estima para cada instante de tiempo t a través de un proceso Gamma. Aunque este enfoque pueda parecer un poco artificial, Spiegelhalter et al. (1996) lo propone y sugiere que pueden incluirse covariables dependientes del tiempo. Estos autores, incorporan en su modelo efectos aleatorios (*frailty*) pero sin ninguna estructura espacial.

En la tabla (3.4) se observa una disminución significativa en el *DIC* al incluir la covariable. El *DIC* mejora mucho más al considerar el *frailty* espacial. Una vez más este resultado sugiere la importancia de modelar la variabilidad espacial en fenómenos de esta naturaleza. Este resultado sugiere además, que el riesgo de enfermar esta determinado tanto por un efecto del pasado (historia del contagio) y un efecto del presente, una relación entre árboles vecinos en el instante t.

	Base	Sólo covariable	Covariable + *Frailty* Espacial
		$Y_i(t)\mathrm{e}^{(\beta_0+\beta_1 x_i(t))}d\Lambda_0(t)$	$Y_i(t)\mathrm{e}^{(\beta_0+\beta_1 x_i(t)+\phi_{it})}d\Lambda_0(t)$
$\overline{D}$	808.4	749.0	715.6
p_D	2.803	6.336	17.5
DIC	811.4	755.3	733.1

Tabla 3.4: Resumen de la bondad de ajuste de los modelos bajo la propuesta CMGPH

Parámetro	Media	D. típica	2.5 %	50 %	97.5 %
β_0	-1.592	0.810	-2.81	-1.753	-0.228
β_1	0.4471	0.076	0.302	0.4516	0.5997
$d\Lambda_0(t=1)$	0.055	0.012	0.034	0.054	0.081
$d\Lambda_0(t=2)$	0.046	0.011	0.027	0.045	0.070
$d\Lambda_0(t=3)$	0.062	0.013	0.040	0.061	0.09)
$d\Lambda_0(t=4)$	0.055	0.012	0.033	0.054	0.082
$d\Lambda_0(t=5)$	0.040	0.010	0.022	0.039	0.063
$\sigma_\phi(t=1)$	0.554	0.145	0.319	0.5432	0.856
$\sigma_\phi(t=2)$	0.641	0.172	0.363	0.620	1.02
$\sigma_\phi(t=3)$	0.723	0.197	0.390	0.700	1.158
$\sigma_\phi(t=4)$	0.626	0.176	0.348	0.599	1.014
$\sigma_\phi(t=5)$	0.521	0.115	0.327	0.515	0.759

Tabla 3.5: Resumen de las distribuciones posteriores para la modelización CMGPH

En la tabla (3.5) se observa como β_1 alcanza un valor siempre positivo, con lo cual se confirma lo significativo de la covariable considerada. Además, se observa como el riesgo acumulado ($d\Lambda_0(t)$) alcanza valores de hasta el 6 %

en $t = 3$. La desviación típica marginal posterior para el *frailty* espacial en los primeros años resulta importante y disminuye en $t = 5$. Al analizar las estimaciones de las tablas (3.4) y (3.5) se evidencia que el modelo con el mejor ajuste manifiesta que los tiempos de supervivencia están determinados por un proceso progresivo de contagio y por una relación de vecindad entre árboles ubicados a distancias menores a los 10 metros.

La figura (3.3) muestra las probabilidades de supervivencia estimadas para cuatro árboles en $t = 5$. Se observa como la probabilidad de supervivencia disminuye con los años. Sin embargo, en el árbol 281 se observa un comportamiento distinto, esto sugiere, que cada árbol tiene un comportamiento diferente en la evolución de sus riesgos, lo que evidencia la existencia de un efecto heterogéneo no observado que el modelo es capaz de capturar.

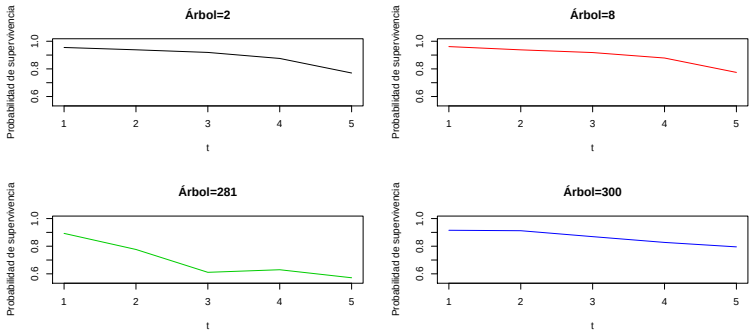


Figura 3.3: Algunas probabilidades de supervivencia bajo la modelización CMGPH

3.4.3. Aplicación del modelo basado en funciones poligonales

La diferencia fundamental de esta propuesta con la modelización previa, es que el tiempo de supervivencia será estimado usando en la función de

riesgo base un polígono con vértices definidos a partir de los intervalos de tiempo t observados (Beamonte y Bermúdez, 2003). Con esta modelización, queremos conocer si es posible obtener mejores estimaciones de los parámetros, ya que Mostafa y Ghorbal (2011) aseguran que asumir procesos Gamma independientes en el riesgo acumulado no es adecuado en la mayoría de las aplicaciones y además sostienen que en algunos casos se pueden obtener estimaciones insesgadas y engañosas. Con esta modelización se suaviza el efecto de la previa para τ al asignarle procesos de auto-correlación de primer orden (Gamerman, 1991).

	Base	Sólo covariable	Covariable + *Frailty* Espacial
		$Y_i(t)\mathrm{e}^{(\beta_0+\beta_1 x_i(t))}d\Lambda_0(t)$	$Y_i(t)\mathrm{e}^{(\beta_0+\beta_1 x_i(t)+\phi_{it})}d\Lambda_0(t)$
$\overline{D}$	808.2	747.5	716.9
p_D	1.475	2.486	17.34
DIC	809.7	750.0	734.3

Tabla 3.6: Resumen de la bondad de ajuste de los modelos bajo la propuesta CMPFH

Parámetro	Media	D. típica	2.5 %	50 %	97.5 %
β_0	-0.9168	0.5631	-2.041	-0.9184	0.2767
β_1	0.4542	0.077	0.302	0.456	0.598
$d\Lambda_0(t=1)$	0.0519	0.012	0.029	0.051	0.079
$d\Lambda_0(t=2)$	0.0419	0.010	0.023	0.041	0.065
$d\Lambda_0(t=3)$	0.0497	0.013	0.027	0.048	0.078
$d\Lambda_0(t=4)$	0.0503	0.012	0.028	0.049	0.077
$d\Lambda_0(t=5)$	0.0405	0.010	0.023	0.039	0.062
$\sigma_\phi(t=1)$	0.491	0.145	0.251	0.476	0.815
$\sigma_\phi(t=2)$	0.714	0.209	0.295	0.730	1.107
$\sigma_\phi(t=3)$	0.645	0.20	0.304	0.633	1.06)
$\sigma_\phi(t=4)$	0.557	0.175	0.26	0.542	0.947
$\sigma_\phi(t=5)$	0.463	0.123	0.242	0.455	0.734

Tabla 3.7: Resumen de las distribuciones posteriores para la modelización CMPFH

Los resultados que presentan las tablas (3.6) y (3.7) son similares a las estimaciones obtenidas bajo la modelización CMGPH, además β_1 sigue siendo positiva, lo que demuestra el carácter significativo de la covariable en el modelo. El modelo que presenta el mejor ajuste es el que contiene tanto la covariable como el *frailty* espacial. Al comparar las estimaciones de esta modelización con las obtenidas a partir de la asignación de procesos Gamma en el riesgo base, discrepamos de la opinión de Mostafa y Ghorbal (2011), pues los resultados que encontramos y su interpretación no difieren de lo hallado con la modelización CMGPH.

El modelo con menor *DIC* nos dice que el tiempo de supervivencia de cada árbol disminuye con el paso del tiempo, es decir se tendrá un mayor riesgo de enfermar. Si bien es cierto que la serie observada es pequeña y aún en ausencia de covariables propias de cada árbol, hemos encontrado que los

tiempos de supervivencia están completamente determinados por un efecto del pasado y por un efecto aleatorio espacial en el instante de tiempo actual t entre árboles ubicados a distancias menores a los 10 metros.
La figura (3.4) al igual que la figura (3.3) muestra la estimación de la supervivencia para el mismo grupo de árboles sanos. Se observa en ambas propuestas curvas de supervivencia similares para cada árbol.

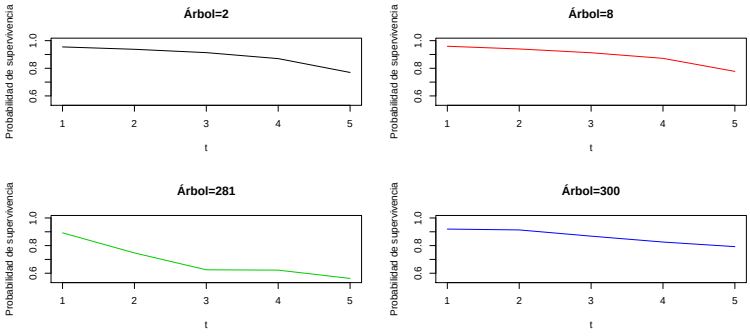


Figura 3.4: Algunas probabilidades de supervivencia bajo la modelización CMPFH

3.5. Comparativa

Después de obtener y evaluar las estimaciones de las tres modelizaciones y los requerimientos de convergencia y de implementación en cada caso, mencionaremos en qué situaciones resulta mejor aplicar una u otra propuesta. Comparando los tiempos de ejecución de los mejores modelos en cada caso, se obtuvo un menor tiempo con la modelización basada en procesos Gamma (CMGPH), con una duración de 31 segundos. Mientras que la modelización basada en funciones poligonales (CMPFH) obtuvo un tiempo de 1 minuto. Siendo la modelización basada en el modelo Weibull la de mayor tiempo, 2.17 minutos.
La convergencia del parámetro espacial σ_ϕ fue lenta, en especial bajo la

modelización WDTM. En las tres modelizaciones fue necesario asignar previas Uniformes con hiperparámetros distintos en la desviación típica que define la varianza del *frailty* espacial. En particular, en la modelización WDTM se usó una distribución Uniforme más informativa en comparación con las otras dos modelizaciones. Siendo la propuesta CMPFH quien necesitó una distribución menos informativa en σ_ϕ. Además, la convergencia de σ_ϕ fue más rápida en la modelización CMGPH en comparación con la modelización CMPFH.

En la primera modelización WDTM correspondiente al modelo completo, compuesto por covariable + *frailty* espacial se obtienen valores negativos en el número efectivo de parámetros, p_D, durante las primeras 3000 iteraciones y la deviance del modelo no se estabiliza hasta después de este número de iteraciones. Sin embargo, la convergencia de la deviance bajo el modelo completo definido en la modelización CMGPH muestra un mejor comportamiento y su convergencia es mucho más rápida que la del modelo CMPFH.

En cuanto a la implementación de cada modelización en lenguaje BUGS, resulta más sencilla la sintaxis y escritura de los modelos CMGPH y CMPFH. Para la modelización WDTM fue necesario utilizar el truco de los ceros propuesto por Spiegelhalter et al. (2003). Este mecanismo de ajuste consiste en generar n-variables latentes Y_i con distribución Poisson de media θ_i. Así Y_i contribuirá en $\exp(-\theta_i)$ a la verosimilitud. Al implementar este mecanismo, se tiene que corregir la contribución de la verosimilitud en el tiempo T_i haciendo $\theta_i = -\log(L_i)$. Luego, fue necesario escribir la función de verosimilitud ligeramente diferente a la definición dada en la sección (3.3.1) y poder así incorporar el truco de los ceros.

En las tres modelizaciones propuestas, se tiene que el modelo que considera el efecto de la covariable dependiente del tiempo con el *frailty* espacial alcanzan los mejores ajustes. A pesar de lo que señalan Mostafa y Ghorbal (2011) de obtener estimaciones engañosas al asumir procesos

Gamma en la previa del riesgo base, hemos encontrado que las propuestas basadas en procesos de conteo generan estimaciones equivalentes y de igual interpretación.

Una ventaja adicional de la modelización CMPFH se refiere al conjunto de vértices para el polígono, pues al no ser crítica su elección, se pueden obtener estimaciones consistentes y similares aún en presencia de otros conjuntos. Lo único que debe considerarse para la escogencia de tales vértices, es que deben ser valores cercanos a los tiempos observados.

En las tres modelizaciones propuestas, se tiene que la convergencia de la deviance en aquellos modelos que sólo consideran el efecto de la covariable fue más rápida que para los otros modelos ajustados. Además, la modelización basada en la distribución Weibull, WDTM, fue la más sensible a la elección de los hiperparámetros asignados a la previa de los parámetros ρ, α, β_1.

La convergencia fue comprobada usando el conjunto de test diagnósticos proporcionados por el paquete CODA y todas las simulaciones fueron realizadas utilizando el software OpenBUGS y el BUGS desde R. Las estimaciones de los parámetros se obtienen a partir de dos cadenas paralelas y después de quemar las 5000 primeras iteraciones. Para reducir la autocorrelación en las cadenas se toma 1 de cada 5 muestras hasta obtener 10000 iteraciones.

3.6. Conclusiones del capítulo

Aun cuando el conjunto de datos empleado para ilustrar la metodología propuesta están espacialmente asociados a una red fija de localizaciones, los resultados encontrados hablan de que es factible aún con datos de esta naturaleza, pensar en técnicas de supervivencia bajo el paradigma Bayesiano fundamentadas en modelos jerárquicos.

Partir del contexto de supervivencia requiere entonces comprender

estadísticamente el proceso, para ello, se necesitan modelos capaces de capturar la heterogeneidad usualmente no observada y que generalmente no es explicada en las covariables disponibles. Pensar que los individuos son extraídos de una población homogénea, no es adecuado, especialmente en fenómenos donde existen factores de riesgo ocultos que gracias a la cercanía entre los individuos son compartidos. Por lo tanto parece razonable, diseñar modelos jerárquicos que permitan tratar la heterogeneidad existente en la población en alguna de sus capas o niveles. De esta forma, aún cuando dos individuos tengan funciones de riesgo similares no serán necesariamente idénticos, a pesar de compartir el mismo vector de covariables.

La representación jerárquica de las modelizaciones vistas desde el paradigma Bayesiano, permite hacer frente a tal heterogeneidad y la convierten en una metodología de trabajo novedosa y de aplicabilidad en cualquier área científica. La metodología que desarrollamos a diferencia de lo que suele encontrarse en trabajos que utilizan el modelo de riesgo proporcionales de Cox, considera una covariable dependiente del tiempo y la influencia de un efecto aleatorio espacial conocido en análisis de supervivencia como *frailty* espacial. Esta última componente configura a la modelización de estructura capaz de capturar la heterogeneidad no observada en cada instante de tiempo t.

Los modelos de supervivencia vistos como procesos de conteo, flexibilizan la condición de proporcionalidad asumida comúnmente en los modelos de riesgos proporcionales de Cox. En todas las modelizaciones se incluye el *frailty* espacial en la función de riesgo (función hazard ó función de intensidad), el cual puede ser fácilmente conocido por medio de la exponenciación y recuperado cuando así se requiera. Para conferir dependencia espacial en los modelos se adoptan procesos autoregresivos ICAR en los *frailties* (Carlin y Banerjee, 2002). Esta elección se debe a su flexibilidad en el acomodo de la autocorrelación espacial y a su beneficio computacional.

A diferencia de lo propuesto por Bastos y Gamerman (2006), las modelizaciones desarrolladas están basadas en la noción de vecindad y no en modelos geoestadísticos. Esta perspectiva resulta más apropiada en el caso de datos agregados por áreas o datos espaciales ubicados en una red de localizaciones. Una vez analizados los resultados y considerados los beneficios de las modelizaciones propuestas, es posible comentar las siguientes observaciones finales:

- La modelización WDTM puede verse como una forma básica de modelar datos de supervivencia, cuando estos datos provengan de un proceso espacial dado en una red fija de localizaciones y en presencia de tiempos discretos. De las tres modelizaciones propuestas, este modelo es el que menos ventajas computacionales ofrece.

- Las estimaciones bajo las propuestas CMGPH y CMPFH son equivalentes entre sí. Además el modelo con mejor ajuste en ambos casos, es el que considera la covariable dependiente del tiempo y el *frailty* espacial. Es importante destacar, que bajo la modelización WDTM también resulta ser el mejor modelo aquel que considera a la covariable y el *frailty* espacial.

- En la modelización CMPFH, cuanto más larga sea la longitud de los intervalos, más se perderá la información. Por tanto, los vértices considerados para el polígono deben ser valores cercanos a los tiempos observados.

- En las tres modelizaciones se obtiene que los riegos están determinados por dos procesos bastante claros, el primero, recoge la evolución del contagio entre árboles infectados en años anteriores (efecto del pasado) y el segundo, recoge la variabilidad espacial en el instante de tiempo t (efecto del presente). Por tanto se espera, que un árbol con un número importante de vecinos infectados a distancias menores

a los 10 metros tenga mayor probabilidad de enfermar o menor probabilidad de supervivencia.

- La distribución autoregresiva ICAR permite introducir dependencia espacial y el intercambio de información a través de árboles vecinos ubicados a distancias menores a los 10 metros.

- A pesar que las curvas de supervivencia obtenidas a partir de las modelizaciones CMGPH y CMPFH muestran un comportamiento similar en los riesgos, estas dan cuenta y distinguen una heterogeneidad no observada explícitamente para el individuo i. Esto sugiere que ambos modelos capturan fuentes de variabilidad que determinan el tiempo de supervivencia de cada individuo.

- Bajo la modelización WDTM se tiene un comportamiento en la estimación de la curva de supervivencia similar al obtenido al asumir el método de Kaplan-Meier y Cox.

- Dentro del contexto Bayesiano, la mayoría de los modelos de supervivencia parten del supuesto de riegos proporcionales. Las modelizaciones basadas en procesos de conteo flexibilizan la condición de proporcionalidad al representar las fallas en los intervalos [t, t+dt]. Por lo tanto, estos modelos se convierten en métodos flexibles a la hora de abordar estudios de supervivencia que parten de datos espaciales observados de una red de localizaciones.

Capítulo 4

Modelización mediante procesos espaciales continuos

Este capítulo comienza presentando las bases teóricas sobre las que se fundamentan los modelos para datos geoestadísticos (Cressie, 1993), en este sentido, se define la estacionariedad, la isotropía, el variograma y sus elementos desde la perspectiva clásica. Del mismo modo se introducen los conceptos asociados con la predicción clásica y Bayesiana.

Existen diversas formas de representar la dependencia espacial, demostraremos como los campos aleatorios Gaussianos de Markov se convierten en la forma más conveniente de representar la dependencia entre los datos observados desde el contexto de modelos jerárquicos Bayesianos. Se explica cómo es posible realizar el *kriging* Bayesiano usando ecuaciones diferenciales parciales estocásticas.

En este capítulo se propone un tipo de modelo perteneciente a la clase de modelos mixtos lineales generalizados conocidos también como modelos latentes. El modelado propuesto involucra un campo Gaussiano afectado por un proceso espacial y representado como un campo aleatorio Gaussiano de Markov. El objetivo principal de este capítulo es presentar

una estrategia general de estimación y predicción efectiva para procesos espaciales continuos desde el paradigma Bayesiano que permita describir el comportamiento de fenómenos asociados con la presencia de cierto evento de interés.

Los modelos lineales mixtos generalizados gozan de una popularidad cada vez mayor debido a su capacidad para modelar observaciones correlacionadas. Su rango de aplicación puede ir más allá de los populares modelos lineales generalizados, pero esto implica cálculos más complejos y difíciles. Varios procedimientos de inferencia se han propuesto, entre ellos, el más popular desde el enfoque Bayesiano ha sido el análisis con métodos Monte Carlo de cadenas de Markov, sin embargo, recientemente Rue el at. (2009) ha introducido un novedoso método de inferencia numérico conocido como Integrated Nested Laplace Approximation (INLA) que junto al enfoque Stochastic Partial Diferential Equation (SPDE) propuesto por Lindgren et al. (2011) ha permitido construir una metodología potente para el análisis de modelos Gaussianos latentes complejos desde la perspectiva Bayesiana capaz de soportar los diferentes tipos de datos espaciales.

Lindgren et al. (2011) demuestra que a través de ecuaciones diferenciales parciales estocásticas (SPDE) encontradas a partir de la familia Matérn, es posible pasar de un campo Gaussiano a un campo aleatorio Gaussiano de Markov. La ventaja principal de esta representación esta en la mejora notable de los tiempos de cálculo y disminuye las dificultades numéricas asociadas con el análisis de modelos lineales mixtos generalizados, ya que los campos aleatorios Gaussianos de Markov están definidos sobre matrices dispersas y no sobre matrices densas.

En el campo referido a Epidemiología en Agricultura son pocos los trabajos que se encuentran dedicados a la cartografía de enfermedades en plantas, y los que existen están enfocados en el análisis frecuentista y no en el enfoque Bayesiano. La metodología basada en la cartografía de enfermedades se ha popularizado en los últimos años, especialmente gracias al trabajo

publicado por Besag et al. (1991), donde generalmente se asigna una distribución Poisson a la variable respuesta. Hasta ahora es poca la literatura enmarcada en la metodología INLA y en especial no hemos encontrado trabajos en Agricultura dedicados al contexto epidemiológico que hagan uso de ella. Por eso creemos interesante, ilustrar la metodología con datos obtenidos de un cultivo agrícola. Se demuestra además las bondades de la metodología en fenómenos con muchos y pocos datos. Y se ilustran los beneficios de emplear algunas estrategias de muestreo.

4.1. Introducción

La Geoestadística es un término acuñado en los años 50 para denominar a las técnicas estadísticas aplicadas al análisis geográfico. Su desarrollo, en esa década y en la siguiente, se debe a su aplicación a la ingeniería de minas, para predecir las reservas de mineral a partir de observaciones espacialmente distribuidas en una región.

Hay una gran variedad de problemas que pueden resolverse utilizando métodos geoestadísticos. La característica común a todos ellos es que los datos pueden verse como una realización, habitualmente parcial, de un proceso estocástico sobre una región espacial continua. Matheron (1963) denomina esta situación como un problema de variables regionalizadas enfatizando la naturaleza continua del conjunto de índices.

Los métodos geoestadísticos ofrecen una manera de describir la continuidad espacial, que es un rasgo distintivo esencial de muchos fenómenos naturales y proporciona adaptaciones de las técnicas clásicas de regresión para tomar ventajas de esta continuidad (Isaaks y Srivastava, 1989).

Cuando el objetivo es hacer predicción, la geoestadística opera básicamente en dos etapas. La primera es el análisis estructural, en el cual se describe la correlación entre puntos en el espacio. En la segunda fase se hace predicción en sitios de la región no muestreados por medio de la técnica *kriging*.

Un proceso estocástico es una colección de variables aleatorias indexadas; esto es, para cada s en el conjunto de índices D, $Y(s)$ es una variable aleatoria. En el caso de que las mediciones sean hechas en una superficie, entonces $Y(s)$ puede interpretarse como la variable aleatoria asociada a ese punto del plano (s representa las coordenadas, planas o geográficas, y Y la variable en cada una de ellas).
La formulación básica de un proceso estocástico se concreta a la situación espacial en la que se toma como conjunto de índices una determinada región continua D del espacio:

$$\{Y(s) : s \in D\} \tag{4.1}$$

donde D es un conjunto fijo en el espacio euclidiano d-dimensional. En el contexto espacial, usualmente se tiene a $d = 2$ ó $d = 3$. En situaciones donde $d > 1$, entonces el proceso será referido como un proceso espacial.

4.1.1. Estacionariedad

La predicción es posible si el proceso tiene, en algún sentido, un comportamiento estable en toda la región de estudio. En adelante asumiremos que nuestro proceso espacial tiene una media, $\mu(s) = E(Y(s))$ y que la varianza de *Y(s)* existe para todo $s \in D$.
El proceso *Y(s)* es Gaussiano, si para cualquier $n \geq 1$ y para cualquier conjunto de sitios $\{s_1, \ldots, s_n\}$, $Y = (Y(s_1), \ldots, Y(s_n))^T$ tiene una distribución normal multivariante. El proceso se dice que es estrictamente estacionario, si para cualquier $n \geq 1$, para cualquier conjunto de sitios $\{s_1, \ldots, s_n\}$ y para cualquier $\mathbf{h} \in \Re^d$, la distribución de $Y = (Y(s_1), \ldots, Y(s_n))$ es la misma que $Y = (Y(s_1 + h), \ldots, Y(s_n + h))$.
La estacionariedad estricta es una condición muy fuerte y poco habitual, pues establece que las distribuciones de probabilidad conjunta permanezcan invariantes ante una traslación. Esta condición se escribe como:

$$F_{s_1+h,\ldots,s_m+h}(y_1,\ldots,y_m) \equiv F_{s_1,\ldots,s_m}(y_1,\ldots,y_m) \tag{4.2}$$

La condición menos exigente es la *estacionariedad de segundo orden*, o *estacionariedad débil*, que implica que la esperanza sea constante y que la función de covarianza sea invariante por traslación. Esto es,

$$E(Y(s)) = \mu, \qquad \forall s \in D \tag{4.3}$$

$$Cov(Y(s_1), Y(s_2)) = C(s_1 - s_2), \quad \forall s_1, s_2 \in D \tag{4.4}$$

De esta forma, la función de covarianza de un proceso estacionario se puede expresar en función del vector de diferencia entre los puntos. A la función *C(.)* se le denomina *covariograma*.

Es claro que si una variable regionalizada es estrictamente estacionaria entonces también será estacionaria débil. El concepto de estacionariedad es muy útil en la modelación de series temporales (Box y Jenkins, 1976). En este contexto es fácil la identificación, puesto que sólo hay una dirección de variación (el tiempo). En el campo espacial existen múltiples direcciones y por lo tanto se debe asumir que en todas el fenómeno es estacionario. Cuando la esperanza de la variable no es la misma en todas las direcciones o cuando la covarianza o correlación dependan del sentido en que se determinan, no habrá estacionariedad.

Si la correlación entre los datos no depende de la dirección en la que esta se calcule se dice que el fenómeno es isotrópico, en caso contrario se hablará de anisotropía. En Isaaks y Srivastava (1989) se definen los posibles tipos de anisotropía y se proponen algunas soluciones. Cressie (1993) discute cual debe ser el tratamiento en caso de que la media no sea constante.

En casos prácticos resulta compleja la identificación de la estacionariedad. Suelen emplearse gráficos de dispersión de la variable respecto a las coordenadas, de medias móviles y de valores clasificados según puntos de

referencia, con el propósito de identificar posibles tendencias de la variable en la región de estudio.
Una perspectiva diferente de la estacionariedad se obtiene al estudiar la variabilidad de los incrementos del proceso, ya que existen algunos fenómenos físicos reales en los que la varianza no es finita. En estos casos se trabaja sólo con la hipótesis en que $[Y(s+h)-Y(s)]$ sean estacionarios (Clark, 1979), esto es

$$E[Y(s+h)-Y(s)]=0,$$
$$Var(Y(s_1)-Y(s_2))=2\gamma(s_1-s_2)=2\gamma(h)=2[C(0)-C(h)],\forall s_1,s_2\in D \tag{4.5}$$

Esta hipótesis se verifica si la varianza de las diferencias entre las variables en dos puntos depende únicamente del vector que los separa. A esta propiedad se denomina *estacionariedad intrínseca* y es una condición más débil que la estacionariedad de segundo orden y es la que se emplea habitualmente en la modelización geoestadística.

4.1.2. Funciones de correlación espacial

La primera etapa en el desarrollo de un análisis geoestadístico es la determinación de la dependencia espacial entre los datos medidos de una variable. Esta fase es también conocida como análisis estructural. Para llevarla a cabo, con base en la información muestral, se usan tres funciones: El semivariograma, el covariograma y el correlograma. A continuación se hace una revisión rápida de los conceptos asociados a cada una de ellas y se describen sus bondades y limitaciones.

- Variograma y semivariograma:
  Cuando se definió la estacionariedad débil se mencionó que se asumía que la varianza de los incrementos de la variable regionalizada era

finita. A esta función denotada por $2\gamma(h)$ se le denomina variograma. Utilizando la definición teórica de la varianza en términos del valor esperado de una variable aleatoria, tenemos:

$$2\gamma(h) = E((Y(s+h) - Y(s))^2) \tag{4.6}$$

La mitad del variograma $\gamma(h)$ se conoce como la función de semivarianza y caracteriza las propiedades de dependencia espacial del proceso. Dada una realización del fenómeno, la función de semivarianza es estimada, por el método de momentos, a través del semivariograma experimental, que se calcula mediante (Wackernagel, 1995):

$$\overline{\gamma}(h) = \frac{\sum(Y(s+h) - Y(s))^2}{2n} \tag{4.7}$$

donde $Y(s)$ es la variable medida en el sitio s, $Y(s+h)$ es otro valor muestral separado del anterior por una distancia h y n es el número de parejas que se encuentran separadas por dicha distancia. La función de semivarianza se calcula para varias distancia h. En la práctica, debido a irregularidades en el muestreo y por ende en las distancias entre los sitios, se toman intervalos de distancia $\{[0,h],(h,2h],(2h,3h],\ldots\}$ y el semivariograma experimental corresponde a una distancia promedio entre parejas de sitios dentro de cada intervalo y no a una distancia h específica.

Varios elementos aparecen diferenciados en el semivariograma: la pepita, el alféizar y el rango.

- Se denomina *efecto pepita*, al término extraído de la aplicación a la minería, a la situación en que el variograma no tiende a 0 al acercarse al origen. Esto puede ser debido a un error de medida o a la variación a muy pequeña escala. En algunas ocasiones

puede ser indicativo de que parte de la estructura espacial se concentra a distancias inferiores a las observadas.

$$\lim_{h \to 0} \gamma(h) = c_0 > 0$$

- De forma lógica, un semivariograma crece con la distancia, recogiendo el fenómeno de que el proceso es similar en puntos próximos, hasta que se estabiliza en un valor llamado *alféizar* que expresa la variabilidad entre puntos distantes. El alféizar puede ser o no finito. Los semivariogramas que tienen alféizar finito cumplen con la hipótesis de estacionariedad estricta; mientras que cuando ocurre lo contrario, el semivariograma define un fenómeno natural que cumple sólo con la hipótesis de estacionariedad intrínseca.

$$\lim_{h \to \infty} \gamma(h) = c_s > 0$$

- El rango es la distancia h_s a la que se alcanza el alféizar, $\gamma(h) = c_s, \forall h > h_s$. En términos prácticos corresponde a la distancia a partir de la cual dos observaciones son independientes. El rango se interpreta como la zona de influencia. Entre más pequeño sea el rango, más cerca se esta del modelo de independencia espacial.

Para interpretar el semivariograma experimental se parte del criterio de que a menor distancia entre los sitios mayor similitud o correlación espacial habrá entre las observaciones. Por ello, en presencia de autocorrelación se espera que para valores de h pequeños el semivariograma experimental tenga magnitudes menores a las que este toma cuando las distancias h se incrementan.

- Covariograma y correlograma:
  La función de covarianza muestral entre parejas de observaciones que

se encuentran a una distancia h se calcula, empleando la formula clásica de la covarianza muestral, por:

$$C(h) = Cov(Y(s+h) - Y(s)) = \frac{\sum_{i=1}^{n}(Y(s+h)Y(s)) - m^2}{n} = C(h)$$

donde m representa el valor promedio en todo punto de la región de estudio y n es el número de parejas de puntos que se encuentran a una distancia h. Asumiendo que el fenómeno es estacionario y estimando la varianza de la variable regionalizada a través de la varianza muestral, se tiene que el correlograma muestral está dado por:

$$r(h) = \frac{Cov(Y(s+h) - Y(s))}{S_{s+h} \cdot S_s} = \frac{C(h)}{S_s^2} = \frac{C(h)}{C(0)}$$

Bajo el supuesto de estacionariedad cualquiera de las tres funciones de dependencia espacial mencionadas, es decir semivariograma, covariograma o correlograma, puede ser usada en la determinación de la relación espacial entre los datos. Sin embargo como se puede observar en las fórmulas, la única que no requiere hacer estimación de parámetros es la función de semivarianza. Por esta razón, fundamentalmente, en la práctica se emplea el semivariograma y no las otras dos funciones.

4.1.3. Isotropía

Si el semivariograma $\gamma(h)$ depende del vector de separación sólo a través de su longitud $\|h\|$, entonces decimos que el proceso es isotrópico. Así para un proceso isotrópico, $\gamma(h)$, es una función de valor real de argumento univariado y se puede escribir como $\gamma(\|h\|)$. Si el proceso es intrínsicamente estacionario e isotrópico entonces el proceso es homogéneo.
Entre los muchos modelos isotrópicos de semivariograma que se han propuesto, los más empleados son el lineal, esférico, exponencial, cuadrático

racional, ondulado, potencial y Gaussiano (Banerjee et al. 2004). Éstos constituyen una amplia batería representativa de diferentes comportamientos de los procesos espaciales.
La isotropía es estudiada a través del cálculo de funciones de autocovarianza o de semivarianza muéstrales en varias direcciones. Si estas tienen formas considerablemente distintas puede no ser válido el supuesto de isotropía.

4.1.4. Función de covarianza

Con el fin de especificar un proceso estacionario se debe proporcionar una función de covarianza válida. Aquí "válida" significa que $c(h) \equiv \text{cov}(Y(s), Y(s+h))$ tal que para cualquier conjunto finito de sitios $s_1, \ldots, s_n$ y para cualesquiera $a_1, \ldots, a_n$,

$$Var[\sum_i a_i Y(s_i)] = \sum_{i,j} a_i a_j \text{Cov}(Y(s_i, Y(s_j))) = \sum_{i,j} a_i a_j c(s_i - s_j) \geq 0 \tag{4.8}$$

la cual es una desigualdad estricta si no todos los a_i son 0. Necesitamos que $c(h)$ sea una función definida positiva, verificar esta condición no es trivial, pero el Teorema de Bochner proporciona una condición suficiente y necesaria para que $c(h)$ lo sea. Este teorema es aplicado para $\mathbf{h}$ en el espacio d-dimensional euclidean.
En general, el Teorema de Bochner establece que $c(h)$ es definida positiva si y sólo si

$$c(h) = \int \cos(w^T h) G(dw), \tag{4.9}$$

donde G es acotada, positiva, simétrica alrededor de 0 medida en $\Re^d$. Entonces $c(0) = \int G d(w)$ se convierte en una constante normalizada y $\frac{G(dw)}{c(0)}$ es referida como la distribución espectral que induce a $c(h)$.

Por otro lado, si $G(dw)$ tiene una densidad con respecto a la medida de Lebesgue, es decir, $G(dw) = g(w)dw$, entonces $\frac{g(w)}{c(0)}$ es referida como la densidad espectral. Evidentemente (4.9) puede ser usada para generar funciones de covarianza válidas.

Ya que $e^{iW^Th} = \cos(w^Th)+i\sin(w^Th)$, tenemos que $c(h) = \int e^{iW^Th}G(dw)$. El término imaginario desaparece debido a la simetría de G alrededor de 0. Por lo tanto, $c(h)$ es una función válida si y sólo si es la función característica de una variable aleatoria simétrica en d-dimensional (variable aleatoria con distribución simétrica). Nótese que si G no se asume simétrica en 0, $c(h) = \int e^{iW^Th}G(dw)$ todavía proporciona una función de covarianza válida (definida positiva), pero ahora para un proceso aleatorio de valores complejos en $\Re^d$.

La transformada de Fourier para $c(h)$ es

$$\widehat{c}(w) = \int e^{-iW^Th}c(h)dh \tag{4.10}$$

Aplicando la fórmula inversa, $c(h) = (2\pi)^{-2}\int e^{iW^Th}\widehat{c}(w)dw$ y se tiene que $(2\pi)^{-d}\widehat{c}(w)/c(0) = g(w)$, la densidad espectral. El cálculo de (4.10) no es posible excepto en casos especiales. La relación uno a uno entre $c(h)$ y $g(w)$ permite examinar los procesos espaciales en el dominio espectral en lugar del dominio observacional.

Banerjee et al. (2004) se limitan al dominio observacional por lo complicado que constituye la construcción de la aproximación al dominio espectral (a través de la transformada rápida de Fourier). Sin embargo, consideran que el análisis usando el dominio espectral puede llevar a un mejor rendimiento computacional cuando se manejan grandes conjuntos de datos.

Las funciones de covarianza isotrópicas son mayormente adoptadas dentro de la clase estacionaria. Es sorprendente que una función de covarianza sea válida (definida positiva) en dimensión d pero no sea válida en dimensión $d+1$. Hay funciones de correlación isotrópicas que son válidas en todas las dimensiones. La función de correlación Gaussiana $k(\|h\|) = \exp(-\phi\|h\|^2)$

es un ejemplo de este tipo de funciones. $k(\|h\|)$ es la función característica asociada con d variables aleatorias normales i.i.d. con varianza $1/(2\phi)$ para cualquier d. En general, la potencia exponencial, $\exp(-\phi\|h\|^{\alpha})$ con $0 < \alpha \leq 2$ es válida para cualquier d.

En lugar de buscar las funciones de correlación isotrópicas que son válidas en todas las dimensiones, se pueden buscar todas las funciones de correlación isotrópicas válidas en una dimensión d particular. Matérn (1960) proporciona un resultado general. Sea $c(\|h\|)$ de la forma

$$c(\|h\|) = \int_0^{\infty} (\frac{2}{w\|h\|})^{\alpha} \Gamma(\nu+1) J_{\nu}(w\|h\|) G(dw) \tag{4.11}$$

donde G es no decreciente e integrable en $\Re^{+}$, J_{ν} es la función Bessel de orden ν y $\nu = (d-2)/2$ ofrece todas las funciones de correlación isotrópicas válidas en $\Re^{d}$.

Si nos limitamos a funciones de covarianza isotrópicas estrictamente monótonas entonces se puede introducir la noción de *rango*. Como ya se ha dicho, el rango es la distancia más allá de la cual la asociación espacial se vuelve insignificante. Si la función de covarianza alcanza el 0 en una distancia finita nos referiremos a esta distancia como el *rango*.

El parámetro ν en la clase Matérn es un parámetro de suavizamiento. En el espacio 2-dimensional, el valor entero más grande de ν indica el número de veces en que el proceso será diferenciable. El uso de la función de covarianza Matérn como modelo permite que los datos disponibles informen sobre ν; podemos aprender sobre el suavizamiento del proceso, a pesar de observar el proceso sólo en un número finito de puntos.

Siguiendo a Stein (1999), la clase Matérn se convierte en una herramienta general para la construcción de modelos espaciales. El cálculo de esta función requiere de una evaluación modificada de la función Bessel. De hecho, la evaluación se llevará a cabo repetidamente para obtener una matriz de covarianza asociada a las n localizaciones y entonces ajustar iterativamente el modelo usando métodos MCMC. Esto parece fuera de

lugar, de hecho, tales cálculos pueden ser realizados usando expansiones para aproximarse a K_ν ó trabajando a través de la fórmula (4.10), la cual en este caso se convierte en

$$2(\frac{\phi\|h\|}{2})^\nu \frac{K_\nu(\phi(\|h\|))}{\phi^{2\nu}\Gamma(\nu+\frac{d}{2})} = \int_{\Re^d} e^{iW^T h}(\phi^2+\|w\|^2)^{(\nu+d/2)}dw, \qquad (4.12)$$

donde K_ν es la función Bessel modificada de orden ν. Más adelante veremos de qué forma usaremos la familia Matérn para realizar predicciones desde el punto de vista Bayesiano.

4.2. Predicción espacial clásica

La predicción espacial en el caso de datos georeferenciados es comúnmente referida como *kriging*. La palabra *kriging* (expresión anglosajona) procede del nombre del geólogo sudafricano D. G. Krige, cuyos trabajos en la predicción de reservas de oro, realizados en la década del cincuenta, suelen considerarse como pioneros en los métodos de interpolación espacial. Kriging encierra un conjunto de métodos de predicción espacial que se fundamentan en la minimización del error cuadrático medio de predicción. Los métodos kriging se aplican con frecuencia con el propósito de predicción, sin embargo estas metodologías tienen diversas aplicaciones, dentro de las cuales se destacan la simulación y el diseño de redes óptimas de muestreo.

De la teoría de decisión se conoce que si Y_0 es una cantidad aleatoria y Y_0^* es su predictor, entonces $L(Y_0, Y_0^*)$ representa la pérdida en que se incurre cuando se predice Y_0 con Y_0^* y el mejor predictor será el que minimice $E\{L(Y_0, Y_0^*)|Y\}$ con $Y = \{Y_1, \ldots, Y_n\}$. Es decir, el predictor óptimo es el que minimice la *esperanza condicional* de la función de pérdida. Si $L(Y_0, Y_0^*) = E(Y_0|Y)$ entonces para encontrar el predictor óptimo se requiere conocer la distribución conjunta de las $n+1$ variables aleatorias.

Un predictor lineal para Y_0 basado en Y debe tener la forma $\sum \ell_i Y(s_i)+\delta_0$. Usando la pérdida el error cuadrático medio, el mejor predictor lineal será el que minimice $E[Y(s_0) - (\sum \ell_i Y(s_i) + \delta_0)^2]$.

Para un proceso de media constante se tendrá que $\sum \ell_i = 1$; en este caso se minimiza la expresión $E[Y(s_0) - (\sum \ell_i Y(s_i))^2] + \delta_0^2$ y δ_0 deberá ser 0.

Un variograma necesariamente debe satisfacer la condición definida negativa. De hecho, para cualquier conjunto de localizaciones $s_1, \ldots, s_n$, conjunto de constantes $a_1, \ldots, a_n$ tales que $\sum a_i = 0$ y $\gamma(h)$ válida, se cumple que $\sum_i \sum_j a_i a_j \gamma(s_i - s_j) \leq 0$.

Por tanto,

$$\sum_i \sum_j a_i a_j \gamma(s_i - s_j) = -E[\sum a_i Y(s_i)]^2 \leq 0$$

Así si hacemos, $a_0 = 1$ y $a_i = -\ell_i$ el predictor se convierte en $E[\sum_{i=0}^{n} a_i Y(s_i)]^2$ con $\sum a_i = 0$. Esta relación revela cómo históricamente el variograma surgió en el *kriging*.

Los ℓ's óptimos pueden obtenerse resolviendo con multiplicadores de Lagrange la condición de optimización definida, los cuales serán funciones de $\gamma(h)$ (Cressie, 1983). Con una estimación de $\gamma(h)$ se obtiene directamente la predicción denominada kriging ordinario. Aparte de la estacionariedad intrínseca del proceso, no es necesario asumir condiciones adicionales en las $Y(s)$.

En el contexto de procesos Gaussianos, si consideramos el caso cuando no se tienen covariables sino solamente variables de respuesta $Y(s_i)$ (kriging ordinario). El modelo para los datos observados viene dada por

$$Y = \mu 1 + \epsilon, \quad \epsilon \sim N(0, \Sigma) \tag{4.13}$$

La estructura espacial de covarianza sin considerar el efecto pepita, se define como

$$\Sigma = \sigma^2 H(\phi), \quad H(\phi)_{ij} = \rho(\phi; d_{ij}) \tag{4.14}$$

donde $d_{ij} = \|s_i - s_j\|$ es la distancia entre s_i y s_j y ρ es una función de correlación válida en $\Re^d$ (Banerjee et al. 2004).
Para un modelo con efecto pepita, se tendrá que Σ viene dada por

$$\Sigma = \sigma^2 H(\phi) + \tau^2 I, \tag{4.15}$$

donde τ^2 es la varianza del efecto pepita.
Cuando se tiene el vector de covariables $x = (x(s_1), \ldots, x(s_n))'$ y $x(s_0)$ disponibles para incorporarse en el análisis, el procedimiento anterior es denominado *kriging universal*. El modelo en este contexto, asume la forma general siguiente

$$Y = X\beta + \epsilon, \quad \epsilon \sim N(0, \Sigma) \tag{4.16}$$

donde, Σ es definida como en (4.15) y (4.14) con o sin efecto pepita.
Nótese que el *kriging ordinario* es un caso particular del *kriging universal* con X (matriz $n \times 1$) como un vector columna $\mathbf{1}$ y β como el escalar μ.
El proceso de predicción se traduce en buscar la función $f(y)$ que minimice el error de predicción cuadrático medio, esto es

$$E[(Y(s_0) - f(y))^2 | y] \tag{4.17}$$

Sumando y restando la media condicional $E[Y(s_0)|y]$ en el término cuadrático y reagrupando los términos en la expresión (4.18) se obtiene

$$E[(Y(s_0) - f(y))^2 | y] = E\{(Y(s_0) - E[Y(s_0)|y])^2\} + \{E[Y(s_0)|y] - f(y)\}^2 \tag{4.18}$$

En (4.18) la esperanza del término del producto cruzado es 0. Ya que el segundo término del lado derecho es no negativo, se obtiene que

$$E[(Y(s_0) - f(y))^2|y] \geq E\{(Y(s_0) - E[Y(s_0)|y])^2|y\} \tag{4.19}$$

para cualquier función $f(y)$. La igualdad se cumple si y sólo si $f(y) = E[Y(s_0)|y]$; así se tiene que el predictor $f(y)$ que minimiza el error es la esperanza condicional de $Y(s_0)$ dado los datos. Este resultado es bastante intuitivo desde el punto de vista Bayesiano, ya que $f(y)$ es justamente la media posterior de $Y(s_0)$, es decir, $f(y)$ minimiza el riesgo posterior (regla de Bayes).
Una vez identificada la *mejor* forma del predictor nos concentraremos en su estimación.

- Consideremos primero la situación irreal en la que todos los parámetros $(\beta, \sigma^2, \phi, \tau^2)$ son conocidos. De la teoría de la normal estándar multivariada tenemos el resultado general siguiente:
  Si $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix}\right)$, con $\Omega_{21} = \Omega_{12}^T$, entonces la distribución condicional $p(Y_1|Y_2)$ es Normal con media y varianza

  $$E[Y_1|Y_2] = \mu_1 + \Omega_{12}\Omega_{22}^{-1}(Y_2 - \mu_2)$$

  $$Var[Y_1|Y_2] = \Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}$$

  En nuestro enfoque $Y_1 = Y(s_0)$ y $Y_2 = y$. De aquí se deduce que $\Omega_{11} = \sigma^2 + \tau^2$, $\Omega_{12} = \gamma^T$ y $\Omega_{22} = \Sigma = \sigma^2 H(\phi) + \tau^2 I$, donde $\gamma^T = (\sigma^2\rho(\phi; d_{01}), \ldots, \sigma^2\rho(\phi; d_{0n}))$. Sustituyendo estos valores en la media y la varianza recién formuladas, obtenemos

  $$E[Y(s_0)|y] = x_0^T\beta + \gamma^T\Sigma^{-1}(y - X\beta), \tag{4.20}$$

$$Var[Y(s_0)|y] = \sigma^2 + \tau^2 - \gamma^T \Sigma^{-1} \gamma \tag{4.21}$$

Esta solución asume que conocemos el valor de la covariable $x_0 = x(s_0)$ en el nuevo sitio s_0. Se puede considerar no hacer predicción en una nueva ubicación, pero si, en uno de los lugares ya observados. En este caso el factor de predicción (4.20) es igual al valor observado en ese lugar si $\tau^2 = 0$.

- Consideremos ahora, el escenario más realista en el que los parámetros del modelo son desconocidos y debemos estimarlos de los datos. Modificamos la expresión de $f(y)$ como sigue

$$\widehat{f(y)} = x_0^T \hat{\beta} + \hat{\gamma}^T \hat{\Sigma}^{-1}(y - X\hat{\beta}) \tag{4.22}$$

donde $\hat{\gamma} = (\hat{\sigma}^2 \rho(\hat{\phi}; d_{01}), \ldots, \hat{\sigma}^2 \rho(\hat{\phi}; d_{0n}))^T$, $\hat{\beta} = (X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} y$, el estimador usual para β de mínimos cuadrados y $\hat{\Sigma} = \hat{\sigma}^2 H(\hat{\phi})$. Así $\widehat{f(y)}$ puede ser reescrita como $\lambda^T y$, con

$$\lambda = \hat{\Sigma}^{-1}\hat{\gamma} + \hat{\Sigma}^{-1} X (X^T \hat{\Sigma}^{-1} X)^{-1}(x_0 - X^T \hat{\Sigma}^{-1}\hat{\gamma}) \tag{4.23}$$

Si x_0 no es observado, podemos estimarlo junto con $Y(s_0)$ iterando esta formula y haciendo $\hat{x}_0 = X^T \lambda$ que surge al multiplicar ambos lados de (4.23) por X^T y simplificando.

4.3. Predicción espacial Bayesiana

En el enfoque clásico se invierte mucho tiempo en determinar las estimaciones de las ecuaciones presentadas anteriormente. Tradicionalmente se emplean métodos restringidos de máxima verosimilitud (REML) para las estimaciones y se alcanzan ciertas propiedades óptimas.

Sin embargo, desde la perspectiva Bayesiana el desarrollo de esta estimación no es un problema. Ya que se imponen distribuciones previas a los parámetros desconocidos y se produce la distribución posterior predictiva $p(Y(s_0)|y)$ y cualquier estimación puntual o por intervalos puede ser calculada a partir de esta distribución.

Existen programas para realizar el kriging desde el enfoque Bayesiano. Además del software WinBUGS, es posible desde el software R invocando a la librería GeoR usar la función krige.bayes para hacer predicción (kriging ordinario y universal). Esta función no es tan versátil como el WinBUGS, ya que es más limitada en los tipos de modelos disponibles y la actualización de los parámetros de interés no se hace usando métodos MCMC. Sin embargo, es una herramienta práctica que proporciona muestras posteriores de todos los parámetros del modelo y de su variabilidad. La función krige.bayes implementa métodos Bayesianos sólo para el modelo Gaussiano, muestreando de la distribución posterior y realizando la predicción.

En un análisis realista, no siempre se parte de un modelo Gaussiano y nunca se conoce el variograma ya que se ignora la incertidumbre real de sus parámetros. Desde la perspectiva Bayesiana se pueden asignar distribuciones previas al alféizar parcial, el rango y la pepita relativa. Si se mantiene como conocidos el rango y la pepita, el análisis es bastante rápido, se puede asignar una previa recíproca ($\propto 1/\sigma^2$) para el alféizar parcial o una χ^2-escalada-inversa.

El análisis completo con todos los parámetros desconocidos requiere de un esfuerzo computacionalmente considerable. Se puede emplear una previa recíproca para el alféizar parcial y discretas uniformes para el rango y para la pepita relativa. Cuando se supone desconocido el variograma es necesario por tanto, la implementación del kriging a través de métodos MCMC.

Desde una perspectiva Bayesiana se debe establecer la estructura

jerárquica probabilística para el modelo. Esta formulación comprende las siguientes definiciones

$$Y|\theta, W \sim N(\mu + W, \tau^2 I) \tag{4.24}$$

donde $\mu = X\beta$ recogerá la variabilidad a gran escala y W es el vector de efectos espaciales.
En el segundo nivel, se especifica a W como una distribución normal multivariante con matriz de covarianzas expresada como función paramétrica de la distancia entre pares de puntos. Así, W se define como

$$W|\sigma^2, \phi \sim N(0, \sigma^2 H(\phi)) \tag{4.25}$$

donde H es una matriz de correlaciones indexada por el parámetro ϕ. La formulación completa del modelo requiere la asignación de previas en los parámetros alféizar parcial, el rango y la pepita relativa (Banerjee et al. 2004).

4.4. Campos Gaussianos y Campos Aleatorios de Markov Gaussianos

Los campos Gaussianos, en adelante abreviados como GF (siglas en inglés), tienen un rol dominante en estadística espacial y en el campo geoestadístico (Cressie, 1993; Stein, 1999; Diggle y Ribeiro, 2007; entre otros) y constituyen un componente importante de los modelos jerárquicos espaciales actuales (Banerjee et al. 2004). Los GFs son uno de los pocos modelos multivariantes apropiado con una constante de normalización explícita y con buenas propiedades analíticas.
En un dominio $D \in \Re^d$ con coordenadas $s \in D$, $x(s)$ es un GF continuamente indexado, si toda la colección finita $\{x(s_i)\}$ tiene conjuntamente distribución Gaussiana. En la mayoría de los casos, el GF se

especifica mediante una función de media $\mu(.)$ y una función de covarianza $C(.)$, así, $\mu = (\mu(s_i))$ y la matriz de covarianza es $\Sigma = (C(s_i, s_j))$.

A menudo, la función de covarianza es sólo una función de la posición relativa de dos localizaciones, en cuyo caso se dice que es estacionaria y es isotrópica, si las funciones de covarianza sólo dependen de la distancia euclídea entre las ubicaciones. Dado que una matriz de covarianza regular es definida positiva, la función de covarianza debe ser una función definida positiva. Esta restricción hace difícil establecer una función de covarianza en forma cerrada. El Teorema de Bochner permite en este contexto, caracterizar a todas las funciones continuas definidas positivas en $\Re^d$.

A pesar de la conveniencia de los GFs desde el punto de vista analítico y práctico, los problemas de cálculo siempre han sido un cuello de botella. Esto se debe al costo computacional $O(n^3)$ que implica la factorización de matrices de covarianza densas de orden $n \times n$. La creciente popularidad de los modelos jerárquicos Bayesianos ha hecho que esta situación sea muy importante debido a la necesidad de repetir simulaciones para el ajuste de los modelos, lo cual puede resultar poco viable (Banerjee et al. 2004); esta situación es informalmente referida como "el problema de n grande".

Se han propuesto diversas metodologías para enfrentar el llamado "problema de n grande". Específicamente en este capítulo, se empleará el enfoque propuesto por Lindgren et al. (2011), en el cual, un GF es reemplazado por un campo aleatorio de Markov Gaussiano (GMRF, siglas en inglés). Consultar a Rue y Held (2005) y a Rue et al. (2009) para más detalles de esta metodología.

Un GMRF x, es un campo Gaussiano indexado discretamente, donde las condicionales completas $\pi(x_i|x_{-i})$ con $i = 1, \ldots, n$, dependen solamente de un conjunto de vecinos ∂_i para cada localización i (si $i \in \partial_j$ entonces $j \in \partial_i$). La notación $x = (x_1, \ldots, x_n)$ con $x \sim \mathrm{N}(\mu, Q^{-1})$ se refiere a un GMRF n-dimensional con media μ y matriz de precisión simétrica y definida positiva Q (inversa de la matriz de covarianza).

Sea G un grafo no dirigido que denota las propiedades de independencia condicional de x, entonces G será un GMRF con respecto a G, si la media de x es μ y la densidad de x esta dada por

$$\pi(x) = (2\pi)^{-n/2}|Q|^{1/2}exp(-\frac{1}{2}(x-\mu)'Q(x-\mu)) \tag{4.26}$$

La distribución condicional completa de x_i $(i = 1, \ldots, n)$ depende sólo de unos pocos componentes de x_{-i}, esto gracias a la propiedad Markoviana relacionada con la estructura de vecindad. Así, ∂_i constituye el conjunto de vecinos de cada unidad i,

$$\pi(x_i|x_{-i}) = \pi(x_i|x_{\partial_i}) \tag{4.27}$$

la notación x_{-i} denota a todos los elementos de x pero sin x_i. Esto es equivalente a decir, que dada la estructura de vecindad ∂_i, los términos x_i y $x_{\{-i,\partial_i\}}$ son independientes. Siguiendo la notación de Rue y Held (2005) se tiene que esta relación de independencia puede ser expresada como

$$x_i \perp x_{-\{i,\partial i\}}|x_{\partial_i} \tag{4.28}$$

para $i = 1, \ldots, n$. El punto clave es que esta propiedad de independencia condicional esta estrictamente relacionada con la matriz de precisión Q. De hecho, para una pareja (i, j) cualquiera con $i \neq j$, se tendrá que

$$x_i \perp x_j|x_{-\{i,j\}} \Longleftrightarrow Q_{ij} = 0 \tag{4.29}$$

lo cual significa que el patrón de no-ceros de Q esta dado por la estructura de vecindad del proceso. Luego, $Q_{ij} \neq 0$ si $j \in \{i, \partial_i\}$. Esta propiedad permite realizar rápidas factorizaciones de Q como LL^T, donde L es la triangular inferior de Cholesky. La forma de Q se hereda de L gracias a la propiedad global de Markov : para $i < j$, tales que i, j están separados por $F(i, j) = \{i+1, \ldots, j-1, \ldots, j+1, \ldots, n\}$ en G, $L_{ij} = 0$. Sólo se calculan

los términos no nulos de L y además los nodos se pueden ordenar para disminuir el número de términos distintos de cero en L.
La densidad $log(\pi(x))$ puede ser fácilmente calculada por la ecuación (4.26) ya que $log|Q| = 2\sum_i logL_{ii}$, además la varianzas marginales también se pueden calcular eficientemente. Estas varianzas se encuentran partiendo de la ecuación $L^T x = z$ con $z \sim N(0, I)$. Si rescribimos esta ecuación tenemos que $L_{ii}x_i = z_i - \sum_{k=i+1}^{n} L_{ki}x_k$ para $i = n, \ldots, 1$. Multiplicando cada lado por x_j y tomando valor esperado, se obtiene

$$\Sigma_{ij} = \delta_{ij}/L_{ij}^2 - \frac{1}{L_{ii}} \sum_{k=i+1}^{n} L_{ki}\Sigma_{kj}, \qquad j \geq i, \qquad i = n, \ldots, 1 \tag{4.30}$$

$\Sigma(= Q^{-1})$ es la matriz de covarianza y $\delta_{ij} = 1$ si $i = j$, 0 en otro caso. Cuando el GMRF es definido con restricciones adicionales como $Ax = e$ para una matriz $A_{k\times n}$ de rango k, entonces partiendo de x sin restricciones se puede obtener un GMRF con restricciones de la siguiente forma

$$x^c = x - Q^{-1}A^T(AQ^{-1}A^T)^{-1}(Ax - e) \tag{4.31}$$

La ganancia computacional de hacer inferencia usando un GMRF se deriva directamente de la matriz de precisión Q. De hecho, las operaciones de algebra lineal pueden realizarse usando métodos numéricos para matrices dispersas. La factorización de una matriz densa que usualmente requiere $O(n^3)$ intentos, se reduce a $O(n)$, $O(n^{3/2})$ y $O(n^2)$ para una matriz dispersa en el caso de un GMRF temporal, espacial y espacio-temporal respectivamente.
En general, los GMRFs tienen propiedades computacionales muy buenas, que son de gran importancia en los métodos inferenciales Bayesianos. Los GMRFs junto con la metodología INLA se convierten en un marco excelente para realizar inferencia bayesiana en forma rápida y precisa en campos Gaussianos latentes.

A pesar de las ventajas computacionales de los GMRFs, hasta el momento no ha existido una forma óptima de parametrizar la matriz de precisión (Besag and Kooperberg, 1995; Rue y Tjelmeland, 2002). La restricción definida positiva de la matriz Q complica los cálculos y puede que no sea evidente cómo esta condición influye en la parametrización de las condicionales completas.

Rue y Tjelmeland (2002) demostraron empíricamente que los GMRFs se pueden aproximar a la mayoría de las funciones de covarianza usadas en geoestadística y propusieron usar a los GMRFs como una aproximación a los GFs por razones de cálculo al hacer el kriging (Hartman y Hössjer, 2008).

Las dificultades numéricas de: modelar un GF mediante la construcción de un GF discretizado con matriz de covarianza Σ; encontrar un GMRF con estructura de vecindad y matriz de precisión Q que represente adecuadamente el GF y el realizar los cálculos utilizando matrices dispersas; se pueden resolver usando miembros de los campos Gaussianos con la función de covarianza Matérn en $\Re^d$, donde la representación GMRF esta disponible explícitamente (Stein, 1999).

La representación de un campo aleatorio Gaussiano de Markov puede ser construida explícitamente mediante una ecuación diferencial parcial estocástica (SPDE en inglés) (Lindgren et al. 2011). Sorprendentemente, la ampliación de este resultado fundamental parece abrir nuevas puertas y oportunidades, y permite dar respuestas muy simples a los problemas de modelado más difíciles.

4.4.1. Modelos Gaussianos Latentes

Los modelos Gaussianos latentes son modelos jerárquicos con estructura de regresión aditiva. En estos modelos las variables respuesta (observaciones) y_i se asumen pertenecen a una familia exponencial, donde, la media μ_i está vinculada a un predictor η_i de estructura aditiva a través de la función

de vínculo $g(\cdot)$ de modo que $g(\mu_i) = \eta_i$. La estructura del predictor η_i incluye el efecto de varias covariables en forma aditiva, de la siguiente manera

$$\eta_i = \beta_0 + \sum_{j=1}^{n_f} w_{ji} f^j(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki} + \varepsilon_i \tag{4.32}$$

Los $\{\beta_k\}$ representan los efectos lineales de las covariables z. Las $\{f^j(\cdot)\}$ son funciones desconocidas de las covariables u. Estas pueden tener diferentes formas: efectos no lineales de covariables continuas, tendencias temporales, interceptos aleatorios i.i.d., grupos específicos de efectos aleatorios y efectos aleatorios espaciales. Los w_{ij} son pesos conocidos definidos por cada dato observado. Finalmente, los ε_i son efectos aleatorios sin estructura espacial.

Un modelo Gaussiano latente se obtiene al asignar a $x = \{\{f^j(\cdot)\}, \{\beta_k\}, \{\eta_i\}\}$ una previa Gaussiana con matriz de precisión $Q(\theta)$ con hiperparámetro θ. La parametrización que usaremos incluye los η_i en lugar de los ε_i. La distribución de las variables $y = \{y_1, \ldots, y_n\}$ es denotada por $\pi(y|x, \theta)$ y asumiremos que las y_i son condicionalmente independientes dado x y θ. Para simplificar denotamos a $\theta = (\theta_1^T, \theta_2^T)$ con $\dim(\theta) = m$. La posterior ($Q(\theta)$ no singular) es de la forma

$$\begin{aligned} \pi(x, \theta|y) &\propto \pi(\theta)\pi(x|\theta)\Pi_i \pi(y_i|x_i, \theta) \\ &\propto \pi(\theta)|Q(\theta)|^{n/2} \exp\left(-\frac{1}{2}x^T Q(\theta)x + \sum_i \log\pi(y_i|x_i, \theta)\right) \end{aligned} \tag{4.33}$$

El objetivo de esta modelización es aproximarse a las marginales posteriores de $\pi(x_i|y)$, $\pi(\theta|y)$ y $\pi(\theta_j|y)$. Los modelos Gaussianos considerados satisfacen las dos propiedades básicas siguientes: La primera, el campo latente x, el cual a menudo es de alta dimensión (n entre 10^2 y 10^5) admite propiedades de independencia condicional; esto gracias a la forma de la

matriz de precisión Q. La segunda propiedad, es que la dimensión m del vector de hiperparámetros θ es pequeña, es decir, $m \leq 6$. Estas propiedades son satisfechas por muchos modelos Gaussianos latentes existentes en la literatura. Existen excepciones y los modelos geoestadísticos son una de ellas. Sin embargo, a través de la metodología INLA se puede aplicar a modelos geoestadísticos empleando un cálculo computacional diferente o utilizando una representación Markov del campo Gaussiano (Eidsvik et al. 2009 y Lindgren et al. 2011).

4.5. El enfoque SPDE

Sea $x(s) \equiv \{x(s), s \in D \subseteq \Re^2\}$ un campo Matérn, es decir, un GF estacionario de segundo orden e isotrópico con función de covarianza Matérn dada por

$$C(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^{\nu} K_{\nu}(\kappa h) \tag{4.34}$$

Como ya se señalo en la sección (4.1.4), K_ν es la función Bessel modificada de segundo tipo y orden $\nu > 0$. El parámetro μ es usualmente fijo y mide el grado de suavizamiento del proceso y su valor entero determina la diferenciabilidad cuadrática media del proceso. κ es un parámetro de escala relacionado con el rango. En particular, se usará la definición empírica de $\rho = \frac{\sqrt{8\nu}}{\kappa}$, en este caso, ρ corresponde a la distancia donde la correlación espacial esta cerca de 0.1 para $\forall\nu$. La función de correlación espacial $C(h)$ dependerá de las localizaciones s_i y s_j sólo a través de la distancia Euclídea $h = \parallel s_i - s_j \parallel \in \Re$.

La función de covarianza Matérn aparece naturalmente en varios campos científicos (Guttorp y Gneiting, 2006). Sin embargo, en esta ocasión se establece una relación entre el campo Gaussiano y la función de covarianza Matérn como una solución de ecuaciones diferenciales parciales estocásticas

de la siguiente forma

$$(\kappa^{2-\Delta})^{\alpha/2}x(u) = W(u), u \in \Re^d, \alpha = \nu + d/2, \kappa > 0, \nu > 0 \tag{4.35}$$

donde $(\kappa^2 - \Delta)^{\alpha/2}$ es un operador pseudo-diferencial definido en la ecuación (4.39) a través de sus propiedades espectrales (Whittle, 1954, 1963). El proceso de innovación W es espacial Gaussiano de ruido blanco con varianza unitaria; Δ es el Laplaciano

$$\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} \tag{4.36}$$

y la varianza marginal es

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}\kappa^{2\nu}} \tag{4.37}$$

En adelante a cualquier solución de la ecuación (4.35) se llamará un campo Matérn. Las soluciones limites bajo el enfoque SPDE cuando $\kappa \to 0$ o $\nu \to 0$ no tienen funciones de covarianza Matérn, sin embargo, existe solución cuando $\kappa = 0$ o $\nu = 0$ si se definen bien las medidas aleatorias. Cuando $\alpha \geq 2$ el espacio nulo del operador diferencial no es trivial, y contiene por ejemplo, las funciones $\exp(\kappa e^T u)$ para todo $\|e\| = 1$. Los campos Matérn son las únicas soluciones estacionarias a las ecuaciones parciales diferenciales estocásticas.

La prueba dada por Whittle (1954, 1963) demostró que el número de ondas del espectro de una solución estacionaria es

$$R(k) = (2\pi)^{-d}(\kappa^2 + \|k\|^2)^{-\alpha} \tag{4.38}$$

usando la definición de la transformada de Fourier del Laplaciano fraccionado en $\Re^d$ se tiene que

$$\{(\kappa^2 - \Delta)^{\alpha/2}\phi\}(k) = (\kappa^2 + \|k\|^2)^{\alpha/2}(\phi)(k) \tag{4.39}$$

donde ϕ es una función en $\Re^d$. El objetivo del enfoque SPDE es encontrar un GMRF con estructura de vecindad y matriz dispersa de precisión Q que mejor represente el campo Matérn. Dada esta representación, es posible hacer inferencia usando el GMRF encontrado y sus buenas propiedades de cálculo.

Básicamente el enfoque SPDE usa una representación finita para definir el campo Matérn como una combinación lineal de funciones base definidas en una triangulación en el dominio D. Esta triangulación consiste en dividir a D en un conjunto de triángulos no interceptados unidos por al menos un borde o esquina común. En primer lugar, los vértices de los triángulos iniciales son colocados en las localizaciones $s_1, \ldots, s_n$ y luego se agregan vértices adicionales en orden para obtener una triangulación útil para la predicción espacial deseada.

Teniendo en cuenta la triangulación, la representación de la función base del campo Matérn $X(s)$ está dada por

$$X(s) = \sum_{i=1}^{n} \psi_l(s) w_l \tag{4.40}$$

donde n es el número total de vértices, $\{\psi_l(s)\}$ son las funciones base y w_l son pesos con distribución Gaussiana. Las funciones $\{\psi_l(s)\}$ son seleccionadas para que sean trozos lineales en cada triángulo, es decir, $\psi_l(s)$ es 1 en el vértice l y 0 en los otros vértices. La altura de cada triángulo (el valor del campo espacial en cada vértice del triángulo) es dada por el peso w_l y los valores en el interior del triángulo son determinados por interpolación lineal.

El punto clave del enfoque SPDE es la representación finita de (4.40) que establece el vínculo entre el GF $X(s)$ y el GMRF definido por los pesos Gaussianos w_l, a los cuales se les puede asignar una estructura Markoviana como lo demuestra Lindgren et al. (2011).

En particular, la matriz de precisión Q del GMRF esta definida por

la ecuación $w_l \sim N(0, Q_S^{-1})$ como función de κ^2, para $\alpha = 1, 2, \ldots$, $\nu = 0, 1, 2, \ldots$ y $\alpha = \nu + 1$.

4.6. Inferencia: un nuevo enfoque

La aproximación de Laplace basada en integrales anidadas (INLA) es una metodología introducida por Rue y Martino (2007) y por Rue et al. (2009) para realizar inferencia estadística en modelos Gaussianos latentes. INLA proporciona una forma rápida y eficiente de hacer inferencia Bayesiana usando aproximaciones precisas de la densidad marginal posterior de los hiperparámetros $\tilde{\pi}(\theta|y)$ y a las condicionales completas de las marginales posteriores de las variables latentes $\tilde{\pi}(x_i|\theta, y)$, $i = 1, \ldots, n$.
El primer paso en la aproximación INLA es realizar una aproximación Laplace a la posterior conjunta

$$\begin{aligned} \pi(\theta|y) =& \frac{\pi(\theta)\pi(x,\theta)\pi(y|x)}{\pi(x|\theta,y)} \\ \propto& \frac{\pi(\theta)\pi(x,\theta)\pi(y|x)}{\pi_G(x|\theta,y)} \end{aligned} \tag{4.41}$$

La aproximación para $\pi(\theta|y)$ está basada en el trabajo propuesto por Tierney y Kadane (1986), mientras que $\pi(x_i|\theta, y)$ puede ser aproximada usando tres enfoques diferentes: la aproximación Gaussiana, una aproximación completa de Laplace y una simplificada de Laplace. Cada uno de estos enfoques tiene características de tiempos de cálculo y precisión diferentes. La aproximación Gaussiana es más rápida en el cálculo, pero pueden ocurrir errores en la localización de la media posterior y/o errores debido a la falta de simetría. La aproximación completa de Laplace es más precisa, pero requiere de mayor tiempo computacional. Mientras que la versión simplificada es rápida de calcular y genera aproximaciones los suficientemente precisas (Rue et al. 2009).

La clave de este nuevo enfoque de inferencia esta en aproximar las marginales posteriores de x_i por las aproximaciones anidadas de

$$\tilde{\pi}(x_i|y) = \int \tilde{\pi}(x_i|\theta, y)\tilde{\pi}(\theta|y)d\theta \approx \sum_{k=1}^{K} \tilde{\pi}(x_i|\theta_k, y)\tilde{\pi}(\theta_k|y)\Delta_k \qquad (4.42)$$

$\tilde{\pi}(\cdot|\cdot)$ es una densidad aproximada condicional. Las aproximaciones de (4.42) son calculadas por aproximaciones en $\pi(\theta|y)$ y $\pi(x_i|\theta, y)$ usando integración numérica (suma finita) sobre θ. Las marginales posteriores para los hiperparámetros $\tilde{\pi}(\theta_j|y)$, $j = 1, \ldots, m$ se determinan en forma similar. La inferencia esta basada en la aproximación $\tilde{\pi}(\theta|y)$ de la marginal posterior de θ:

$$\tilde{\pi}(\theta|y) \propto \frac{\pi(x, \theta, y)}{\tilde{\pi}_G(x|\theta, y)}|_{x=x^*(\theta)} \qquad (4.43)$$

donde $\tilde{\pi}_G(x|\theta, y)$ es la aproximación Gaussiana de la condicional completa de x y $x^*(\theta)$ es la moda de la condicional completa de x para un θ dado. El signo de proporcionalidad se debe a que la constante de normalización para $\pi(x, \theta|y)$ es desconocida. Esta expresión es equivalente a la aproximación de Laplace de Tierney y Kadane (1986) y esto sugiere que el error de aproximación es relativo y de orden $O(n^{-3/2})$ después de la renormalización.

Nótese que $\tilde{\pi}(\theta|y)$ tiende a alejarse demasiado de la Gaussianidad, por lo tanto, este enfoque determina las aproximaciones de $\tilde{\pi}(\theta|y)$ y $\tilde{\pi}(x_i|y)$ en forma no paramétrica. La herramienta principal para realizar inferencia es la aplicación de la aproximación de Laplace a $\pi(x_i|\theta, y)$.

La estrategia de integración sobre los puntos de θ_k propuesta por Rue et al. (2009) se denomina Central Composite Design (CCD), la cual consiste en la colocación de una pequeña cantidad de "puntos" en un espacio m-dimensional con el fin de estimar la curvatura de $\tilde{\pi}(\theta|y)$. Dicha estrategia es por lo general lo suficientemente precisa para el cálculo de $\tilde{\pi}(x_i|y)$. Las aproximaciones de las marginales posteriores con esta estrategia pueden

ser usadas para calcular resúmenes de medidas estadísticas posteriores (medias, varianzas y cuantiles). Usaremos la estrategia CCD para la obtención de resultados.

Al implementar la metodología INLA en el software R es posible encontrar, como subproducto de los cálculos principales, otras cantidades de interés como el Criterio de Información de Deviance (DIC), verosimilitudes marginales y medidas predictivas.

Implementación del DIC

El DIC (Spiegelhalter et al. 2002) bajo este enfoque puede ser calculado en dos pasos: primero, se calcula la media condicionada en θ usando integración numérica en cada i ($i = 1, \ldots, n$ =total de marginales); segundo, se integra con respecto a $\pi(\theta|y)$. La deviance de la media requiere la media posterior de cada x_i, las cuales son calculados de las marginales posteriores de x_i. En cuanto a los hiperparámetros, se prefiere usar la moda posterior θ^*, ya que la marginal posterior para θ puede estar severamente sesgada.

La exactitud de $\tilde{\pi}(\theta|y)$ parece estar directamente relacionada con la "verdadera" dimensión de x. Rue et al. (2009) recomiendan evaluar el número efectivo de parámetros condicionado en θ. Ya que x dado y y θ es aproximadamente Gaussiano, $p_D(\theta)$ es aproximado convenientemente por

$$p_D(\theta) = n - \text{Trace}\{Q(\theta)Q^*(\theta)^{-1}\} \tag{4.44}$$

n, se refiere al total de marginales. El cálculo de $p_D(\theta)$ no es computacionalmente costoso, ya que las covarianzas de los vecinos son obtenidas como un producto de las varianzas marginales en la aproximación Gaussiana basada en la expresión (4.30). $p_D(\theta)$ también mide en qué medida la Gaussianidad y la estructura de dependencia de la previa se conservan en la posterior

de x dado θ. Así por ejemplo, para datos no informativos se tendrá que $p_D(\theta) = 0$ y el error de aproximación es cero, ya que la posterior es igual a la previa Gaussiana. Rue et al. (2009) observaron en todas sus aplicaciones que $p_D(\theta)$ era relativamente más pequeño a n_d para valores de θ en las proximidades de la moda posterior. n_d se refiere al número de datos observados. Si $p_D(\theta)$ es menor a n_d entonces se tendrá un modelo con buen ajuste.

Verosimilitudes marginales

La verosimilitud marginal $\pi(y)$ es una cantidad útil para comparar modelos, así como los factores de Bayes, se define como el radio de verosimilitudes marginales de dos modelos en competencia. Es evidente a partir de (4.43) que la aproximación natural a la verosimilitud marginal es la constante de normalización de $\tilde{\pi}(\theta|y)$,

$$\tilde{\pi}(y) = \int \frac{\pi(\theta, x, y)}{\tilde{\pi}_G(x|\theta, y)}|_{x=x^*(\theta)} d\theta \tag{4.45}$$

$\pi(\theta, x, y) = \pi(\theta)\pi(x|\theta)\pi(y|x, \theta)$. Una estimación más cruda de la verosimilitud marginal se obtiene asumiendo que $\theta|y$ es Gaussiana; entonces (4.45) se convierte en una constante conocida, determinada por la matriz Hessiana $|H|^{-1/2}$ (con $\Sigma = H^{-1}$). La aproximación (4.45) no requiere de esta condición, pues $\tilde{\pi}(\theta|y)$ es tratada en forma no paramétrica. Kass y Vaidyanathan (1992) usan una expresión similar a (4.45) para aproximarse a la verosimilitud marginal en un contexto diferente.

Medidas predictivas

Las medidas predictivas pueden ser usadas para validar y comparar modelos (Gelfand, 1996; Gelman et al. 2003) así como un mecanismo para detectar posibles outliers u observaciones sospechosas (Pettit, 1990).

Usualmente la densidad predictiva se calcula para y_i basada en el resto de observaciones, es decir, $\pi(y_i|y_{-i})$. Ahora explicaremos cómo aproximar esta cantidad. Primero, nótese que remover a y_i del conjunto de datos afecta a las marginales de x_i y a θ como sigue:

$$\pi(x_i|y_{-i},\theta) \propto \frac{\pi(x_i|y,\theta)}{\pi(y_i|x_i,\theta)} \tag{4.46}$$

$$\pi(\theta|y_{-i}) \propto \frac{\pi(\theta|y)}{\pi(y_i|y_{-i},\theta)} \tag{4.47}$$

Se requiere calcular la integral

$$pi(y_i|y_{-i},\theta) = \int \pi(y_i|x_i,\theta)\pi(x_i|y_{-i},\theta)dx_i \tag{4.48}$$

Usualmente valores pequeños de $\pi(y_i|x_i,\theta)$ indican observaciones sospechosas, pero lo que se entiende por "pequeño", debe calibrarse con el nivel de x_i. Pettit (1990) sugiere calibrar con el valor máximo de $\pi(\cdot|y_{-i})$, pero una alternativa es calcular la probabilidad integral transformada PIT_i =Prob($y_i^{\text{nuevo}} \leq y_i|y_{-i}$) utilizando el mismo mecanismo anterior (4.48).

Un valor de PIT_i inusualmente pequeño (cerca de 0) o grande (cerca de 1) asumiendo observaciones continuas, indica una posible observación sospechosa, la cual requiere de mayor atención. Además, si el histograma de las PIT_i esta demasiado lejos de una distribución uniforme, el modelo puede ser cuestionado (Czado et al. 2007).

4.7. Modelización bajo la metodología INLA

En esta sección se propone una modelización general basada en la metodología INLA y en el enfoque SPDE. El modelo jerárquico con

estructura espacial que desarrollamos será capaz de predecir en lugares no observados la presencia de cierto evento de interés, evento que en principio puede originarse en fenómenos de distinta naturaleza.

El enfoque Bayesiano es apropiado en el caso de modelos jerárquicos espaciales, porque permite que tanto los datos observados como los parámetros del modelo sean variables aleatorias resultando en una estimación más realista y precisa de la incertidumbre (Banerjee et al. 2004). Otra ventaja de este paradigma es la facilidad para incorporar información a priori; esta información puede ser útil en la discriminación de los efectos espaciales de autocorrelación de aquellos efectos lineales ordinarios no espaciales (Gaudard et al. 2006).

En situaciones donde interesa conocer la ocurrencia de un evento de interés y el proceso espacial puede ser visto como un continuo se puede seguir a Diggle et al. (1998) y pensar en un modelo jerárquico para datos geoestadísticos que indique la presencia/ausencia del evento. Específicamente, si Y_i es una variable Bernoulli que representa la presencia (1) o ausencia (0) en la localización i ($i = 1, \ldots, n$) y π_i es la probabilidad de la presencia, entonces

$$Y_i \sim \text{Ber}(\pi_i) \tag{4.49}$$

En la segunda capa del modelo definimos el logit de la probabilidad π_i

$$\text{logit}(\pi_i) = \beta_0 + W_i \tag{4.50}$$

donde, β_0 representa el intercepto del predictor lineal para la observación i; W_i representa el efecto aleatorio con estructura espacial. Mientras que la relación entre π_i, las covariables de interés y el efecto aleatorio es modelada a través de la función de vínculo logit. En esta propuesta no se incluye el efecto de ninguna covariable, por tanto, la probabilidad del evento viene determinada sólo por el intercepto y el efecto aleatorio espacial.

W_i se asume como una distribución Gaussiana con matriz de covarianza $\sigma_W^2 H(\phi)$, dada por la distancia entre localizaciones y con hiperparámetros σ_W^2 y ϕ que representan, respectivamente, la varianza (alféizar parcial en terminología del kriging clásico) y el rango del efecto espacial. De esta forma W_i tendrá la distribución probabilística siguiente,

$$W \sim N(0, \sigma_W^2 H(\phi)) \tag{4.51}$$

La estructura para $H(\phi)$ viene determinada por la función Matérn

$$C(h) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(\kappa h)^\nu K_\nu(\kappa h)$$

Como ya se señalo en la sección (4.1.4), K_ν es la función Bessel modificada de segundo tipo y orden $\nu > 0$. El parámetro μ es usualmente fijo y mide el grado de suavizamiento del proceso y su valor entero determina la diferenciabilidad cuadrática media del proceso. κ es un parámetro de escala relacionado con el rango. En particular, se usará la definición empírica de $\rho = \frac{\sqrt{8\nu}}{\kappa}$, en este caso, ρ corresponde a la distancia donde la correlación espacial esta cerca de 0.1 para $\forall\nu$. La función de correlación espacial $C(h)$ dependerá de las localizaciones s_i y s_j sólo a través de la distancia Euclídea $h = \parallel s_i - s_j \parallel \in \Re$.

La modelización propuesta puede ser aumentada incorporando un término puro de error conocido como efecto pepita (en terminología del kriging clásico). Este efecto describe el "ruido" asociado a la replica de medición en cada localización, usualmente cuando se emplea el enfoque Bayesiano es común asignarle una distribución Gaussiana.

Bajo el paradigma Bayesiano es necesario asignar distribuciones previas a cada parámetro involucrado en el modelo $(\beta_0, \sigma_W^2, \phi)$. En este sentido, la selección usualmente es tratar con previas independientes para los parámetros (Banerjee et al. 2004), es decir,

$$p(\beta_0, \sigma_W^2, \phi) = p(\beta_0)p(\sigma_W^2)p(\phi) \tag{4.52}$$

Cuando se quiere expresar un conocimiento vago, pero útil acerca de los parámetros, se elige como distribuciones candidatas, una previa no-informativa Gaussiana para β y distribución inversa gamma para σ_W^2. La especificación de la distribución para ϕ dependerá de la función de correlación elegida (Banerjee et al. 2004), la cual determina la matriz de covarianza H. La selección final de las previas dependerá del tipo de modelado elegido y de la parametrización definida.
Las expresiones de (4.49) a (4.52) contienen todo nuestro conocimiento sobre la distribución posterior, pero no producen expresiones cerradas para las distribuciones posteriores de los parámetros. La forma general de la distribución posterior para las variables $y = \{y_1, \ldots, y_n\}$ denotada por $\pi(y|x, \theta)$ con $\theta = (\theta_1^T, \theta_2^T)$ con $\dim(\theta) = 2$ es la siguiente

$$\begin{aligned} \pi(x, \theta|y) &\propto \pi(\theta)\pi(x|\theta)\Pi_i\pi(y_i|x_i, \theta) \\ &\propto \pi(\theta)|Q(\theta)|^{n/2}\exp\left(-\frac{1}{2}x^T Q(\theta)x + \sum_i \log\pi(y_i|x_i, \theta)\right) \end{aligned} \tag{4.53}$$

con $Q(\theta)$ no singular. El objetivo de esta modelización es aproximarse a las marginales posteriores de $\pi(x_i|y)$, $\pi(\theta|y)$ y $\pi(\theta_j|y)$. Esta aproximación se hará usando la metodología INLA (Rue et al. 2009) y el kriging Bayesiano se hará siguiendo el enfoque SPDE propuesto por Lindgren et al. (2011).
La idea clave que subyace en esta modelización, es darse cuenta que estos modelos jerárquicos pueden ser vistos como modelos estructurados aditivos de regresión ($STAR$) (Fahrmeir et al. 2001). En otras palabras, modelos en los que la media de la variable respuesta Y_i está vinculada a un predictor que representa los efectos de diversas covariables en forma aditiva.
En contraste a lo que hace el WinBUGS (Spiegelhalter et al. 2003) en cuanto a la asignación de las previas, cuando se trata del enfoque SPDE la función de correlación no se modela directamente. En este caso, la solución numérica al campo Gaussiano se encuentra como una solución débil a

través de ecuaciones diferenciales parciales estocásticas (SPDE). Esta solución exige definir dos nuevos parámetros, κ y τ, los cuales determinan el rango del efecto espacial y la varianza total. El rango es aproximado por la expresión $\phi \approx \sqrt{8}/\kappa$ mientras que la varianza es $\sigma_W^2 = \frac{1}{4\pi\kappa^2\tau^2}$.
Por defecto, el software INLA en R especifica una previa impropia plana en el intercepto β_0, a κ y τ especificadas por la reparametrización $\theta_1 = \log\tau$ y $\theta_2 = 2\log\kappa$ se les asignan distribuciones Gaussianas independientes. De forma predeterminada, la media para θ_2 se elige en forma razonable, en función al tamaño de la región; mientras que la media para θ_1 se elige de manera que la variación del campo sea 1. Con estas últimas consideraciones queda completa la especificación del modelo jerárquico propuesto.

4.8. Modelización de enfermedades en cultivos agrícolas

En sus orígenes la Epidemiología se limitaba al estudio de las enfermedades infecciosas, pero su método es aplicable a otras patologías e incluso a campos no estrictamente sanitarios. En la actualidad la Epidemiología se presenta como una disciplina integradora y ecléctica que estudia la enfermedad en los grupos humanos, aprovechando conceptos y métodos de otras ciencias: bioestadística, biológicas (ecología), ciencias veterinarias, sociales (demografía, antropología), económicas (gestión sanitaria), Sistemas de Información Geográficos, Defensa, etc.
No existen modelos de predicción en Epidemiología Agrícola que usen modelos jerárquicos espaciales bajo la metodología INLA. De esta forma se propone un modelo con estructura jerárquica espacial, a partir del cual se recogen y representan patrones asociados con la presencia de enfermedades en cultivos agrícolas.
En particular, sea Y_i una variable Bernoulli definida como en (4.49), que recoge la presencia (1) o ausencia (0) de enfermar de un individuo (planta

o árbol) en la localización i $(i = 1, \ldots, n)$ y sea $\pi_i = \text{logit}^{-1}(\beta_0 + W_i)$ la probabilidad de enfermar, entonces

$$\begin{aligned} Y_i &\sim \text{Ber}(\pi_\text{i}) \\ \text{logit}(\pi_i) &= \beta_0 + W_i \end{aligned} \tag{4.54}$$

Luego, π_i en (4.54) viene dada por las componentes aleatorias descritas en las ecuaciones (4.51) y (4.52).

Nuestra modelización no considera aumentar la estructura aditiva propuesta, ya que al comprobar la influencia del "ruido" (efecto pepita), se obtiene que este efecto aleatorio presenta una variabilidad muy grande. Esto demuestra que en situaciones como la analizada, este efecto no distingue fuentes de variabilidad diferentes a la espacial. Este resultado coincide con los hallazgos de los autores Roos y Held (2011), ellos demuestran que cuando se usan modelos Binomiales existe una marcada sensibilidad derivada de la elección de las previas asignadas a los parámetros de precisión que definen a los efectos aleatorios. Señalan que agregar nuevas fuentes de variabilidad puede afectar las estimaciones de los parámetros de interés.

Demostramos con nuestra propuesta que es posible aún en ausencia de covariables y sólo considerando la presencia del individuo enfermo junto con su localización geográfica, conocer el comportamiento y distribución de enfermedades en plantas. Además gracias a la metodología, será posible dibujar mapas de predicción con las probabilidades de enfermar y la estimación de la incertidumbre tanto en lugares observados como en los no observados. Con estos mapas y con las estimaciones de los parámetros involucrados en el modelo, será posible establecer estrategias y políticas de vigilancia para controlar la distribución de enfermedades en cultivos agrícolas. Este tipo de modelización fue aplicada en el contexto de pesquería con muy buenos resultados (Muñoz et al. 2012).

4.8.1. Presencia de CTV en *Citrus macrophylla*

La utilidad de la metodología propuesta, se ilustra a través de un conjunto de datos proveniente de un cultivo constituido por 10920 árboles de *Citrus macrophylla*. Este conjunto de datos proviene de un cultivo que contiene un vivero con plantas de *Citrus macrophylla*. Las 10920 plantas están distribuidas en 40 filas con 273 plantas cada una. Las plantas están dispuestas sobre 20 caballones, compuestos de dos filas de plantas cada uno. La distancia entre dos plantas cualesquiera de la misma fila está entre 15 y 18 centímetros, sin embargo, la distancia final considerada entre cada dos plantas fue el punto medio entre 15 y 18, es decir, 16.5 centímetros. Por otro lado, la distancia entre dos filas de un mismo caballón es de 40 centímetros y entre dos filas contiguas de distinto caballón es de 70 centímetros.

El análisis se realizó sobre los 10920 árboles en búsqueda del virus de la tristeza. La figura (4.1) muestra la distribución de la enfermedad en todo el cultivo y se observa un total de 443 árboles enfermos (puntos rojos), lo que representa una tasa de infección del 4.05 %. El cultivo es visto como una región continua en la que puede aparecer un árbol enfermo con el virus en cualquier punto de la misma dada la cercanía de los árboles. Esta consideración puede hacerse gracias a la gran cantidad de árboles plantados y a la baja proporción de árboles infectados con el virus de la tristeza.

La figura (4.1) representa la triangulación definida en la que se fundamenta el enfoque SPDE implementado y a partir del cual se hace el kriging Bayesiano. Cada vértice de la malla es un punto observado o un punto de predicción, los puntos rojos indican árboles infectados y los negros representan los árboles no infectados.

Al aplicar el modelo definido en (4.8) sobre el conjunto de datos observado, se encuentra la estimación de los parámetros de interés presentados en la tabla (4.1).

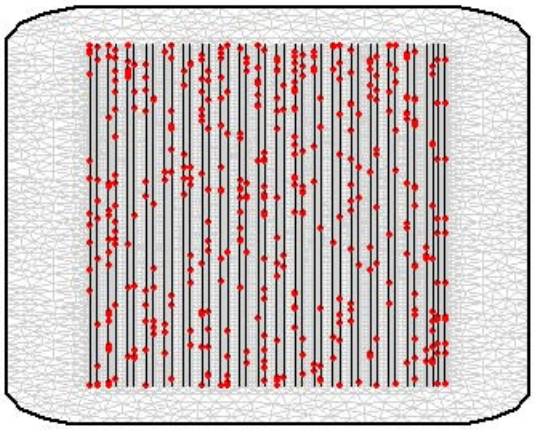


Figura 4.1: Lugares muestreados con la presencia y ausencia del virus CTV sobre la maya construida para la predicción

Parámetro	Media	Desv.típica	Cuartiles
Intercepto(β_0)	-3.13	0.31	(-3.77, -3.14, -2.45)
κ	0.00217	0.0063	(0.00074, 0.0019, 0.0049)
τ	355.46(3.55 cm)	0.0153	(1.198, 3.12, 8.38)

Tabla 4.1: Distribución posterior de los parámetros para el cultivo completo

De acuerdo a la definición de ϕ (sección 4.7) se tiene entonces que el rango es igual a 1302.869, es decir aproximadamente 13 centímetros. Ya que esta es la distancia en la cual la correlación se acerca a 0.10, se puede inferir que los datos se caracterizan por una fuerte correlación a distancias $\leq$ a los 13 centímetros. De igual forma, se puede concluir que la correlación decrece después de esta distancia. Es evidente que la presencia de la enfermedad, esta determinada claramente por el efecto espacial, en particular, el contagio se produce entre plantas ubicadas en una misma

fila a distancias menores e iguales a los 13 centímetros en cualquiera de los caballones.

La figura (4.2) muestra la media posterior del efecto espacial W_i. Se observa como la componente espacial alcanza valores positivos en la parte norte y sur del vivero, así como valores negativos y cercanos al cero en el centro. Es posible reconocer en este mapa que las zonas con mayores riesgos están en las orillas del vivero, específicamente en el lado norte y sur. Esto puede explicarse debido a la acción del viento que ingresa al vivero llevando consigo pulgones contagiados con el virus y que introducen la enfermedad en el cultivo.

La varianza del efecto aleatorio espacial viene dada por la ecuación definida en (4.8). El valor para σ^2_W en centímetros es igual a 0.13, al ser una variabilidad pequeña se tiene entonces que la componente espacial determina el patrón de contagio entre árboles próximos ubicados en una misma fila.

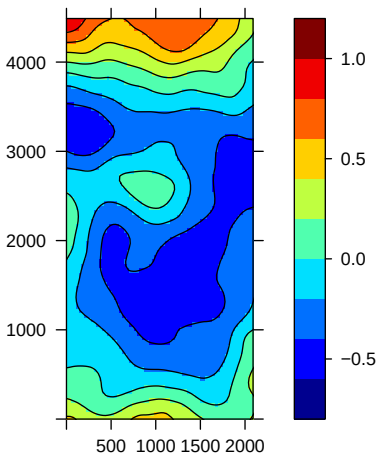


Figura 4.2: Media posterior del efecto espacial correspondiente al cultivo completo

Con el fin de entender el comportamiento del virus en el cultivo, se han generado mapas con la estimación de las probabilidades $(\pi_i|Y)$ tanto en los sitios observados como en los no observados. La figura (4.3) muestra la media posterior de la probabilidad $\pi_i|Y$, mientras que las figuras (4.4) y (4.6) muestran los cuartiles para $\pi_i|Y$. De esta forma se obtiene no sólo una estimación puntual de la probabilidad de enfermedad de un individuo sino una evaluación de la incertidumbre de esta estimación. Estas figuras confirman que la probabilidad de encontrar el virus de la tristeza es mayor en las orillas del vivero donde la influencia del viento esta presente.

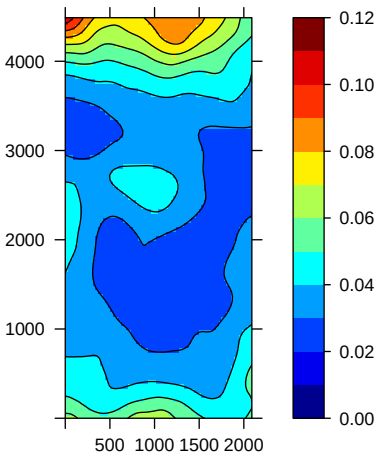


Figura 4.3: Media posterior para $\pi_i|Y$ correspondiente al cultivo completo

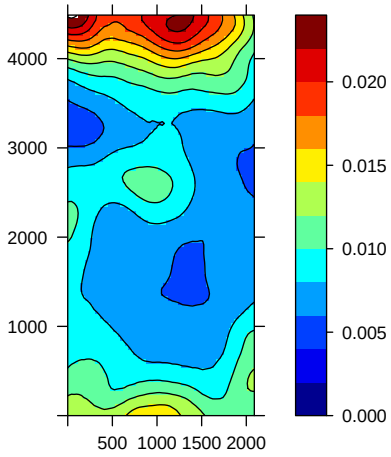


Figura 4.4: Primer cuartil para $\pi_i|Y$ correspondiente al cultivo completo

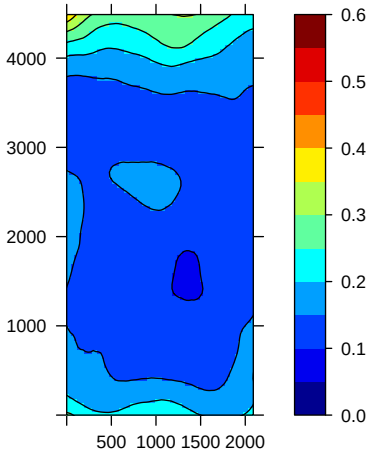


Figura 4.5: Tercer cuartil para $\pi_i|Y$ correspondiente al cultivo completo

4.8.2. Estrategias de muestreo

En algunas áreas podemos necesitar información sobre una población para determinar la prevalencia, la tasa de infección de un virus, o estar interesados en la presencia/ausencia de una enfermedad o simplemente deseamos conocer las causas o factores posibles de riesgo mediante estudios epidemiológicos. En cualquiera de estos casos, analizar a toda la población resulta muy costoso o en algunos casos imposible.

En lugar de examinar a todos los individuos de la población objeto de estudio, se prefiere medir variables en una parte de ella, es decir, obtener una muestra. Trabajar con una muestra tiene la ventaja de ser más rápido y barato. Además si la muestra es elegida correctamente, la información que se obtiene conduce a estimaciones razonables y confiables.

Cuando el objetivo es conocer la dinámica de una enfermedad o cuando se desea realizar un estudio epidemiológico cuyos resultados puedan extrapolarse a una población general, un requisito indispensable es que la muestra sea representativa. La mejor opción para obtener una muestra representativa es elegir a los individuos al azar mediante un método de muestreo aleatorio.

El muestreo aleatorio puede realizarse de distintas maneras, los métodos más frecuentes son el muestreo simple, el sistemático, el estratificado y el muestreo por conglomerados. En el muestreo aleatorio todos los individuos tienen la misma probabilidad de ser elegidos. Los elementos que forman parte de la muestra se eligen al azar mediante números aleatorios.

Los 10920 árboles analizados se cultivan con el propósito de estudiar el comportamiento del virus de la tristeza en cítricos en un entorno controlado. El ideal para lograr este propósito es analizar sólo una parte de los individuos, es decir, una muestra de todo el cultivo. Por lo tanto, en situaciones como esta es muy beneficioso determinar estrategias de muestreo que ayuden a comprender la dinámica de la enfermedad con

la menor inversión de recursos y sin necesidad de analizar a todos los árboles. En este sentido, proponemos un procedimiento de calibración para la muestra y una comparación de la eficiencia entre varios métodos de muestreo.

Calibración de la muestra

En situaciones similares a la analizada donde se cuenta con una importante cantidad de árboles a estudiar y en aras de determinar el método aleatorio más adecuado, se considera conveniente realizar como procedimiento previo a la elección del muestreo, una calibración de la muestra. A través de este proceso se podrá conocer qué zonas del cultivo tienen los mayores riesgos de infección y al mismo tiempo permitirá proponer métodos de muestreo combinados que puedan mejorar la elección de la muestra.

El proceso de calibración se inicia dividiendo el cultivo en 9 franjas horizontales, cada una de ellas, esta compuesta por 500 puntos en función a los valores de la coordenada x del árbol i con $i = 1, \ldots, n = 10920$. La figura (4.6) ilustra esquemáticamente la configuración de las franjas.

Para encontrar las probabilidades en todas las localizaciones del cultivo (punto observado, punto de predicción) se ajusta el modelo propuesto con el enfoque INLA-SPDE y se utiliza la malla o triangulación presentada en la figura (4.1). Gracias a la proyección construida sobre la triangulación fue posible concebir el proceso espacial estudiado como un proceso continuo.

El resumen con las medidas estadísticas descriptivas obtenidos a partir del proceso de calibración, se presentan en la tabla (4.2). De acuerdo a esta tabla, se observa que las mayores probabilidades o mayores riesgos de enfermar se encuentran localizados en las franjas 1 y 9 del cultivo.

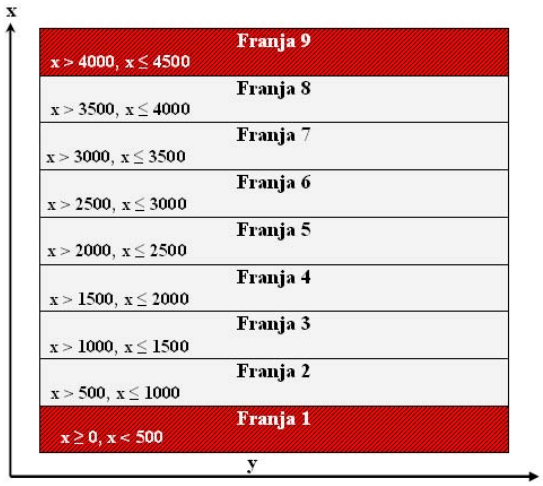


Figura 4.6: Configuración del cultivo para el proceso de calibración

	Media	Desv.típica	Cuantiles
Franja 1	0.046	0.007	(0.041, 0.046, 0.050, 0.064)
Franja 2	0.035	0.005	(0.031, 0.034, 0.039, 0.052)
Franja 3	0.029	0.003	(0.026, 0.028, 0.032, 0.041)
Franja 4	0.029	0.004	(0.026, 0.028, 0.031, 0.044)
Franja 5	0.033	0.005	(0.028, 0.031, 0.036, 0.046)
Franja 6	0.034	0.006	(0.028, 0.035, 0.039, 0.048)
Franja 7	0.030	0.003	(0.028, 0.030, 0.033, 0.038)
Franja 8	0.042	0.006	(0.038, 0.042, 0.047, 0.047)
Franja 9	0.067	0.011	(0.058, 0.067, 0.076, 0.104)

Tabla 4.2: Medidas estadísticas obtenidas a partir del proceso de calibración

Después de calibrar la muestra se determina cuál de los métodos aleatorios puede generar el mayor beneficio en función al tamaño de la muestra

y a los menores errores de predicción. Para esto se calculan medidas de discrepancia entre las probabilidades obtenidas en todo el cultivo (proyección basada en la triangulación definida en 4.1) y aquellas obtenidas bajo los diferentes métodos de muestreo. Los puntos de proyección de cada muestreo se definen sobre la triangulación definida para todo el vivero (figura 4.1). De esta forma, se pueden comparar las probabilidades en los mismos puntos de proyección.
Las medidas de discrepancia o errores de predicción se obtienen después de realizar simulaciones sucesivas de cada muestreo aleatorio. Entre las medidas definidas están, el error cuadrático medio (e.c.m.), el error absoluto (e.abs.) y el coeficiente de variación (c.v).

Resultados de los métodos de muestreo considerados

La tabla (4.3) muestra el resumen con las medidas obtenidas usando un muestreo aleatorio simple. Este método conceptualmente es el más sencillo y consiste en extraer todos los individuos al azar de una lista. En nuestro caso, la lista esta formada por los 10920 árboles originalmente observados. Con el objetivo de proponer el mejor diseño de muestreo, se prueban varios porcentajes o tamaños de muestra bajo las diferentes estrategias de muestreo analizadas. En la tabla (4.3) se presenta tanto los parámetros estimados como las diferentes medidas de error para muestras aleatorias simples del 10 %, 15 %, 20 % y 25 % consideradas.
Bajo este muestreo se tiene que los menores errores se presentan en aquellas muestras aleatorias simples del 25 %. Con el 25 % de representatividad se obtienen menores discrepancias en el error cuadrático medio y en el coeficiente de variación. Sin embargo, no presenta mejoras significativas en el error absoluto.
La tabla (4.4) presenta el resumen con las medidas obtenidas usando un muestreo sistemático. En este caso se elige el primer individuo al azar y el resto viene condicionado por esta elección. Este método es simple de

Medida	Muestra 10 %	Muestra 15 %	Muestra 20 %	Muestra 25 %
κ	0.0058	0.0084	0.0044	0.0029
ϕ(cm)	6.26	3.83	8.39	12.44
τ(cm)	19.44	11.81	6.76	4.82
e.c.m.	3.1419	2.8638	2.8448	2.4847
e.abs.	0.0134	0.0116	0.0118	0.0106
c.v	0.2025	0.2183	0.2075	0.1689

Tabla 4.3: Distribución posterior de los parámetros y errores de predicción para las muestras aleatorias simples consideradas

aplicar en la práctica y tiene la ventaja de que no hace falta disponer de un marco de encuesta. Puede aplicarse en la mayoría de las situaciones, la única precaución que debe considerarse es comprobar que la característica que estudiamos no tenga una periodicidad que coincida con la del muestreo.

En las muestras que constituyen el 50 % se ha definido un salto sistemático igual a dos, es decir se considera una de cada 2 plantas. En las muestras del 25 % se ha considerado una de cada cuatro plantas, mientras que para el resto de los porcentajes se ha elegido una de cada cinco y una de cada once respectivamente.

Es evidente que con una muestra sistemática del 50 % se obtendrán los menores errores debido al tamaño tan grande de la muestra. Sin embargo, al comparar las otras muestras sistemáticas se tiene que las muestras del 25 % son las que menores errores presentan. Las medidas de discrepancia de estas muestras comparadas con las muestras del 20 % son parecidas entre sí, sin embargo, el error cuadrático medio (e.c.m.) para estas últimas es mayor. La medida de discrepancia que reconoce mayores diferencias sigue siendo el error cuadrático medio.

Finalmente, de acuerdo a los resultados de la tabla (4.4) y en virtud de

Medida	Muestra 50 %	Muestra 25 %	Muestra 20 %	Muestra 9 %
κ	0.0032	0.0046	0.0036	0.0026
ϕ(cm)	8.58	6.03	7.70	10.81
τ(cm)	7.98	4.33	4.20	4.80
e.c.m.	2.4409	2.7979	2.9392	3.7203
e.abs.	0.0109	0.0109	0.0110	0.0130
c.v	0.1775	0.2017	0.2391	0.3313

Tabla 4.4: Distribución posterior de los parámetros y errores de predicción para las muestras sistemáticas consideradas

estudiar muestras más pequeñas y representativas, las mejores muestras bajo este esquema se obtienen al seleccionar el 25 % de todos los árboles del vivero.

En la tabla (4.6) se presenta el resumen con las medidas obtenidas usando un muestreo mixto, se combina el muestreo estratificado con el aleatorio. En este caso, se construyen bloques o estratos considerando el valor de la coordenada x de cada árbol i con $i = 1, \ldots, 10920$. Se forman tres bloques o subconjuntos de datos; el primer bloque esta conformado por aquellos individuos localizados en la franja 1 (figura 4.6), el segundo por los árboles ubicados en las franjas 2,3,4,5,6,7 y 8 (figura 4.6) y el tercer bloque esta constituido por los árboles ubicados en la franja 9 (figura 4.6). Una vez conformados los bloques, se toman muestras aleatorias de diferentes tamaños en cada uno de ellos (tabla 4.5).

Los resultados de este método se resumen en la tabla (4.6), se observa que a medida que n aumenta disminuye progresivamente la medida de error considerada. Sin embargo, las diferencias entre los errores de las últimas muestras no es significativa. Por lo tanto, en este tipo de situaciones se puede usar un muestreo combinado con porcentajes del 30 % en el bloque

	Bloque 1	Bloque 2	Bloque 3	Total
100 %	20 %	20 %	10 %	12.23 %
N	1240	1200	8480	10920
n	248	240	848	1336
100 %	25 %	25 %	15 %	16.67 %
N	1240	1200	8480	10920
n	310	300	1211	1821
100 %	30 %	30 %	20 %	22.97 %
N	1240	1200	8480	10920
n	413	400	1696	2509
100 %	35 %	35 %	25 %	30.58 %
N	1240	1200	8480	10920
n	620	600	2120	3340

Tabla 4.5: Porcentajes usados en las muestras estratificadas aleatorias consideradas

Medida	20 %,20 %,10 %	25 %,25 %,15 %	30 %,30 %,20 %	35 %,35 %,25 %
κ	0.0042	0.0031	0.0024	0.0091
ϕ(cm)	6.98	10.36	14.68	9.35
τ(cm)	9.82	3.80	3.917	2.91
e.c.m.	8.2617	4.0836	3.4858	3.0376
e.abs.	0.0195	0.0140	0.0130	0.0124
c.v	0.3985	0.3264	0.2969	0.2505

Tabla 4.6: Distribución posterior de los parámetros y errores de predicción en las muestras estratificadas aleatorias consideradas

1, 30 % en el bloque 2 y 20 % en el bloque 3 o muestras en cada bloque del 35 %, 35 % y 25 % respectivamente. Con ambos esquemas de muestreo se tendrá aproximadamente el 23 % y el 31 % de todo el vivero. Al igual que en los otros métodos de muestreo, el error cuadrático medio (e.c.m.) es quien reconoce mayores diferencias.

Estimación y predicción a partir del muestreo recomendado

Una vez evaluado el efecto de elegir muestras aleatorias bajo los distintos métodos de muestreo considerados, es posible concluir que el método más indicado cuando se manejan datos de la naturaleza estudiada, es aquel que toma muestras aleatorias simples del 25 %. Con este tipo de muestras se obtienen las menores errores de predicción. Este resultado parece adecuado ya que constituye un esquema de muestreo intermedio entre el porcentaje de muestra recomendado por Gottwald et al. (2007) y el 10 % señalado por la mayoría de literatura dedicada al muestreo.

Sólo mostraremos los mapas del efecto espacial y las incertidumbres estimadas de las probabilidades posteriores para muestras aleatorias del 25 %. Al comparar los mapas obtenidos con muestras aleatorias del 25 % (figuras 4.7,4.8) con los mapas de todo el vivero (figuras 4.2, 4.3) se observan comportamientos similares. Es decir, el patrón observado en los datos originales se conserva y se suaviza en las muestras aleatorias simples del 25 %.

El modelo es capaz de capturar el comportamiento del conjunto de datos original aún con menos muestra. Esto no sólo lo observamos con muestras simples del 25 % sino también en los mapas obtenidos a partir de muestras simples del 10 %, 15 % y 20 %. Igual comportamiento se obtuvo en los mapas generados a partir de los otros métodos de muestreo considerados.

La modelización es robusta en presencia de pocos datos y en ausencia de variables explicativas. Una vez más se reconoce la importancia del efecto aleatorio espacial y su efecto determinante en la dinámica de la enfermedad

(figura 4.7). Los mayores riesgos siguen presentándose en las orillas del cultivo, fronteras norte y sur, donde es mayor la acción del viento (figura 4.8).

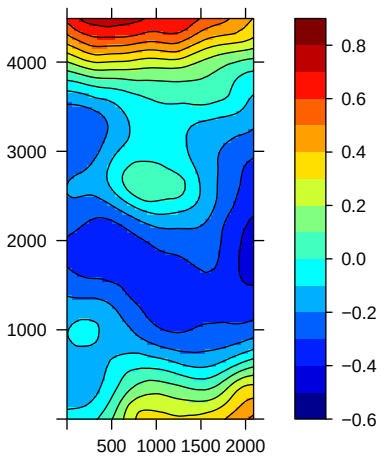


Figura 4.7: Media posterior del efecto espacial para muestras aleatorias simples del 25 %

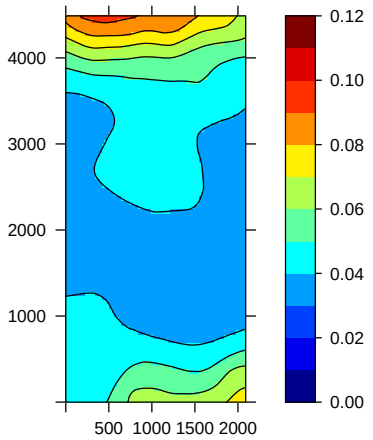


Figura 4.8: Media posterior de $\pi_i|Y$ para muestras aleatorias simples del 25 %

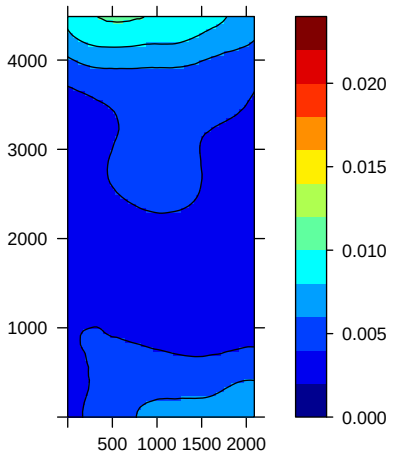


Figura 4.9: Primer cuartil para $\pi_i|Y$ correspodiente a muestras aleatorias simples del 25 %

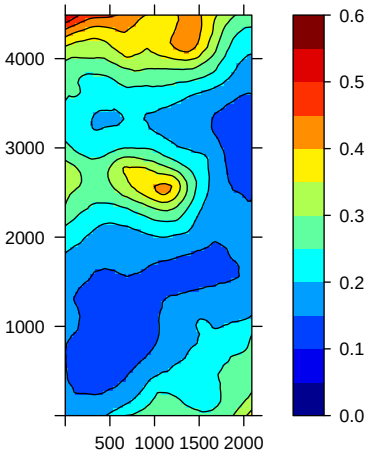


Figura 4.10: Tercer cuartil para $\pi_i|Y$ correspodiente a muestras aleatorias simples del 25 %

Siguiendo las definiciones para el rango, ϕ y para la varianza total del efecto espacial, σ_W^2 se tiene respectivamente que, la máxima correlación espacial se alcanza en 975.3197, en aproximadamente 9.75 centímetros, mientras que la variabilidad en centímetros es igual a 0.04072. Al comparar estas estimaciones con las obtenidas al usar la muestra completa del cultivo, se tiene que tanto el rango (ϕ) como la varianza espacial (σ_W^2) es menor en el caso de muestras aleatorias del 25 %. Este resultado parece lógico y es producto del método de muestreo. Sin embargo, a pesar de estas diferencias, se puede afirmar que el rango sigue describiendo la existencia de un patrón de contagio entre árboles próximos, es decir, entre árboles ubicados en una misma fila. Efecto que se repite en todos los caballones del cultivo y en especial en aquellas filas localizadas al norte y sur del cultivo. La validez de la modelización se comprueba a través del número efectivo de parámetros p_D, derivado del cálculo del DIC y con las medidas predictivas generadas a partir de los cálculos principales del INLA. En particular, la

medida predictiva evaluada fue la probabilidad integral transformada PIT definida en la sección (4.6). En todos los modelos ajustados se obtienen valores de p_D menores al número de datos considerados. Por otro lado, los histogramas de las PIT en todos los casos, mostraron un comportamiento cercano a una distribución uniforme.

4.9. Conclusiones del capítulo

Combinar el enfoque SPDE con la metodología INLA, ofrece un marco excepcional de cálculo para realizar inferencia Bayesiana en modelos complejos con estructura espacial y al mismo tiempo facilita el manejo de grandes conjuntos de datos con excelentes tiempos de computación. Esta metodología de trabajo permite construir mapas de predicción e incertidumbre de una forma relativamente sencilla y rápida. La mayor ventaja de nuestra modelización es el beneficio computacional para el ajuste y la predicción en comparación con los métodos geoestadísticos clásicos y con modelos Bayesianos definidos en WinBUGS.

Tradicionalmente en geoestadística clásica se tienen numerosos parámetros a estimar. En la mayoría de sus aplicaciones la incertidumbre no siempre esta medida y se consideran conocidos algunos de los parámetros o estimados mediante modelos estadísticos lo que termina en evaluaciones optimistas de la predicción (Diggle y Ribeiro, 2007). Usando el kriging Bayesiano es posible incorporar en el modelo fuentes de incertidumbre asociadas a los parámetros de predicción y de esta forma encontrar estimaciones más realistas.

Hemos demostrado la robustez de la metodología en presencia de pocos datos y en ausencia de variables explicativas. La utilización de la metodología INLA es posible para cualquier tipo de dato espacial inclusive puede abordar fenómenos continuos no estacionarios y anisotrópicos. Finalmente después de ilustrar la utilidad de la metodología con el conjunto

de datos como ejemplo, se pueden agregar las siguientes observaciones finales:

- En algunos casos, en especial en cultivos donde los individuos están muy cerca geográficamente, es posible estudiar el fenómeno como un proceso espacial continuo a pesar de contar con datos en localizaciones fijas.
- Se han obtenido mapas de predicción que estiman la incertidumbre de la probabilidad de enfermar tanto en lugares observados como en aquellos sin observar.
- En cultivos agrícolas con datos de naturaleza similar a los analizados, se pueden emplear muestras aleatorias simples del 25 % para estudiar fenómenos asociados a un gran número de datos.
- La metodología propuesta por Lindgren et al. (2011) ofrece un marco teórico excelente para abordar problemas en Agricultura relacionados con estudios epidemiológicos, a partir de la cual es posible predecir.
- Los resultados encontrados muestran que la distribución de la presencia del virus CTV está determinada por el efecto aleatorio espacial y por la acción del viento.
- Es importante resaltar que la modelización desarrollada y el análisis realizado puede ser extendido a otras especies de árboles y a otros cultivos.
- La metodología propuesta demuestra, que usando métodos de muestreo es posible capturar patrones de comportamiento y dibujar mapas de riesgos, similares a los que se presentan en la población de donde se ha extraído la muestra. Esta bondad es una ventaja importante, en especial en el contexto epidemiológico agrícola.

Capítulo 5

Conclusiones generales y líneas futuras

5.1. Conclusiones

A lo largo del trabajo hemos presentado en cada capítulo modelizaciones que pueden aplicarse de forma general en estudios epidemiológicos no sólo en el contexto agrícola sino en otras áreas de la ciencia. No se pretende encontrar el mejor modelo a partir de un conjunto de datos, ni proponer la mejor metodología, sólo queremos ofrecer nuevas herramientas de modelización capaces de adaptarse a problemas reales y que permitan estudiar fenómenos asociados con un proceso espacial en una red fija de localizaciones o en un espacio continuo.

En general, las distintas modelizaciones propuestas reconocen la existencia de correlación espacial a pequeña escala. Al ilustrar la metodología con datos reales, se reconoce la importancia de la variabilidad espacial y es gracias a ella que puede comprenderse la dinámica de contagio y el patrón de movilidad de los agentes causantes de la enfermedad. De cada capítulo se extraen conclusiones interesantes que a continuación mencionaremos.

Del capítulo 2 se concluye, que el uso de modelos jerárquicos Bayesianos espaciales constituyen una metodología de trabajo novedosa y capaz de

capturar fuentes de variabilidad no observadas. Gracias a la representación jerárquica del modelo basado en datos espaciales ubicados en una red de localizaciones y a los métodos MCMC es posible encontrar estimaciones posteriores de los parámetros de interés. Al usar la metodología en un ejemplo, se ha encontrado que la covariable con información del pasado que recoge la historia del contagio entre individuos es determinante en la dinámica de la enfermedad.

Los modelos con mejores ajusten contienen en su estructura no sólo el efecto de la covariable sino la influencia del efecto aleatorio espacial dinámico. El modelo con mejor ajuste además incluye en su estructura el efecto aleatorio heterogéneo, esto evidencia que la probabilidad de un individuo enfermar dependerá de un proceso espacial determinado tanto por el pasado como por el presente y por fuentes de variabilidad ajenas al proceso espacial. Con la modelización aplicada, hemos encontrado la presencia de patrones de contagio entre árboles ubicados a distancias menores a los 10 metros y de forma implícita el reconocimiento del patrón de movimiento del principal vector transmisor del virus *A. gossypii*.

En general, los modelos jerárquicos Bayesianos espaciales con estructura dinámica pueden ser herramientas muy útiles en estudios epidemiológicos en cualquier contexto ya que permiten estudiar la incidencia y extensión de fenómenos asociados a un proceso espacial. En particular, su utilidad queda demostrada en Agricultura.

El hecho de que los modelos con mejores ajustes contengan en su estructura alguno de los efectos aleatorios o ambos, coincide con lo afirmado por Leroux et al. (1999). Estos autores demostraron haciendo un estudio de simulación que si los datos son realmente independientes, un modelo con sólo efectos ICAR sin efectos de heterogeneidad, tendrá una seria sobrestimación en el parámetro de precisión del modelo ICAR. Por tanto, proponer modelos espaciales que incluyan sólo la estructura ICAR sin considerar otra fuente de variabilidad conducirá a pobres estimaciones en

los coeficientes de regresión.

Abordando la metodología de modelos jerárquicos Bayesianos con estructura espacial desde el contexto de supervivencia es posible plantear otras formas de modelizaciones. Partiendo de este punto de vista, se requiere entonces entender estadísticamente el proceso, para ello se necesitan modelos capaces de capturar heterogeneidad usualmente no observada y que generalmente no es explicada en las covariables disponibles. Pensar que los individuos son extraídos de una población homogénea, no es adecuado, especialmente en fenómenos donde existen factores de riesgo ocultos que gracias a la cercanía entre los individuos son compartidos. Por tanto parece adecuado, diseñar modelos jerárquicos que permitan tratar la heterogeneidad existente en la población en alguna de sus capas o niveles. De esta forma, aún cuando dos individuos tengan funciones de riesgo similares no serán necesariamente idénticos, a pesar de compartir el mismo vector de covariables.

La representación jerárquica de las modelizaciones propuestas en el capítulo 3 permiten manejar la heterogeneidad subyacente en cualquier fenómeno y la convierte en una metodología de trabajo novedosa y de aplicabilidad en cualquier área científica. En estudios de supervivencia, la variación espacial puede ser explicada apropiadamente a través de los *frailties*. La estructura espacial que consideramos es dinámica en el tiempo y se incluye en la función de riesgo, también conocida como función hazard ó función de intensidad, de esta forma, una vez estimados los parámetros puede ser fácilmente conocida por medio de la exponenciación y recuperada cuando así se requiera.

Las modelizaciones desarrolladas en el capítulo 3 están basadas en la noción de vecindad, que resulta más apropiada en el caso de datos agregados por áreas o ubicados en una red de localizaciones. Este enfoque se popularizó en la comunidad estadística después del artículo de Besag et al. (1991). Para conferir dependencia espacial en los modelos se adoptan

procesos autoregresivos ICAR en los *frailties* (Carlin y Banerjee, 2002). Esta elección se debe a su flexibilidad en el acomodo de la dependencia espacial y a su aplicabilidad tanto en datos espaciales continuos como en datos ubicados en una red fija de localizaciones.

Una vez ilustradas las modelizaciones propuestas en el capítulo 3, es posible señalar que la modelización basada en la distribución Weibull para tiempos discretos puede verse como una primera forma de modelar datos de supervivencia provenientes de una red de localizaciones. De las tres modelizaciones propuestas, este modelo es el que menos ventajas computacionales ofrece, sin embargo, puede considerarse como una herramienta básica para estudiar fenómenos determinados por procesos espacio-temporales, además puede ser usada para obtener hiperparámetros con los que definir las previas de otros modelos de supervivencia.

Gracias a los modelos basados en riesgos proporcionales de Cox con procesos Gamma y funciones poligonales en la función de riesgo base es posible flexibilizar la condición de proporcionalidad generalmente asumida en este tipo de modelos. Esto permite construir modelos más reales y de mayor alcance.

Al igual que obtuvimos en el capítulo 2, los modelos de supervivencia con mejores ajustes son aquellos que consideran la covariable dependiente del tiempo y el *frailty* espacial. Es importante mencionar, que bajo la modelización Weibull con tiempos discretos resulta ser mejor modelo también aquel que considera la covariable y el *frailty* espacial. A pesar de ser tres formas distintas de estimar el tiempo de supervivencia todas recogen el mismo comportamiento.

Las estimaciones bajo las dos propuestas basadas en procesos de conteo son consistentes entre sí. Las curvas de supervivencia en ambas casos presentan evoluciones similares en los riegos y ambos modelos son capaces de reconocer fuentes de variabilidad propias de cada individuo. Los resultados de las modelizaciones desarrolladas en el contexto de supervivencia

evidencian que los riegos están determinados por dos procesos bastante claros, el primero, recoge la evolución del contagio entre árboles infectados en años anteriores (efecto del pasado) y el segundo, recoge la variabilidad espacial no observada en el instante de tiempo t (efecto del presente). Por tanto se espera, que un individuo con un número importante de vecinos enfermos tenga mayor probabilidad de enfermar o menor probabilidad de supervivencia.

La modelización basada en la distribución Weibull con tiempos discretos presenta un comportamiento similar al encontrado al usar los métodos de Kaplan-Meier y Cox. Luego esta modelización puede ser vista en algún sentido como equivalente a los métodos mencionados.

La metodología discutida y evaluada en el capítulo 4, se convierte en un marco excepcional de cálculo para realizar inferencia Bayesiana en modelos complejos con estructura espacial y al mismo tiempo facilita el manejo de grandes conjuntos de datos con excelentes tiempos de computación. Esta metodología de trabajo permite construir mapas de predicción e incertidumbre de una forma relativamente sencilla y rápida. En general, la mayor ventaja de esta metodología es el beneficio computacional para el ajuste y la predicción en comparación con los métodos geoestadísticos clásicos y los modelos Bayesianos para datos geoestadísticos, ya que no requieren de métodos MCMC para la estimación.

Tradicionalmente en geoestadística clásica se tienen numerosos parámetros a estimar. En la mayoría de sus aplicaciones la incertidumbre no siempre esta medida y se consideran conocidos algunos de los parámetros o estimados mediante modelos estadísticos (Diggle y Ribeiro, 2007) lo que termina en evaluaciones optimistas de la predicción. Usando el kriging Bayesiano es posible incorporar en el modelo fuentes de incertidumbre asociadas a los parámetros de predicción y de esta forma encontrar estimaciones más realistas.

La metodología INLA combinada con el enfoque SPDE ofrece un marco

teórico excelente para fenómenos que necesitan predicción. La ilustración de la metodología con datos reales permite reconocer su utilidad en estudios epidemiológicos no sólo en el contexto agrícola. Los mapas de estimación en puntos observados y de predicción en lugares sin muestra, demuestran que la presencia de la enfermedad dependerá del efecto del viento y reconocen que la entrada del agente transmisor del virus ocurre por la acción del viento.

Es importante resaltar que para la estimación desde el enfoque Bayesiano bien sea usando métodos MCMC o bajo la metodología INLA se debe recordar que la elección de las previas puede ser crucial en la estimación de los parámetros, en especial en modelos con estructura jerárquica compleja.

Los métodos MCMC requieren de mayor tiempo para realizar las simulaciones cuando se quiere hacer predicción, mientras que INLA produce aproximaciones rápidas y precisas de la distribución posterior en menor tiempo, aún en el caso de modelos complejos. Otra ventaja de combinar la metodología INLA con el enfoque SPDE, es su generalidad, lo cual hace posible desarrollar análisis Bayesianos en forma sencilla y hacer predicción aún en presencia de pocos datos. La metodología propuesta en el Capítulo 4, es robusta incluso en ausencia de variables explicativas y puede ser utilizada con cualquier tipo de dato espacial e inclusive es capaz de abordar fenómenos continuos no estacionarios y anisotrópicos tan difíciles de manejar en otros contextos.

Otro aporte importante en el Capítulo 4, es que se demuestra, que usando métodos de muestreo es posible capturar patrones de comportamiento y dibujar mapas de riesgos, similares a los que se presentan en la población de donde se ha extraído la muestra. Esta bondad es una ventaja importante, en especial en el contexto epidemiológico agrícola.

5.2. Líneas futuras de investigación

Entre las líneas futuras se pueden mencionar algunos temas de investigación interesantes originados de las distintas modelizaciones desarrolladas que a continuación mencionamos.

Como primera línea de investigación parece interesante proponer un proceso auto-regresivo CAR-propio como alternativa a la distribución ICAR empleada en los modelos propuestos en los capítulos 2 y 3. Esta idea permitirá a diferencia de lo que sucede con la distribución ICAR, modular la dependencia espacial y no depender de la estructura espacial de los datos. La distribución CAR-propia depende de dos parámetros σ y ρ. El parámetro ρ será quien controle la estructura de correlación, así, si $\rho > 0$ individuos cercanos tomarán valores similares (dependencia espacial), si $\rho = 0$, entonces las observaciones serán independientes y cuando $\rho < 0$ individuos próximos tenderán a tomar valores opuestos (dependencia espacial negativa o de inhibición). Otra alternativa para inducir dependencia espacial diferente a la ofrecida por la distribución ICAR puede ser considerar una Distribución Auto-regresiva simultánea (SAR), en este caso, se haría la autoregresión del vector ϕ en sí mismo y no en cada una de las distribuciones condicionales de ϕ.

Conjugar las herramientas espaciales con las herramientas generalmente empleadas en la modelización temporal puede ayudar a abordar desde otro contexto, es decir, desde la modelización espacio-temporal hipótesis sobre los posibles factores de riesgo. En este sentido, se puede extender la propuesta de Besag et al. (1991) a un proceso auto-regresivo de primer orden, de esta forma las estimaciones del riesgo para cada individuo y periodo compartirán información con aquellos individuos cercanos y con el propio individuo en periodos de tiempo colindantes. La modelización resultante de la combinación de las técnicas espacio-temporales ayudará a identificar no sólo los individuos o periodos de mayor o menor riesgo sino

que se podrá explorar la interacción de ambos procesos.
Las modelizaciones propuestas reconocen dependencia espacial a pequeña escala, pero puede ser interesante agregar en los modelos una componente o superficie de tendencia que permita recoger la variabilidad a gran escala y mejorar la suavización de los riesgos.
Otra cuestión interesante que surge de los capítulos 2 y 3, es considerar a la distancia no en forma fija sino como variable aleatoria. Esto supondrá mayor cantidad de parámetros a estimar y el diseño de otros algoritmos para la estimación de los parámetros. En este caso la distribución del efecto aleatorio espacial ϕ estará condicionada a la distribución previa de la distancia.
En el contexto de análisis de datos de supervivencia, existen dos modelos de riesgo generalmente empleados, los multiplicativos y los aditivos. En el capítulo 3, sólo se plantean modelizaciones bajo la primera suposición, sin embargo, puede ser interesante estudiar la asociación entre los factores de riesgo y los tiempos de supervivencia en forma aditiva. Este tipo de modelos es más complicado que los modelos basados en riesgos proporcionales. Otra modelización interesante en el contexto de supervivencia, es plantear el modelo con función de riesgo en forma dinámica y con efectos aleatorios espaciales y no espaciales (frailties) para datos geoestadísticos (Bastos y Gamerman, 2006).
Ninguna de las modelizaciones propuestas consideró covariables observadas del fenómeno, es decir, no se consideran variables como: temperatura, gradiente del viento, tipo de suelo, altura del árbol, tipo de naranjo, etc. En los modelos del capítulo 2 y 3 se incluye una covariable y esta se obtiene a partir de los datos observados; mientras que en el modelo desarrollado en el capítulo 4 no se considera ninguna covariable. Por lo tanto, la inclusión de covariables seguramente mejorará la interpretación del fenómeno y aportará mayor generalidad a la modelización.

Bibliografía

.

Andersen, P.K. y Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *Annals of Statistics*, 10, 1100–1120.

Anselin, L. (2001b). A Companion to Theoretical Econometrics. In Baltagi, B. Blackwell, Oxford.

Anselin, L., Bongiovanni, R. y Lowenberg-DeBoer, J. (2002). «A spatial econometric approach to the economics of site-specific nitrogen management in corn production». Technical Report 02-T-2, Laboratory (REAL), University of Illinois, Urbana-Champaign, IL.

Banerjee, S. (2007). Bayesian Inference. Intermediate Bayesian Data Analysis Using WinBUGS and BRugs. University of Minnesota.

Banerjee, S., Carlin, B.P. y Gelfand, A.E. (2003a). Hierarchical modelling and anlysis for spatial data.

Banerjee, S., Carlin, B.P. y Gelfand, A.E. (2004). Hierarchical Modeling and Analysis for Spatial Data. Chapman & Hall/CRC. Monographs on Statistics & Applied Probability.

Bar-Joseph, M., Marcus, R. y Lee, R.F. (1989). The Continuous Challenge of Citrus Tristeza Virus Control. *Phytopathology*, 27, 291–316.

Barceló, M., Saez, M., Cano Serral, G., Martínez-Beneito, M.A., Martínez, J.M., Borrell, C., Ocaña Riola, R., Montoya, I., Calvo, M., López-Abente, G., Rodriguez Sanz, M., Toro, S., Alcalá, J.T., Saurina, C., Sánchez-Villegas, P. y Figueiras, A. (2008). Métodos para la suavización de indicadores de mortalidad: aplicación al análisis de desigualdades en mortalidad en ciudades del Estado Españñl (Proyecto MEDEA). *Gaceta Sanitaria*, 6, **22**, 596–608.

Bastos, L.S. y Gamerman, D. (2006). Dynamic survival models with spatial frailty. *Lifetime Data Analysis*, 4, **12**, 441–460.

Beamonte, E. y Bermúdez, J.D. (2003). A Bayesian semiparametric analysis for additive hazards models with censored observations. Test, 2, **12**, 101–117.

Bell, K.P. y Bockstael, N.E. (2000). Applying the Generalized Moments Estimation Approach to Spatial Problems Involving Microlevel Data. *Review of Economics and Statistics*, 1, **87**, 72–82.

Benedetti, R., Piersimoni, F., Bee, M. y Espa, G. (2010). Agricultural Survey Methods. John Wiley & Sons Inc, 1ª edición. ISBN 9780470743713.

Benirschka, M. y J.K., Binkley (1994). Land price volatility in a geographically dispersed market. *American Journal of Agricultural Economics*, 76, 185–195.

Berger, J.O. (1985). Statistical Decision Theory and Bayesian Analysis. Springer-Verlag, New York, 2ª edición.

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society:* Series B, 36, 192–225.

Besag, J. y Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82, 733–746.

Besag, J., York, J.C. y Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 1, **43**, 1–59.

Best, N., Ickstadt, K., Wolpert, R.L. y Briggs, D.J. (2000a). «Combining models of health and exposure data: The SAVIAH Study». Oxford University Press. 393-414.

Best, N.G., Ickstadt, K. y Wolpert, R.L. (2000). Ecological modelling of health and exposure data measured at disparate spatial scales. *Journal of the American Statistical Association*, 95, 1076–1088.

Biggeri, A., Catelan, D., Rinaldi, L., Lagazio, C. y Cringoli, G. (2006). Disease mapping in veterinary epidemiology: a Bayesian geostatistical approach. *Statistical Methods in Medical Research*4, **15**, 337–352.

Bockstael, N.E. (1996). Modeling economics and ecology: the importance of a spatial perspective. *American Journal of Agricultural Economics*, 5, **78**, 1168–1180.

Botella-Rocamora, P. (2010). Suavización Espacio-Temporal en cartografía de enfermedades. Tesis, Universitat de València, Facultat de Matemàtiques., Departament d'Estadística i Investigació Operativa.

Box, G., G. y Jenkins (1976). Time Series Analysis: Forecasting and Control. Honden Day.

Breslow, N.E. y Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 421, **88**, 9–25.

Brooks, S.P. (1998). Markov chain Monte Carlo methods and its application. *Journal The American Statistician*, 47, 69–100.

Cambra, M., Gorris, M.T., Marroquín, C., Román, M.P., Olmos, A., Martínez, M.C., Hermoso De Mendoza, A., López, A. y Navarro, L. (2000a). Incidence and epidemiology of Citrus tristeza virus in the Valencian Community of Spain. *Virus Research*, 1-2, **71**, 85–95.

Carlin, B.P. y Banerjee, S. (2002). Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion). *Bayesian Statistics 7*, 7, 45–63. Oxford University Press.

Carlin, B.P. y Louis, T.A. (2000). Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall/CRC, 2ª edición.

Casella, G. y George, E.I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46, 167–174.

Chandler, R.E., Scott, E.M., Kneib, T. y Fahrmeir, L. (2011). Statistical Methods for Trend Detection and Analysis in the Environmental Sciences. John Wiley & Sons. ISBN 10.1002/9781119991571.

Clark, I. (1979). Practical geostatistics. Applied Science Publishers. Ltd. London.

Clayton, D.G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika:* Series B, 34, 187–220.

Clayton, D.G. y Kaldor, J. (1987a). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*43, 671–681.

Clayton, D.G. y Kaldor, J.M. (1987b). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671–681.

Clifford, P. (1990). Markov random fields in statistics. In Disorder in Physical Systems. Oxford University Press. 20-32.

Congdon, P. (2007). Mixtures of spatial and unstructured effects for spatially discontinuous health outcomes. *Computational Statistics & Data Analysis*, 51, 3197–3212.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society:* Series B, 34, 187–202.

Cox, D.R. y Oakes, D. (1984). Analysis of Survival Data. London, Chapman & Hall.

Cressie, N. (1993a). Statistics for Spatial Data. Wiley, New York.

Cressie, N. y Chan, N.H. (1989). Spatial Modeling of Regional Variables. *Journal of the American Statistical Association*, 406, **84**, 393–401.

Cressie, N.A. (1993b). Statistics for Spatial Data. New York: Jhon Wiley & Sons, 2ª edición.

Czado, C., Erhardt, V., Min, A. y Wagner, S. (2007). Zero-inflated generalized Poisson models. *Statistical Modelling*, 7, 125–153.

De Santis, F. y Spezzaferri, F. (1999). Methods for robust and default Bayesian model comparison: the fractional Bayes factor approach. *International Statistical Review*, 67, 267–286.

Denison, D. y Holmes, C. (2001). Bayesian partitioning for estimating disease risk. *Biometrics*, 57, 143–149.

Diggle, P.J. (2003). Statistical Analysis of Spatial Point Patterns. Arnold, 2ª edición. London.

Diggle, P.J. y Ribeiro, P.J. (2007). Model-based Geostatistics. Springer-Verlag.

Diggle, P.J., Tawn, J.A. y Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Journal of the Royal Statistical Society:* Series C (Applied Statistics), 47, 299–350.

Earnest, A., Morgan, G., Mengersen, R., K. Louise, Richard, S. y Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (CAR) models. *International Journal of Health Geographics*, 54, **6**, 1–12.

Eidsvik, J., Martino, S. y Rue, H. (2009). Approximate Bayesian inference for spatial generalized linear mixed models. *Scandinavian Journal of Statistics*, 36, 1–22.

Fahrmeir, A., Morgan, G., Mengersen, K., Louise, R., Richard, S. y Beard, J. (2001). Multivariate Statistical Modelling Based on Generalised Linear Models. Springer, Berlin.

Ferrándiz, J., Abellán, J.J., López, A., Sanmartín, P., Vanaclocha, H., Zurriaga, O., Martínez-Beneito, M.A., Melchor, I. y Calabuig, J. (2002). «Geographical distribution of the cardiovascular mortality in Comunidad Valenciana (Spain)». GIS for Emergency preparedness and health risk reduction. D. Briggs, P. Forer, L. Jarup, R. Stern (Eds). Springer-Verlag. Capítulo 15.

Ferrándiz, J., López Quílez, A., Llopis, A., Morales, M. y Tejerizo, M.L. (1995). Spatial interaction between neighbouring counties: Cáncer mortality data in Valencia (Spain). *Biometrics*, 2, **51**, 665–678.

Finley, A.O., Banerjee, S. y McRoberts, R.E. (2009). Hierarchical spatial models for predicting tree species assemblages across large domains. *Annals of Applied Statistics*, 3, **3**, 1052–1079.

Florax, R., Folmer, H. y Rey, S.J. (2003). Specification searches in spatial

econometrics: the relevance of Hendry's methodology. *Regional Science and Urban Economics*, 33, 557–579.

Fong, Y., Rue, H. y Wakefield, J. (2010). Bayesian inference for generalized linear mixed models. *Biostatistics*, 3, **11**, 397–412.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Journal of the Royal Statistical Society:* Series C *(Applied Statistics)*, 1, **40**, 63–79.

Gamerman, D. (1997). Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference. Chapman & Hall.

Gangnon, R. y Clayton, M. (2000). Bayesian detection and modeling of spatial disease clustering. *Biometrics*, 922-935, **56**.

Garnsey, S.M. (1999). Systemic diseases. Citrus Health Management. L. W. Timmer and L. W. Duncan (Eds). 95-106.

Garrigues, S., Allardb, D., Baretc, F. y Weissd, M. (2006). Quantifying spatial heterogeneity at the landscape scale using variogram models. *Remote Sensing of Environment*, 1, **103**, 81–96.

Gaudard, M., Ramsey, P. y Stephens, M. (2006). Interactive Data Mining and Design of Experiments: the JMP®Partition and Custom Design Plataforms. *Group*, 26.

Gelfand, A. E. (1996). Model determination using sampling based methods in Markov Chain Monte Carlo in Practice. London: Chapman & Hall. W.R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds).

Gelfand, A.E., Kim, H.J., Sirmans, C.F. y Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98, 387–396.

Gelfand, A.E., Schmidt, A.M., Wu, S., Silander, J., Latimer, A. y Rebelo, A.G. (2005). Modelling species diversity through species level hierarchical modeling. *Journal of the Royal Statistical Society:* Series C (Applied Statistics), 1, **54**, 1–20.

Gelfand, A.E. y Smith, A.F.M. (1990). Sampling-based approach to calculating marginal densities. *Journal American Statistics Association*, 85, 398–409.

Gelman, A. (2003). Bugs.R: functions for calling Bugs from R. `www.stat.columbia.edu/∼gelman/bugsR/`

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Journal of the American Statistical Association*,1, **3**, 515–533.

Gelman, A., Carlin, J.B., Stern, H.S., y Rubin, D.B. (2003). Bayesian Data Analysis. London: Chapman & Hall, 2ª edición.

Gelman, A. y Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457–511.

Geman, S. y Geman, D. (1984). Stochactic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine intelligence*, 6, 721–741.

Geweke, J. (1992). Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments. *Bayesian Statistics*, *J. M. Bernardo, J. O. Berger, A. P. Dawiv, and A. F. M. Smith, eds.*4, 169–193.

Gilks, W.R., Richardson, S. y Spiegelhalter, D.J. (1996). Markov Chain Monte Carlo in Practice. Chapman & Hall/CRC, London.

Gill, R.D. (2005). Product-integration. *In Encyclopedia of Biostatistics*, 6, 4246–4250. P. Armitage and T. Colton (Eds).

Givens, G.H. y Hoeting, J.A. (2005). Computational Statistics. Wiley New Jersey.

Gottwald, T. R., Da Graça, J.V. y Bassanezi, R.B. (2007). Citrus huanglongbing: The pathogen and its impact. *Plant Health Progress*, 0906–01.

Green, P.J. y Richardson, S. (2002). Hidden Markov models and disease mapping. *Journal of the American Statistical Association*, 97, 1055–1070.

Guttorp, P. y Gneiting, T. (2006). Studies in the history of probability and statistics XLIX on the Matern correlation family. *Biometrika*, 4, **93**, 989–995.

Hartman, L. y Hössjer, O. (2008). Fast kriging of large data sets with Gaussian Markov random fields. *Computational Statistics & Data Analysis*, 5, **52**, 2331–2349.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57, 97–109.

Heagerty, P.J. y Lele, S.R. (1998). A Composite Likelihood Approach to Binary Spatial Data. *Journal of the American Statistical Association*, 443, **93**, 1099–1111.

Heidelberger, P. y Welch, P.D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, 6, **31**, 1109–1144.

Henderson, R. y Shimakura, S. (2003). A serially correlated gamma frailty model for longitudinal count data. *Biometrika*, 2, **90**, 355–366.

Henderson, R., Shimakura, S. y Grost, D. (2002). Modelling spatial variation in Leukaemia survival data. *Journal of the American Statistical Association*, 97, 965–972.

Hobert, J.P. y Casella, G. (1996). The effect of improper priors on Gibbs Sampling in hierarchical linear mixed models. *Journal of the American Statistical Association*, 436, **91**, 1461–1473.

Illian, J.B., Møller, J. y Waagepetersen, R.P. (2009). Hierarchical spatial point process analysis for a plant community with high biodiversity. *Environmental and Ecological Statistics*, 3, **16**, 389–405.

Irwin, E.G. y Bockstael, N.E. (2002). Interacting agents, spatial externalities, and the endogenous evolution of residential land use pattern. *Journal of Economic Geography*, 1, **2**, 31–54.

Isaaks, E. y Srivastava, R.M. (1989). An Introduction to Applied Geostatistics. Oxford University Press, New York, USA.

Kalbfleisch, J.D. (1978). Nonparametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society:* Series B, 40, 214–221.

Kalbfleisch, J.D. y Prentice, R.L. (2002). The Statistical Analysis of Failure Time Data. Wiley Series in Probability and Statistics, 2ª edición.

Kaplan, E. y Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53, 457–481.

Karasev, A.V., Boyko, V.P., Gowda, S., Nikolaera, O.V., Hilf, M.E., Koonin, E.V., Niblett, C.L., Cline, K., Gumpf, D.J., Lee, R.F., Garnsey, S.M. y Dawson, W.O. (1995). Complete sequence of the citrus tristeza virus RNA genome. *Virology*, 2, **208**, 511–520.

Kass, R.E. y Vaidyanathan, S.K. (1992). Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society:* Series B, 54, 129–144.

Kensall, J. E. y Wakefield, J.C. (2002). Modeling spatial variation in disease risk: a geostatistical approach. *Journal of the American Statistical Association*, 459, **97**, 692–701.

Kim, Tae-W., Valdés, J.B y Aparicio, J. (2002). Frequency and spatial characteristics of droughts in the Conchos River Basin. *Water International*, 3, **27**, 420–430.

Kleinbaum, D.G. (1995). Survival Analysis. Springer-Verlag, New York.

Kneib, T. y Fahrmeir, L. (2006). Structured additive regression for multicategorical space-time data: A mixed model approach. *Biometrics*, 1, **62**, 109–118.

Kneib, T., Müller, J. y Hothorn, T. (2008). Spatial smoothing techniques for the assessment of habitat suitability. *Environmental and Ecological Statistics*, 3, **15**, 343–364.

Knorr-Held, L. y Rasser, G. (2000). Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 13-21, **56**, 2045–2060.

Lambert, D. (1992). Zero-inflated Poisson regression, with application to detects on manufacturing. *Technometrics*, 34, 1–14.

Lawson, A., Biggeri, A., Bohning, E., D. Lesaffre, Viel, J.F. y Bertollini, R. (1999). Disease Mapping And Risk Assessment For Public Health. Wiley.

Lawson, A.B. (2006). Statistical Methods in Spatial Epidemiology. John Wiley & Sons, 2ª edición. New York.

Lawson, A.B. (2008). Bayesian Disease Mapping. Chapman & Hall/CRC.

Lawson, A.B. y Clark, A. (2002). Spatial mixture relative risk models applied to disease mapping. *Statistics in Medicine*, 21, 359–370.

Lee, Dae-J. y Durban, M. (2009). Smooth-CAR mixed models for spatial count data. *Computational Statistics & Data Analysis*, 8, **53**, 2968–2979.

Lee, Y. y Nelder, J.A. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society:* Series B, 4, **58**, 619–678.

Leroux, B.G. (2000). Modeling spatial disease rates using maximum likelihood. *Statistics in Medicine*, 19, 2321–2332.

Leroux, B.G., Lei, X. y Breslow, N. (1999). Estimation of disease rates in small areas: A new mixed model for spatial dependence. *Statistical models in epidemiology, the environment and clinical trials*, 116, 179–192. Springer, Berlin Heidelberg. Halloran M.E, Berry, D. (Eds).

Li, Y. y Ryan, L. (2001). Modelling spatial survival data using semi-parametric frailty models. *Biometrics*, 58, 287–292.

Lindgren, F., Rue, H. y Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the SPDE approach (with discussion). *Journal of the Royal Statistical Society:* Series B, 73, 423–498.

Lope, V., Pollán, M., Pérez-Gómez, B., Aragonés, N., Ramis, R., Gómez-Barroso, D. y López-Abente, G. (2006). Municipal mortality due to thyroid cancer in Spain. *BMC Public Health*, 6, 302.

Lu, H. y Carlin, B.P. (2005). Bayesian areal wombling for geographical boundary analysis. *Geographical Analysis*, 35, 265–285.

Lu, H., Reilly, C.S., Banerjee, S. y Carlin, B.P. (2007). Bayesian areal wombling via adjacency modelling. *Environmental and ecological statistics*, 14, 433–452.

Lunn, D., Spiegelhalter, D., Thomas, A. y Best, N. (2009a). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 25, **28**, 3049–3067.

Lunn, D., Spiegelhalter, D., Thomas, A. y Best, N. (2009b). Rejoinder to commentaries on The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 25, **28**, 3081–3082.

Majumdar, A., Kaye, J. y Gries, C. (2008). Hierarchical Spatial Modeling and Prediction of Multiple Soil Nutrients and Carbon Concentrations. *Communications in Statistics-Simulation and Computation*, 2, **37**, 434–453.

Maritz, J.S. y Lwin, T. (1989). Empirical Bayes Methods. Chapman & Hall, London, 2ª edición.

Matérn, B. (1960). *Meddelanden från Statens Skogsforskningsinstitut.* volumen 49. 2ª edición.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58, 1246–1266.

McKinley, T.J. (2007). Spatial survival analysis of infectious animal diseases. Doctor of philosophy in mathematics, University of Exeter. Advisor: Bailey, Trevor C. http://hdl.handle.net/10036/27033.

Metropolis, N., Rosembluth, A.W., Sosembluth, M.N., Teller, A.H. y Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21, 1087–1092.

Möller, J. (2003). Spatial Statistics and Computational Methods. Springer Verlag, New York.

Möller, J., Syversveen, A.R. y Waagepetersen, R.P. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25, 451–482.

Mostafa, A. y Ghorbal, A.B. (2011). Using WinBUGS to Cox Model with Changing from the Baseline Hazard Function. *Applied Mathematical Sciences*, 45, **5**, 2217–2240.

Muñoz, F., Pennino, M.G., Conesa, D., López-Quílez, A. y Bellido, J.M.. Estimation and prediction of the spatial occurrence of fish species using Bayesian latent Gaussian models. En prensa.

Mugglin, A.S., Carlin, B.P. y Gelfand, A.E. (2000). Fully model-based approaches for spatially misaligned data. *Journal of the American Statistical Association*, 451, **95**, 877–887.

Naes, T. (1982). The asymptotic distribution of the estimator for the regression Parameter in Cox's regression model. *Scandinavian Journal of Statistics*, 9, 107–115.

Navarro, L., Juarez, J., Pina, J.A. y Ballester, J.F. (1984). «The Citrus quarantine station in Spain». En: Timmer L.W. Doods J.A. Garnsey, S.M. (Ed.), Proceedings of the 9th Conference International Organization of Citrus Virologists, 365–370. Departament of Plant Pathology, University of California, Riverside, USA.

Nelson, G.C. y Hellerstein, D. (1997). Do roads cause deforestation? Using satellite images in econometric analysis of land use. *American Journal of Agricultural Economics*, 1, **79**, 80–88.

Pettit, L.I. (1990). The conditional predictive ordinate for the normal distribution. *Journal of the Roal Statistical Society:* Series B, 52, 175–184.

Plummer, M., Best, N., Cowles, K. y Vines, K. (2006). CODA: Convergence Diagnosis and Output Analysis for MCMC. *R News*, 1, **6**, 7–11. http://CRAN.R-project.org/doc/Rnews/.

R Development Core Team (2009). R: A Language and Environment for Statistical Computing, v. 2.9.0. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.

Raftery, A.E. y Lewis, S.M. (1992). How many iteraciones in the Gibbs sampler? *Bayesian Statistics*, 4, 763–773. Oxford University Press, J.M. Bernardo, A.F.M. Smith, A.P. Dawid and J.O. Berger (Eds).

Ramis, P.R., García-Perez, J., Pollán, M., Aragonés, N., Pérez-Gómez, B. y López-Abente, G. (2007). Modelling of municipal mortality due to haematological neoplasis in Spain. *Journal of Epidemiology and Community Health*, 61, 165–171.

Richardson, S. (2003). Spatial models in epidemiological applications. capítulo Highly Structured Stochastic Systems, 237–259. Oxford University Press. Green, P.J. and Hjort, N.L. and Richardson, S. (Eds).

Robert, C.P. y Casella, G. (1999). Monte Carlo Statistical Methods. Springer, New York.

Robert, C.P. y Smith, A.F.M. (1994). Simple condtions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms. University of Minnesota.

Román, M.P., Cambra, M., Juárez, J., Moreno, P., Duran-Vila, N., Tanaka, F.A.O., Alves, E., Kitajina, E.W., Yamamoto, P.T., Basanezi, R.B., Teixeira, D.C., Jesús-Junior, W. C., Ayres, A.J., Gimenes-Fernandes, N., Rabenstein, F., Giroto, L.F. y Bovo, J. M. (2004). Sudden death of Citrus in Brazil: A graft-transmissible bud union disease. *Plant Disease*, 88, 453–467.

Roos, M. y Held, L. (2011). Sensitivity analysis in Bayesian generalized linear mixed models for binary data. *International Society for Bayesian Analysis*, 2, **6**, 259–278.

Rue, H. y Held, L. (2005). Gaussian Markov Random Fields. Chapman & Hall/CRC.

Rue, H. y Martino, S. (2007). Approximate Bayesian inference for hierarchical Gaussian Markov random fields models. *Journal of statistical planning and inference*, 137, 3177–3192.

Rue, H., Martino, S. y Chopin, N. (2009). 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.

Rue, H. y Tjelmeland, H. (2002). Fitting Gaussian Markov random fields to Gaussian fields. *Scandinavian Journal of Statistics*, 29, 31–49.

Schwarz, G. (1978). Estimating the dimension of a model. *Journal Annals of Statistics*, 2, **6**, 461–464.

Spiegelhalter, D., Thomas, A., Best, N. y Lunn, D. (1996). BUGS 0.5: Examples Volume 1, MRC Biostatistics Unit. Institute of Public Health, Cambridge, UK.

Spiegelhalter, D.J., Best, N., Carlin, B.P. y Van der Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society:* Series B, 64, 583–639.

Spiegelhalter, D.J., Thomas, A., Best, N. y Lunn, D. (2003). WinBUGS User Manual, Version 1.4, MRC Biostatistics Unit. Institute of Public Health and Department of Epidemiology and Public Health, Imperial College School of Medicine, UK. http://www.mrc-bsu.cam.ac.uk/bugs.

Spiegelhalter, D.J., Thomas, A., Best, N.G. y Lunn, D. (2007). OpenBUGS User Manual version 3.0.2. MRC Biostatistics Unit, Cambridge, England.

Stein, M.L. (1999). Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tanner, M.A. y Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82, 528–550.

Tsiatis, A.A. (1981). A large sample study of Cox's regression model. *Annals of Statistics*, 9, 93–108.

Ugarte, M.D., Ibañez, B. y Militino, A.F. (2004). Testing for Poisson zero inflation in disease mapping. *Biometrical Journal*, 46, 526–539.

Ugarte, M.D., Ibañez, B. y Militino, A.F. (2006). Modelling riks in disease mapping. *Statistical Methods*, 15, 21–35.

Wackernagel, H. (1995). Multivariate Geostatistics: An Introduction with Applications. Springer-Verlag, Berlin.

Wakefield, J. (2009). Comments on The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 25, **28**, 3079–3080.

Werner Hartman, L. (2006). Bayesian Modelling of Spatial Data Using Markov Random Fields,With Application to Elemental Composition of Forest Soil. *Mathematical Geology*, 2, **38**, 113–133.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41, 434–449.

Whittle, P. (1963). Stochastic process in several dimensions. *Bulletin of the International Statistical Institute*, 40, 974–985.

Yokomi, R.K., Lastra, R., Stoetzel, M.B., Damsteegt, V.D., Lee, R.F., Garnsey, S.M., Gottwald, T.R., Rocha-Peña, M.A. y Niblett, C.N.

(1994). Establishment of the brown citrus aphid (Homoptera: Aphididae) in Central American and the Caribbean Basin and its transimission of citrus tristeza virus. *Journal of Economic Entomology*, 87, 1078–1085.

Apéndices

Apéndice 1: Elección de previas para los parámetros de precisión

Un problema crucial en la formulación de modelos lineales mixtos generalizados (GLMM) desde la perspectiva Bayesiana es la especificación de las distribuciones previas en los parámetros de precisión definidos en los efectos aleatorios. Lunn et al. (2009a) argumentan que la elección de previas Gamma $G(\epsilon, \epsilon)$ con valores en ϵ pequeños son generalmente inapropiadas. No obstante, Wakefield (2009) recomienda una derivación probabilística de las previas Gamma, considerando las probabilidades residuales para datos binarios. Por otro lado, Fong et al. (2010) proponen una elección particular de distribuciones Gamma como previas para las precisiones de los efectos aleatorios cuando analizan los datos del apareamiento de la salamandra.

Revisando la literatura relacionada con la sensibilidad en las estimaciones debido a las previas elegidas en los parámetros incluidos en las capas de un modelo jerárquico visto desde el enfoque Bayesiano, hemos encontrado en primer lugar, que en el caso de los coeficientes que acompañan a las covariables, es decir, β, los autores Hobert y Casella (1996) demostraron que en modelos mixtos linealmente generalizados la falta de conocimiento de este parámetro, se puede afrontar asignando una distribución Normal

con media 0 y varianza muy grande o en su defecto recomiendan asignar una distribución Inversa-Gamma con valores muy pequeños en la desviación típica que define a su varianza.

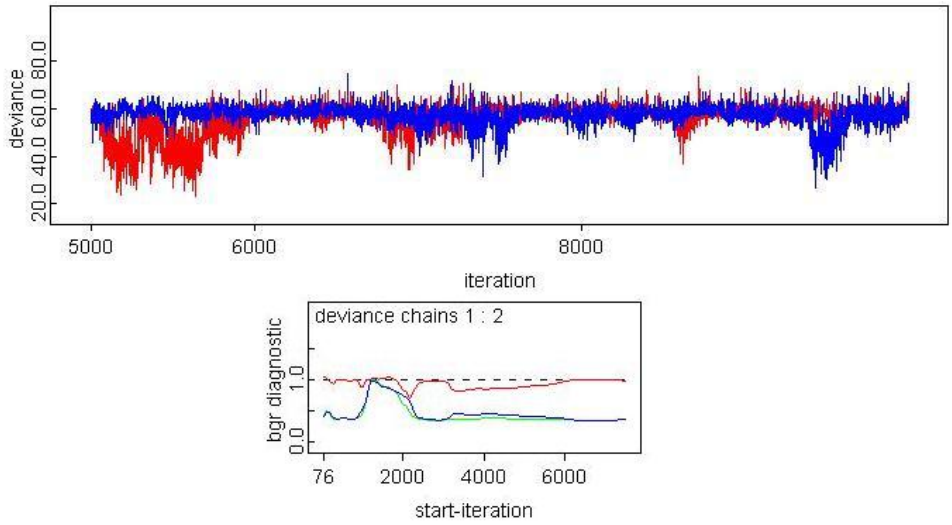
Revisando más literatura al respecto, en segundo lugar hemos encontrado que Gelman (2006) demuestra la desventaja numérica que conlleva el asignar distribuciones Gamma a los parámetros de varianza. Este autor demuestra al comparar las estimaciones cuando asigna distribuciones Gamma y cuando asigna distribuciones Uniformes a la desviación típica de parámetros que definen la varianza, que las estimaciones alcanzan mejores resultados cuando se emplean distribuciones Uniformes. En este sentido, recomienda usar distribuciones Uniformes con valores muy pequeños en sus hiperparámestros para definir dicha desviación típica.

De trabajos recientes, como el publicado por Ross y Held (2011), se sabe que se obtienen mejores estimaciones cuando se asignan distribuciones Normal-Half en los parámetros que definen la varianza. Estos autores comparan las estimaciones obtenidas al usar previas Gammas con las alcanzadas cuando asignan distribuciones Normal-Half y concluyen que las estimaciones a partir de estas últimas, son más estables y menos sensibles a la elección de los valores iniciales elegidos a los hiperparámetros que definen a estas distribuciones.

Para comprobar numéricamente el efecto de asignar previas Gamma o Uniformes sobre las estimaciones de los parámetros posteriores que definen la varianza de los efectos aleatorios, hemos hecho una comparación entre ambas. Para saber cuáles previas asignar a los parámetros de precisión (alternativamente varianza o desviación típica) se hizo un estudio comparativo basado en la deviance de los modelos propuestos en el capítulo 2 con previas Gammas y Uniformes. Se obtuvo que al asignar previas Gammas a los parámetros de precisión, los valores de la deviance eran más elevados y la estabilidad en su convergencia se alcanzaba después de 30000 iteraciones. Mientras que al usar previas Uniformes la convergencia de la

deviance se alcanzó sobre las 10000 iteraciones.

Por otra parte, se observó que al asignar previas Uniformes a la desviación típica que define la varianza del intercepto, β_0 y respectivamente al parámetro que recoge la variabilidad del coeficiente que acompaña a la covariable β_1, se encontró que sus estimaciones eran menos sensibles a los valores iniciales asignados en sus hiperparámetros que los obtenidos al usar previas Gammas. En la figura siguiente se muestran dos gráficas con la convergencia de dos cadenas para la deviance del modelo HDSM propuesto en el capítulo 2. En este caso, se asignaron previas Gammas a todos los parámetros de precisión involucrados en el modelo y se tiene que el diagnóstico propuesto por Gelman y Rubin (1992) no se supera y ambas cadenas tienen una pobre convergencia.



Después de revisar la literatura relacionada con las ventajas y desventajas de asignar una previa u otra en los parámetros e hiperparámetros de los modelos jerárquicos desarrollados en los capítulos 2 y 3, y luego de haber comprobado las diferencias en las estimaciones al comparar distribuciones previas Gamma y Uniformes, estamos de acuerdo con Gelman (2006)

y con Lunn et al. (2009b) sobre el cuidado que debe tenerse a la hora de elegir las previas que definirán a las precisiones de los efectos aleatorios. La asignación de distribuciones Uniformes con valores pequeños en los hiperparámetros que definen las precisiones permite identificar y diferenciar las variabilidades procedentes de cada efecto, evitando así el problema de no identificabilidad que suele presentarse en la modelización Bayesiana. Coincidimos con lo señalado por Gelman (2006) en cuanto a que es mejor asignar previas Uniformes con hiperparámetros pequeños en la desviación estándar que define la precisión de los efectos aleatorios, ya que se obtienen menos sesgos en las estimaciones y los datos inciden mucho más en la distribución posterior.

Apéndice 2: Sintaxis en OpenBUGS para el modelo HDSM

```
model{
for (t in 1:k[1]) {weights1[t]<-1}

for (t in 1:k[2]) {weights2[t]<-1}

for (t in 1:k[3]) {weights3[t]<-1}

for (t in 1:k[4]) {weights4[t]<-1}

phi1[1,1:N[1]]~car.normal(adj1[],weights1[],num1[],tau.phi)
phi2[1,1:N[2]]~car.normal(adj2[],weights2[],num2[],tau.phi)
phi3[1,1:N[3]]~car.normal(adj3[],weights3[],num3[],tau.phi)
phi4[1,1:N[4]]~car.normal(adj4[],weights4[],num4[],tau.phi)

for (j in 1:N[1]) {phi[1,j]<-phi1[1,j]}
phi1[1,300]<-0
for (j in 1:N[2]) {phi[2,j]<-phi2[1,j]}
for (j in 1:7) {phi[2,j+N[2]]<-0}

for (j in 1:N[3]) {phi[3,j]<-phi3[1,j]}
for (j in 1:30) {phi[3,j+N[3]]<-0}
for (j in 1:N[4]) {phi[4,j]<-phi4[1,j]}
for (j in 1:50) {phi[4,j+N[4]]<-0}
for (i in 1:4){
   for (j in 1:N[i]) {
      theta[i,j]~dnorm(0.0,tau.t)
      O[j,i]~dbern(p[j,i])
logit(p[j,i])<-beta0+(beta1*vaux[j,i])+theta[i,j]+phi[i,j]
} # for de j
```

```
sd.theta[i]<-sd(theta[i,1:N[i]])
sd.phi[i]<-sd(phi[i,1:N[i]])
alpha[i]<-sd.phi[i]/(sd.phi[i] + sd.theta[i])
} # for i

# Asignaciones de previas
tau.phi<-1/(sigma.phi*sigma.phi)
#sigma.phi: desviación típica
#para efecto espacial
tau.b<-1/(sigma.b*sigma.b)
#sigma.b: desviación típica
#para beta1
tau.t<-1/(sigma.tau.t*sigma.tau.t)
#sigma.tau.t: desviación
#típica para theta
#(efecto heterogeneidad)
beta0 ~ dnorm(0.0,0.001)
beta1 ~ dnorm(0.0, tau.b)
#Desviaciones típicas para varianza de cada parámetro
sigma.phi~dunif(0,1)
sigma.b~dunif(0,1)
sigma.tau.t~dunif(0,1)
} # end model
```

Apéndice 3: Sintaxis en OpenBUGS para el modelo WDTM

```
model{
for (w in 1:3204) {weights[w]<-1}
for (w in 1:k[1]) {weights2[w]<-1}
for (w in 1:k[2]) {weights3[w]<-1}
for (w in 1:k[3]) {weights4[w]<-1}
for (w in 1:k[4]) {weights5[w]<-1}

phi1[1,1:N]~car.normal(adj[],weights[],num[],tau)
phi2[1,1:K[1]]~car.normal(adj2[],weights2[],num2[],tau)
phi3[1,1:K[2]]~car.normal(adj3[],weights3[],num3[],tau)
phi4[1,1:K[3]]~car.normal(adj4[],weights4[],num4[],tau)
phi5[1,1:K[4]]~car.normal(adj5[],weights5[],num5[],tau)

for (j in 1:N) {phi[1,j]<-phi1[1,j]}

for (j in 1:K[1]) {phi[2,j]<-phi2[1,j]}
phi[2,300]<-0

for (j in 1:K[2]) {phi[3,j]<-phi3[1,j]}
for (j in 1:7) {phi[3,j+K[2]]<-0}

for (j in 1:K[3]) {phi[4,j]<-phi4[1,j]}
for (j in 1:30) {phi[4,j+K[3]]<-0}

for (j in 1:K[4]) {phi[5,j]<-phi5[1,j]}
for (j in 1:50) {phi[5,j+K[4]]<-0}
```

```
for (j in 1:N) {phi[6,j]<-0}
for(i in 1:N) {
      for(j in 1:t[i]) {
      lambda[i,j]<-exp(beta0+beta1*vaux[i,j]+phi[j,i])
      h[i,j]<-lambda[i,j]*(pow(j,rho)-pow(j-1,rho))
      } # fin for j interno
for(j in (t[i]+1):Nyear) {
    lambda[i,j]<-0
    h[i,j]<-0
}
 S[i]<-exp(-sum(h[i,])) # Supervivencia
 # Verosimilitud
 f[i]<-(1-exp(-h[i,t[i]]))*exp(h[i,t[i]]-sum(h[i,]))

zeros[i]<-0  # Truco de los ceros
theta[i]<-(-1)*(delta[i]*log(f[i])+(1-delta[i])*log(S[i]))
zeros[i] ~ dpois(theta[i]) # Distribución Poisson
} # fin de for i

sd.phiT<-sd(phi[,]) # marginal posterior para el
                    # efecto espacial
for(j in 1:Nyear) {sd.phi[j]<-sd(phi[j,])}
# Desv.típica marginal
# posterior del efecto
# espacial en cada año

# Asignación de previas
sigma ~ dunif(0.5,1)
tau<-1 / (sigma*sigma)
beta0~dnorm(-1,1)
beta1~dnorm(0,1)
rho~dgamma(0.1,0.1)
} # fin de modelo
```

Apéndice 4: Sintaxis en OpenBUGS para el modelo CMGPH

```
model{
for (w in 1:3204) {weights[w]<-1}
for (w in 1:k[1]) {weights2[w]<-1}
for (w in 1:k[2]) {weights3[w]<-1}
for (w in 1:k[3]) {weights4[w]<-1}
for (w in 1:k[4]) {weights5[w]<-1}

phi1[1,1:N]~car.normal(adj[],weights[],num[],tau)
phi2[1,1:K[1]]~car.normal(adj2[],weights2[],num2[],tau)
phi3[1,1:K[2]]~car.normal(adj3[],weights3[],num3[],tau)
phi4[1,1:K[3]]~car.normal(adj4[],weights4[],num4[],tau)
phi5[1,1:K[4]]~car.normal(adj5[],weights5[],num5[],tau)

for (j in 1:N) {phi[1,j]<-phi1[1,j]}

for (j in 1:K[1]) {phi[2,j]<-phi2[1,j]}
phi[2,300]<-0

for (j in 1:K[2]) {phi[3,j]<-phi3[1,j]}
for (j in 1:7) {phi[3,j+K[2]]<-0}

for (j in 1:K[3]) {phi[4,j]<-phi4[1,j]}
for (j in 1:30) {phi[4,j+K[3]]<-0}

for (j in 1:K[4]) {phi[5,j]<-phi5[1,j]}
for (j in 1:50) {phi[5,j+K[4]]<-0}
```

```
for(j in 1:T) {
 for(i in 1:N) {
  dN[i,j]     ~ dpois(Idt[i,j]) # Poisson
  # Intensidad
  Idt[i,j] <- Y[i,j]*exp(beta0+
              beta1*vaux[i,j]+phi[j,i])*dL0[j]
  # Supervivencia
  S[i,j]<- pow(exp(-sum(dL0[1:j])),exp(beta0+
           beta1*vaux[i,j]+phi[j,i]))
 }# fin for N
dL0[j] ~ dgamma(mu[j], C)
mu[j] <- dL0.star[j] * C    # prior mean hazard
} # fin for T

sd.phiT<-sd(phi[,])# Desv.típica marginal posterior
                   # para el efecto espacial

for(j in 1:T) {sd.phi[j]<-sd(phi[j,])}
# Desv.típica
# marginal posterior
# del efecto espacial
# en cada año

for (j in 1:T) {dL0.star[j] <- r * (t[j+1]-t[j])}

# Asignacion de previas
sigma ~ dunif(0.5,3)
tau<-1 / (sigma*sigma)
beta0 ~ dnorm(-1,1)
beta1 ~ dnorm(0.0,0.01)
} # fin modelo
```

APÉNDICE 5: Sintaxis en OpenBUGS para el modelo CMPFH

```
model{
eps<- 0.000001
for (w in 1:3204) {weights[w]<-1}

for (w in 1:k[1]) {weights2[w]<-1}

for (w in 1:k[2]) {weights3[w]<-1}

for (w in 1:k[3]) {weights4[w]<-1}

for (w in 1:k[4]) {weights5[w]<-1}

phi1[1,1:N]~car.normal(adj[],weights[],num[],tau.phi)
phi2[1,1:K[1]]~car.normal(adj2[],weights2[],num2[],tau.phi)
phi3[1,1:K[2]]~car.normal(adj3[],weights3[],num3[],tau.phi)
phi4[1,1:K[3]]~car.normal(adj4[],weights4[],num4[],tau.phi)
phi5[1,1:K[4]]~car.normal(adj5[],weights5[],num5[],tau.phi)

for (j in 1:N) {phi[1,j]<-phi1[1,j]}
for (j in 1:K[1]) {phi[2,j]<-phi2[1,j]}
phi[2,300]<-0
for (j in 1:K[2]) {phi[3,j]<-phi3[1,j]}
for (j in 1:7) {phi[3,j+K[2]]<-0}
for (j in 1:K[3]) {phi[4,j]<-phi4[1,j]}
for (j in 1:30) {phi[4,j+K[3]]<-0}
for (j in 1:K[4]) {phi[5,j]<-phi5[1,j]}
for (j in 1:50) {phi[5,j+K[4]]<-0}
```

```
for(j in 1:T) {
  for(i in 1:N) {
    dN[i,j]~ dpois(Idt[i,j]) # Intensidad
    Idt[i,j] <- Y[i,j]*exp(beta0+
                beta1*vaux[i,j]+phi[j,i])*dL0[j]
    # Supervivencia
    S[i,j]<- pow(exp(-sum(dL0[1:j])),exp(beta0+
             beta1*vaux[i,j]+phi[j,i]))
   } # fin for N
tau[j+1] <- tau[j]*exp(e[j+1])
e[j+1] ~ dnorm(0,1)
dL0[j] <- interp.lin(t[j],a[j],tau[j])
*step(a[j+1] - t[j] + eps)
* step(t[j] - a[j] - eps)+ tau[j]
*step(a[T+1] - t[j] + eps)
}# fin for T

# Asignacion de previas
sd.phiT<-sd(phi[,]) # Marginal posterior del
                    # efecto espacial
for(j in 1:T) {sd.phi[j]<-sd(phi[j,])}

sigma ~ dunif(0,3)
tau.phi<-1 / (sigma*sigma)

tau[1] ~ dgamma(0.01,0.01)
beta0  ~ dnorm(-1,3)
beta1  ~ dnorm(0,0.01)
} # End model
```

APÉNDICE 6: Funciones programadas en R (Capítulo 2)

Algunas pruebas estadísticas descriptivas

```
# Analisis descriptivo y grafico de la evolución
# del CTV en cada parcela estudiada

# Creacion del conjunto de datos para PARCELA:BURRIANA
bu<-matrix(scan("bu.txt"),300,5,byrow=T)

# Localizaciones muestreadas en Burriana
buxy<-matrix(scan("burrianaN.txt"),300,2,byrow=T)

#Archivo con las localizaciones muestreadas en Burriana
#buxy<-read.table("burrianaN.txt")
buxy<-matrix(scan("burrianaN.txt"),300,2,byrow=T)

bu<-cbind(bu,buxy) # Datos de BURRIANA y sus localizaciones
anyobu<-c(1994,1995,1996,1997,1998)

#Calcula la proporcion de infectados en parcela BURRIANA
apply(bu[,1:5],2,sum)
pinfbu<-apply(bu[,1:5],2,mean)

# Graficos con la configuracion del CTV en BURRIANA
burri<-function(i)
{ll<-(bu[,i]==1)
plot(bu[,7],-bu[,6],type="n",xlab="",ylab="",xaxt="n",yaxt="n")
points(bu[,7][!ll],-bu[,6][!ll],pch=".")
points(bu[,7][ll],-bu[,6][ll],pch=16, col=2)
title(paste(as.character(anyobu[i]),"
(",as.character(round(pinfbu[i]*100)),"}
```

```
par(mfrow=c(2,3),mar=c(1,1,2,1))
for (i in 1:length(anyobu)) {burri(i)}
# Grafica de incidencia en BURRIANA
plot(anyobu,pinfbu,ylim=c(0,1),xlab="Year",
  ylab="Incidence",pch=16,type="o")

#Funciones para el cálculo de las distancias
#entre los árboles de la parcela BURRIANA
dist2full <- function(dis)
{
  n <- attr(dis, "Size")
  full <- matrix(0, n, n)
  full[lower.tri(full)] <- dis
  full + t(full)
}
distabu<-dist2full(dist(bu[,6:7]))

# Calculo del Nro. de vecinos infectados
# a menos de una distancia x
# en un tiempo determinado año para BURRIANA
# En la ultima columna se almacena el numero de
# vecinos a esa distancia
nvecinbu<-function(x){
  cbind((distabu<x)%*%bu[,1:5]-bu[,1:5],
  apply((distabu<x),2,sum)-1)}
### GLM's para BURRIANA
# GLM con el efecto del Nro. de vecinos infectados a
# una determinada distancia, considerando los árboles
# enfermos del año anterior
glm.bu1<-function(x){
        glm.bu<-matrix(0,11,4)
        nvec<-nvecinbu(x)
        for(i in 1:4)
```

```
{vaux<-nvec[,i][bu[,i]==0]
m<-glm(bu[,i+1][bu[,i]==0]~vaux,family=binomial)
glm.bu[,i]<-c(coef(summary(m)),deviance(m),
summary(m)$null.deviance,m$df.residual)
}
  glm.bu
}
#calculo de los p-valores y
# significancia de la covariable = vaux
c.anyobu<-seq(6,55,2)
p.valor.glm.bu1<-sapply(c.anyobu,function(x){glm.bu1(x)[8,]})
colnames(p.valor.glm.bu1)<-c(rep("p.valor",length(c.anyobu)))
rownames(p.valor.glm.bu1)<-c("95|94","96|95","97|96","98|97")

# GLM con el efecto de la proporcion de
# árboles enfermos a una determinada distancia,
# considerando los enfermos del año anterior
glm.bu2<-function(x){
  glm.bu<-matrix(0,11,4)
  nvec<-nvecinbu(x)
  for(i in 1:4)
  {vaux<-(nvec[,i][bu[,i]==0]/nvec[,6][bu[,i]==0])*100
  m<-glm(bu[,i+1][bu[,i]==0]~vaux,family=binomial)
  glm.bu[,i]<-c(coef(summary(m)),deviance(m),
  summary(m)$null.deviance,m$df.residual)
  }
   glm.bu
}
```

```
c.anyobu<-seq(6,35,2)
#calculo de los p-valores y
# significancia de la covariable = vaux
glmbu2<-glm.bu2(c.anyobu)
p.valor.glm.bu2<-sapply(c.anyobu,function(x)
{glm.bu2(x)[8,]})
colnames(p.valor.glm.bu2)<-
c(rep("p.valor",length(c.anyobu)))
rownames(p.valor.glm.bu2) <-
c("95|94","96|95","97|96","98|97")

# Listado con Nro. de vecinos TOTALES por AÑO
# considerando distancias
# entre x e y
nvec.6<-nvecinbu(6)
nvec.10<-nvecinbu(10)
nvec.12<-nvecinbu(12)
nvec.20<-nvecinbu(20)
nvec.30<-nvecinbu(30)
nvec.40<-nvecinbu(40)
nvec.6a12<-nvec.12-nvec.6
nvec.10a20<-nvec.20-nvec.10
nvec.20a30<-nvec.30-nvec.20
nvec.30a40<-nvec.40-nvec.30
# GLM con el efecto de la proporcion de
# árboles enfermos a varias distancias,
# considerando los enfermos del año anterior
glm.bu2<-function(){
glm.bu<-matrix(0,5,4)
for(i in 1:4){
vaux.6<-(nvec.6[,i][bu[,i]==0]/nvec.6[,6][bu[,i]==0])*100
vaux.6a12<-((nvec.12[,i][bu[,i]==0]-nvec.6[,i][bu[,i]==0])/
(nvec.12[,6][bu[,i]==0]-nvec.6[,6][bu[,i]==0]))*100
```

```
vaux.10a20<-((nvec.20[,i][bu[,i]==0]-nvec.10[,i][bu[,i]==0])/
(nvec.20[,6][bu[,i]==0]-nvec.10[,6][bu[,i]==0]))*100
vaux.20a30<-((nvec.30[,i][bu[,i]==0]-nvec.20[,i][bu[,i]==0])/
(nvec.30[,6][bu[,i]==0]-nvec.20[,6][bu[,i]==0]))*100
vaux.30a40<-((nvec.40[,i][bu[,i]==0]-nvec.30[,i][bu[,i]==0])/
(nvec.40[,6][bu[,i]==0]-nvec.30[,6][bu[,i]==0]))*100
m<-glm(bu[,i+1][bu[,i]==0]~(vaux.6+vaux.6a12+vaux.10a20+
vaux.20a30+ vaux.30a40),family=binomial)
glm.bu[,i]<-c(coef(summary(m))[20],coef(summary(m))[21],
coef(summary(m))[22],coef(summary(m))[23],
coef(summary(m))[24])
}
glm.bu
} # fin función glm.bu2
#calculo de los p-valores y
# significancia de la covariable = vaux
p.valor.glm.bu3<-glm.bu2()
colnames(p.valor.glm.bu3)<-c(rep("p.valor",4))
rownames(p.valor.glm.bu3) <- c("6","6-12",
"10-20","20-30","30-40")
# Prueba de Hipotesis para medir diferencias
# significativas entre enfermos año actual y anterior
bu.test<-NULL
testbu<-function(x){
  for (i in 2:5){
  nvec<-nvecinbu(x)
  vaux<-nvec[,i][bu[,i]==0]
  bu.test<-c(bu.test,t.test(x=vaux[bu[,i-1]==0],
  y=vaux[bu[,i]==0],groups.p=T,mu=0,
  alternative="two.sided",t.paired="Two-sample t",
  var.equal=T,conf.level=0.95,print.object.p=T)$p.value)
}bu.test}
c.anyobu<-seq(6,50,4)
ttest.bu<-sapply(c.anyobu,function(x){testbu(x)})
```

Funciones para graficar los riesgos posteriores

```
# Lectura del conjunto de datos BURRIANA
bu<-matrix(scan("bu.txt"),300,5,byrow=T)

#Localizaciones muestreadas en Burriana
buxy<-matrix(scan("burrianaN.txt"),300,2,byrow=T)

bu<-cbind(bu,buxy) # BURRIANA y sus localizaciones
anyobu<-c(1994,1995,1996,1997,1998)

# Graficos con la configuracion espacial
burriana<-function(bu)
{n<-dim(bu)[2]
 plot(bu[,n],bu[,n-1],type="n",xlab="",ylab="",
 xlim=c(0,45),ylim=c(0,145))
 ll<-(bu[,1]==1)
 points(bu[,n][!ll],bu[,n-1][!ll],pch=".")
 points(bu[,n][ll],bu[,n-1][ll],pch=16,cex=0.4,col=2)
 i<-1
 while (i<=5){
 inc<-1+i
 ll<-(bu[,inc]==1)
 points(bu[,n][!ll],bu[,n-1][!ll],pch=".")
 points(bu[,n][ll],bu[,n-1][ll],pch=16,cex=0.4,col=2)
 i<-i+1
 } # fin While
} # fin function
bxy<-burriana(bu) ## Representación de la data
                  ## observada en Burriana
                  ## en sus localizaciones reales
```

```
p<-read.table("ProbMConjunto.txt") # Contiene las
# probabilidades posteriores a partir de HDSM
colnames(p)<-c("arb","t","p","sd","pc","median","tc")
p<-as.matrix(p)
datos<-matrix(p[p[,2]==4],ncol=7)
p<-datos[,3]
bu2<-read.table("BurrianaNueva.txt")
#.txt Contiene sólo los árboles
# considerando enfermos en pasado
datos<-cbind(bu2[,5:7],p)
colnames(datos)<-c("Y","x","y","p")
cortes.x<-datos[,2]
cortes.y<-datos[,3]
xunif<-(cortes.x-min(cortes.x))/(max(cortes.x)-min(cortes.x))
yunif<-(cortes.y-min(cortes.y))/(max(cortes.y)-min(cortes.y))

datos<-data.frame(datos$Y$,xunif,yunif,p)
colnames(datos)<-c("Y","x","y","p")
## Coordenada transformadas de la Parcela
## Original
cortesx<-bu[,6]
cortesy<-bu[,7]
x.unif<-(cortesx-min(cortesx))/(max(cortesx)-min(cortesx))
y.unif<-(cortesy-min(cortesy))/(max(cortesy)-min(cortesy))
# Creación de la paleta de colores de acuerdo a las
# probabilidades posteriores estimadas
require(graphics)
X<-c(datos$p)
a<-findInterval(X,sort(datos$p))
color<-NULL
cl <-heat.colors(5, alpha = 1)
```

```
for (i in 1:length(a)){
 if (datos[a[i],4]>=min(datos[,4]) & datos[a[i],4]<0.10)
 {color[i]<-cl[5]}
 else
  if (datos[a[i],4]>=0.10 & datos[a[i],4]<0.15)
  {color[i]<-cl[4]}
    else
      if (datos[a[i],4]>=0.15 & datos[a[i],4]<0.20)
      {color[i]<-cl[3]}
        else
          if (datos[a[i],4]>=0.20 & datos[a[i],4]<0.25)
          {color[i]<-cl[2]}
            else
              if (datos[a[i],4]>=0.25){color[i]<-cl[1]}
} # for
dat<-cbind(datos[,1][a],datos[,2][a],datos[,3][a],
datos[,4][a],a,color)
colnames(dat)<-c("Y","x","y","p","ind","color")
dat<-as.matrix(dat)

plot(x.unif,y.unif,type="n",,xlab="",ylab="",
xaxt="n",yaxt="n")
points(x.unif,y.unif,pch=0,cex=1.8)
#ll<-(dat[,1]==0)#arboles en riesgo en último año
for (i in 1: dim(dat)[1]){
points(dat[i,2],dat[i,3],pch=15,col=color[i],cex=1.8)
# arboles en riesgo en 98 dado enfermos en 97
} # for
points(x.unif,y.unif,pch=0,cex=1.8)
```

```
# Construcción de la leyenda para el Mapa
# con las probabilidades pposteriores estimadas
# bajo el Modelo HDSM

## # Construcción de las escalas
# para los riesgos posteriores estimados
# Grafico version Español
p.no<-datos$p[datos$Y==0]
t.riesgo<-rep(NA,length(p.no))
t.riesgo[p.no>=min(p.no) & p.no<0.10]<-"0.05<=pi<0.10"
t.riesgo[p.no>=0.10 & p.no<0.15]<-"0.10<=pi<0.15"
t.riesgo[p.no>=0.15 & p.no<0.20]<-"0.15<=pi<0.20"
t.riesgo[p.no>=0.20 & p.no<0.25]<-"0.20<=pi<0.25"
t.riesgo[p.no>=0.25]<-"pi>=0.25"
frec.riesgo<-table(t.riesgo)
## Diagrama de barras SOLO para árboles SANOS en
## el último AÑO considerado
barplot(frec.riesgo,col=c(cl[5],cl[4],cl[3],cl[2],cl[1]),
xlab="",ylab="Frecuencia")
legend(3.8,100,legend=c("riesgo bajo", "riesgo moderado",
"riesgo medio", "riesgo alto", "riesgo máximo"),
col=c(cl[5],cl[4],
cl[3],cl[2],cl[1]), text.col = "black",
lty= c(pch=0,pch=0,pch=0,pch=0,pch=0),
pch = c(15,15,15,15,15), bty = "n", cex = 1)
```

APÉNDICE 7: Funciones desarrolladas en R para el Capítulo 3

Análisis de supervivencia usando Kaplan-Meier y Cox

```
nvecinbu<-function(x){
 cbind((distabu<x)%*%bu[,1:5]-bu[,1:5])}
# Distancia considerada x<=10 metros
inf<-nvecinbu(10)
# Matriz con total de arboles enfermos para
# arbol j,considerando todos los años estudiados
tot<-apply(inf,1,sum)
inf<-cbind(inf,tot)

# conteo de árboles infectados para cada árbol,
# considerando todos los años
dput(file="totInfSuperv.txt",inf[,6])
data<-read.table("busuperv.txt", header=T)
data<-cbind(data,tot)
library(survival)
library(splines)
## Construccion del Estimador de Kaplan y Meier
attach(data)
km1<-survfit(Surv(tiempo,censor)~1)
summary(km1)
plot(km1,xlab="Year",
ylab="Survival Probabilities", main="")
```

```
# Construccion del Estimador de Kaplan y Meier,
# considerando la variable nro. de árboles
# infectados en torno a cada árbol j. Esta variable
# recoge el total de árboles infectados para
# el árbol j
#km2<-survfit(Surv(tiempo,censor)~tot)
km2<-survfit(Surv(tiempo,censor)~tot)
summary(km2)
plot(km2,xlab="Year",ylab="Survival Probabilities",main="")
## Comparar las funciones de Supervivencia en función a la
#covariable construida (nro. de árboles infectados
#correspondientes al árbol j-ésimo)
survdiff(Surv(tiempo,censor)~tot)

## Asignemos un modelo de Cox con la covariable tot
cox1<-coxph(Surv(tiempo,censor)~tot, na.action=na.exclude)
# Permite conocer la significancia del modelo usando los
# tres criterios siguientes:test de razón de
## verosimilitudes; test de Wald  y;
## test de los puntajes (score o logrank)
summary(cox1)
```

```
## Con el siguiente comando se obtiene la función de
## supervivencia ajustada mediante el modelo Cox
summary(survfit(cox1))
plot(survfit(cox1),xlab="Year",
ylab="Survival Probabilities",
main="",mark.time=FALSE)

##Comparación de la función de supervivencia obtenida
## mediante el estimador de Kaplan y Meier y la
## obtenida mediante el modelo Cox
plot(survfit(cox1),conf.int=FALSE, main="",xlab="Year",
ylab="Survival Probabilities")
#lines(km1,lty=2)
lines(km2, lty=2)
legend(0.10,0.3,legend=c("Método Cox PH con covariate",
"Método Kaplan-Meier sin covariable"),lty=c(1,2),
bty = "n", cex = 1)

# Comprobacion del supuesto de riesgos proporcionales
cox.zph(cox1)
```

Construcción de la estructura de vecindad

```
############################
## Construcción de la covariable para cada arbol-j
## considerando aquellos vecinos ubicados a
## distancias <= x  sin dejar de considerar a los
## enfermos de años pasados.
## Modelos de Supervivencia Propuestos
CoRes<-id.arbol[t.cen>0]

nvecinbu.2<-function(x,distabu.2,bu.2){
 cbind((distabu.2<x)%*%bu.2[,1]-bu.2[,1])}

x<-10 # Distancia inicial
for (t in 1:5){
    dista.t<-dist2full(dist(bu[,6:7]))
    if (t==1) {bu.t<-read.table("busup94.txt",header=T)
               attach(bu.t)
               nvinf<-nvecinbu.2(x,dista.t,
               as.matrix(arb.obs))
               vauxt94<-nvinf[,1]
               dput(file="datos94.txt",
               list(t2=as.numeric(obs.t),
               t2.cen=as.numeric(t.cen),
               CoRes2=as.numeric(id.arbol)))
               detach(bu.t)
               }
    if (t==2) {bu.t<-read.table("busup95.txt",header=T)
               attach(bu.t)
               nvinf<-nvecinbu.2(x,dista.t,
               as.matrix(arb.obs))
```

```
vauxt95<-nvinf[,1]
                dput(file="datos95.txt",
                list(t3=as.numeric(obs.t),
                t3.cen=as.numeric(t.cen),
                CoRes3=as.numeric(id.arbol)))
                detach(bu.t)
                }
if (t==3){bu.t<-read.table("busup96.txt",header=T)
attach(bu.t)
nvinf<-nvecinbu.2(x,dista.t,as.matrix(arb.obs))
vauxt96<-nvinf[,1]
dput(file="datos96.txt",list(t4=as.numeric(obs.t),
t4.cen=as.numeric(t.cen),
CoRes4=as.numeric(id.arbol)))
detach(bu.t)
}
if (t==4) {bu.t<-read.table("busup97.txt",header=T)
attach(bu.t)
nvinf<-nvecinbu.2(x,dista.t,as.matrix(arb.obs))
vauxt97<-nvinf[,1]
dput(file="datos97.txt",list(t5=as.numeric(obs.t),
t5.cen=as.numeric(t.cen),
CoRes5=as.numeric(id.arbol)))
detach(bu.t)}
if (t==5) {bu.t<-read.table("busup98.txt",header=T)
attach(bu.t)
nvinf<-nvecinbu.2(x,dista.t,as.matrix(arb.obs))
vauxt98<-nvinf[,1]
dput(file="datos98.txt",
list(obs.t=obs.t,t.cen=as.numeric(t.cen),
CoRes=as.numeric(id.arbol),vaux=vauxt98))
detach(bu.t)}}
```

```
## Creacion de las estructuras de adyacencias y de
## numero de arboles vecinos para cada arbol-j
## ubicados a distancias <= x
distmat<-as.matrix(dist(cbind(c.x,c.y)))
dist.ind<-(distmat<=x)*1
diag(dist.ind)<-0
num<-as.vector(apply(dist.ind,1,sum))

n<-length(distmat[,1])
adj<-NULL
C<-NULL
for (i in 1:n){
   neigh<-as.vector((1:n)*dist.ind[i,])
   neigh<-neigh[neigh>0]
   neigh.C<-as.vector(dist.ind[i,neigh])
   adj<-as.vector(c(adj,neigh))
   C<-as.vector(c(C,neigh.C))
}
list(adj=adj,num=num,weights=C)

totvec<-sum(num)
dput(file="vecsup.txt",list(adj=adj,
num=num,weights=C))
```

APÉNDICE 8: Funciones desarrolladas en R para el Capítulo 4

Configuración del vivero

```
colvivero <- function(n)
{ col<-matrix(0,nrow=273,ncol=10)

  cxy<-matrix(0,nrow=n,ncol=2)
  cxy2<-matrix(40,nrow=n,ncol=2)  # Caballon 1

  cxy3<-matrix(110,nrow=n,ncol=2)
  cxy4<-matrix(150,nrow=n,ncol=2) # Caballon 2

  cxy5<-matrix(220,nrow=n,ncol=2)
  cxy6<-matrix(260,nrow=n,ncol=2) # Caballon 3

  cxy7<-matrix(330,nrow=n,ncol=2)
  cxy8<-matrix(370,nrow=n,ncol=2) # Caballon 4

  cxy9<-matrix(440,nrow=n,ncol=2)
  cxy10<-matrix(480,nrow=n,ncol=2) # Caballon 5

  cxy11<-matrix(550,nrow=n,ncol=2)
  cxy12<-matrix(590,nrow=n,ncol=2) # Caballon 6

  cxy13<-matrix(660,nrow=n,ncol=2)
  cxy14<-matrix(700,nrow=n,ncol=2) # Caballon 7

  cxy15<-matrix(770,nrow=n,ncol=2)
  cxy16<-matrix(810,nrow=n,ncol=2) # Caballon 8
```

```
cxy17<-matrix(880,nrow=n,ncol=2)
cxy18<-matrix(920,nrow=n,ncol=2) # Caballon 9

cxy19<-matrix(990,nrow=n,ncol=2)
cxy20<-matrix(1030,nrow=n,ncol=2) # Caballon 10

cxy21<-matrix(1100,nrow=n,ncol=2)
cxy22<-matrix(1140,nrow=n,ncol=2) #Caballon 11

cxy23<-matrix(1210,nrow=n,ncol=2)
cxy24<-matrix(1250,nrow=n,ncol=2) # Caballon 12

cxy25<-matrix(1320,nrow=n,ncol=2)
cxy26<-matrix(1360,nrow=n,ncol=2) # Caballon 13

cxy27<-matrix(1430,nrow=n,ncol=2)
cxy28<-matrix(1470,nrow=n,ncol=2) # Caballon 14

cxy29<-matrix(1540,nrow=n,ncol=2)
cxy30<-matrix(1580,nrow=n,ncol=2) # Caballon 15

cxy31<-matrix(1650,nrow=n,ncol=2)
cxy32<-matrix(1690,nrow=n,ncol=2) # Caballon 16

cxy33<-matrix(1760,nrow=n,ncol=2)
cxy34<-matrix(1800,nrow=n,ncol=2) # Caballon 17

cxy35<-matrix(1870,nrow=n,ncol=2)
cxy36<-matrix(1910,nrow=n,ncol=2) # Caballon 18

cxy37<-matrix(1980,nrow=n,ncol=2)
cxy38<-matrix(2020,nrow=n,ncol=2) # Caballon 19
```

```
  cxy39<-matrix(2050,nrow=n,ncol=2)
  cxy40<-matrix(2090,nrow=n,ncol=2) # Caballon 20

  cxy2[1,2]<-cxy3[1,2]<-cxy4[1,2]<-cxy5[1,2]<-0
  cxy6[1,2]<-cxy7[1,2]<-cxy8[1,2]<-cxy9[1,2]<-0
  cxy10[1,2]<-cxy11[1,2]<-cxy12[1,2]<-cxy13[1,2]<-0
  cxy14[1,2]<-cxy15[1,2]<-cxy16[1,2]<-cxy17[1,2]<-0
  cxy18[1,2]<-cxy19[1,2]<-cxy20[1,2]<-cxy21[1,2]<-0
  cxy22[1,2]<-cxy23[1,2]<-cxy24[1,2]<-cxy25[1,2]<-0
  cxy26[1,2]<-cxy27[1,2]<-cxy28[1,2]<-cxy29[1,2]<-0
  cxy30[1,2]<-cxy31[1,2]<-cxy32[1,2]<-cxy33[1,2]<-0
  cxy34[1,2]<-cxy35[1,2]<-cxy36[1,2]<-cxy37[1,2]<-0
  cxy38[1,2]<-cxy39[1,2]<-cxy40[1,2]<-0
for (i in 1:n) {cxy[i,2]<-16.5*i-16.5}
cxy2[,2]<-cxy3[,2]<-cxy4[,2]<-cxy5[,2]<-c(cxy[,2])
cxy6[,2]<-cxy7[,2]<-cxy8[,2]<-cxy9[,2]<-c(cxy[,2])
cxy10[,2]<-cxy11[,2]<-cxy12[,2]<-cxy13[,2]<-c(cxy[,2])
cxy14[,2]<-cxy15[,2]<-cxy16[,2]<-cxy17[,2]<-c(cxy[,2])
cxy18[,2]<-cxy19[,2]<-cxy20[,2]<-cxy21[,2]<-c(cxy[,2])
cxy22[,2]<-cxy23[,2]<-cxy24[,2]<-cxy25[,2]<-c(cxy[,2])
cxy26[,2]<-cxy27[,2]<-cxy28[,2]<-cxy29[,2]<-c(cxy[,2])
cxy30[,2]<-cxy31[,2]<-cxy32[,2]<-cxy33[,2]<-c(cxy[,2])
cxy34[,2]<-cxy35[,2]<-cxy36[,2]<-cxy37[,2]<-c(cxy[,2])
cxy38[,2]<-cxy39[,2]<-cxy40[,2]<-c(cxy[,2])
col<-cbind(cxy,cxy2,cxy3,cxy4,cxy5,cxy6,cxy7,cxy8,cxy9,
cxy10,cxy11,cxy12,cxy13,cxy14,cxy15,cxy16,cxy17,cxy18,
cxy19,cxy20,cxy21,cxy22,cxy23,cxy24,cxy25,cxy26,cxy27,
cxy30,cxy31,cxy32,cxy33,cxy34,cxy35,cxy36,cxy37,cxy38,
cxy28,cxy29,cxy39,cxy40)
return(col)
} # fin function
# Leer archivo de datos
viv<-read.csv2('vivero.csv',header=F)
```

```
vivero<-function(cxy,viv)
{n<-dim(cxy)[2]
 plot(cxy[,n-1],cxy[,n],type="n",xlab="",ylab="",
 xlim=c(0,2040),ylim=c(0,4488))
 ll<-(viv[,1]==1)
 points(cxy[,1][!ll],cxy[,2][!ll],pch=".")
 points(cxy[,1][ll],cxy[,2][ll],pch=16,cex=0.4,col=2)
 #title("Representación de vivero analizado")
 i<-2
 while (i<=dim(viv)[2]){
        inc<-2*i
        ll<-(viv[,i]==1)
        points(cxy[,inc-1][!ll],cxy[,inc][!ll],pch=".")
        points(cxy[,inc-1][ll],cxy[,inc][ll],
        pch=16,cex=0.4,col=2)
        i<-i+1
 } # fin While
} # fin function

cxy<-colvivero(273)
vivero(cxy,viv)  ## Representación de
                 ## data observada

#Representación
vivinf<-function(cxy,viv)
{n<-dim(cxy)[2]
 plot(cxy[,n-1],cxy[,n],type="n",xlab="",ylab="",
 xlim=c(0,2040),ylim=c(0,4488))
 ll<-(viv[,1]==1)
 points(cxy[,1][ll],cxy[,2][ll],pch=16,cex=0.4,col=2)
 title("Representación de plantas infectadas")
```

```
i<-2
 while (i<=dim(viv)[2]){
        inc<-2*i
        ll<-(viv[,i]==1)
        points(cxy[,inc-1][ll],cxy[,inc][ll],pch=16,
        cex=0.4,col=2)
        i<-i+1
 } # fin While
} # fin function

dev.print(pdf, file="Vivero.pdf")
```

Aplicación de INLA y SPDE

```
# Uso de INLA y SPDE con el modelo Binomial
# para las plantas en el vivero analizado
require(INLA)
require(rgl)
require(lattice)
require(sp)
require(pixmap)
require(Matrix)
require(orthopolynom)
# Construye la configuracion
# espacial de las localizaciones
# de todas las plantas del vivero estudiado
source("Configuracionviv.r")
coordxy<-colvivero(273)
source("utils.R")
library(fields)
# Construcción de la data
dat1<-cbind(viv[,1],coordxy[,1:2])
dat2<-cbind(viv[,2],coordxy[,3:4])
dat3<-cbind(viv[,3],coordxy[,5:6])
dat4<-cbind(viv[,4],coordxy[,7:8])
dat5<-cbind(viv[,5],coordxy[,9:10])
dat6<-cbind(viv[,6],coordxy[,11:12])
dat7<-cbind(viv[,7],coordxy[,13:14])
dat8<-cbind(viv[,8],coordxy[,15:16])
dat9<-cbind(viv[,9],coordxy[,17:18])
dat10<-cbind(viv[,10],coordxy[,19:20])
dat11<-cbind(viv[,11],coordxy[,21:22])
dat12<-cbind(viv[,12],coordxy[,23:24])
dat13<-cbind(viv[,13],coordxy[,25:26])
dat14<-cbind(viv[,14],coordxy[,27:28])
dat15<-cbind(viv[,15],coordxy[,29:30])
```

```
dat16<-cbind(viv[,16],coordxy[,31:32])
dat17<-cbind(viv[,17],coordxy[,33:34])
dat18<-cbind(viv[,18],coordxy[,35:36])
dat19<-cbind(viv[,19],coordxy[,37:38])
dat20<-cbind(viv[,20],coordxy[,39:40])
dat21<-cbind(viv[,21],coordxy[,41:42])
dat22<-cbind(viv[,22],coordxy[,43:44])
dat23<-cbind(viv[,23],coordxy[,45:46])
dat24<-cbind(viv[,24],coordxy[,47:48])
dat25<-cbind(viv[,25],coordxy[,49:50])
dat26<-cbind(viv[,26],coordxy[,51:52])
dat27<-cbind(viv[,27],coordxy[,53:54])
dat28<-cbind(viv[,28],coordxy[,55:56])
dat29<-cbind(viv[,29],coordxy[,57:58])
dat30<-cbind(viv[,31],coordxy[,59:60])
dat31<-cbind(viv[,32],coordxy[,61:62])
dat32<-cbind(viv[,33],coordxy[,63:64])
dat33<-cbind(viv[,34],coordxy[,65:66])
dat34<-cbind(viv[,35],coordxy[,67:68])
dat35<-cbind(viv[,36],coordxy[,69:70])
dat36<-cbind(viv[,37],coordxy[,71:72])
dat37<-cbind(viv[,38],coordxy[,73:74])
dat38<-cbind(viv[,39],coordxy[,75:76])
dat39<-cbind(viv[,40],coordxy[,77:78])
dat40<-cbind(viv[,40],coordxy[,79:80])
data<-rbind(dat1,dat2,dat3,dat4,dat5)
data<-rbind(data, dat6,dat7,dat8,dat9,dat10)
data<-rbind(data, dat11,dat12,dat13,dat14,dat15)
data<-rbind(data, dat16,dat17,dat18,dat19,dat20)
data<-rbind(data, dat21,dat22,dat23,dat24,dat25)
data<-rbind(data, dat26,dat27,dat28,dat29,dat30)
data<-rbind(data, dat31,dat32,dat33,dat34,dat35)
data<-rbind(data, dat36,dat37,dat38,dat39,dat40)
cxy<-dim(data)[1]
```

```
data<-data.frame(data)
names(data)<-c("Y","y","x")

formula <- data$Y ~ 1 + f(spatial, model=spde)

mesh.dummy<-inla.mesh.create(loc=matrix(c(0,0, 2090,0,
            2090,4488, 0,4488), 4,2,byrow=TRUE),
            refine=FALSE)

boundary<- inla.mesh.boundary(mesh.dummy)

mesh.v = inla.mesh.create(cbind(data$y,data$x,0),
                          boundary=boundary,
                          #extend=TRUE,
                          refine=list(max.edge=150))

## Graficar el Mesh junto con los
# datos observados
plot(mesh.v,col="lightgrey")
for (i in 1:dim(data)[1]){
        ll<-(data[i,1]==1)
        points(data[i,2][ll],data[i,3][ll],pch=16,
        cex=0.4,col=2)
        points(data[i,2][!ll],data[i,3][!ll],pch=".",
        cex=0.1, col=1)
}

## Create the SPDE/GMRF model,
## (kappa^2-Delta)(tau x) = W:

data$spatial <- mesh.v$idx$loc
spde = inla.spde.create(mesh.v, model="matern",
param=list(alpha=2))
```

```
### Se crean todas lsa combinaciones
### lineales posibles
all.lc <- c()
for(i in 1:dim(data)[1])
{
  lc <- inla.make.lincomb("(Intercept)"=1,
  spatial=c(rep(NA, i-1), 1))
  names(lc) <- paste('lc',
  formatC(i, flag='0', width=3), sep='')
  all.lc <- c(all.lc, lc)
}
r <- inla(formula, family="binomial", Ntrials=1,
      data = data, lincomb=all.lc,control.compute=
      list(return.marginals=TRUE,
      dic=TRUE, cpo=TRUE),control.predictor=
      list(compute=TRUE),
      control.inla=list(lincomb.derived.only=TRUE))

#En la matriz report$summary.hyperpar se
#retorna la media posterior estimada para el log(kappa^2)
#Asi que para obtener la media posterior
#de kappa es necesario hacer lo siguiente:
kappa.marg<-inla.tmarginal(function(x) exp(x)^0.5,
r$marginals.hyperpar$"K.0 for spatial-basisK")
kappa.m1<-inla.emarginal(function(x) x, kappa.marg)
```

```
rho<- sqrt(8*1)/kappa.m1
## da 1302.869 = aprox. 13 cms
tau.marg<-inla.tmarginal(function(x) exp(x),
r$marginals.hyperpar$"T.0 for spatial-basisT")
tau.m1<-inla.emarginal(function(x) x, tau.marg)
# da 355.4694 = 3.55 cms
## calculo de probabilidades a partir
### del mesh del vivero completo
proj.v<-inla.mesh.projector(mesh.v,dims=c(110,110))
### Estimacion de la media posterior del
### Efecto Espacial
pdata<-inla.mesh.project(proj.v,
r$summary.random$spatial[,"mean"])

map.plot(pdata,proj.v)
trellis.focus("panel",1,1,highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="MediaEfEsVivero.pdf")
###### Probab. posteriores para Muestra completa
pdata<-r$summary.fixed['(Intercept)',"mean"]+
        r$summary.random$spatial[,"mean"]
e<-exp(pdata)
pdata<-e/(1+e)
pdata<-inla.mesh.project(proj.v,pdata)
map.plot(pdata,proj.v,at=seq(0,0.12,0.01))
trellis.focus("panel",1,1,highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="MediaProbVivero.pdf")
```

```
### Cuartiles para Vivero Completo
### Cuartil 1
pdata<-r$summary.fixed['(Intercept)',"0.025quant"]+
        r$summary.random$spatial[,"0.025quant"]

e<-exp(pdata)
pdata<-e/(1+e)
pdata<-inla.mesh.project(proj.v,pdata)

map.plot(pdata,proj,at=seq(0,0.025,0.002))

trellis.focus("panel",1,1,highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="Q1ProbVivero.pdf")

### Cuartil 3
pdata<-r$summary.fixed['(Intercept)',"0.975quant"]+
        r$summary.random$spatial[,"0.975quant"]

e<-exp(pdata)
pdata<-e/(1+e)
pdata<-inla.mesh.project(proj.v,pdata)

map.plot(pdata,proj.v,at=seq(0,0.60,0.05))

trellis.focus("panel",1,1,highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="Q3ProbVivero.pdf")
```

```
### Graficar Mapa con las predicciones
map.plot <- function(pdata,p,palette=my.palette, ...)
{
bbb=(levelplot(row.values=p$x, column.values=p$y,
 x=pdata, col.regions=tim.colors(64),ylim=c(0,4488),
 xlim=c(0,2090),aspect="iso",contour=TRUE, cuts=11,
 labels=FALSE, pretty=TRUE,xlab='',ylab='', ...))
print(bbb)}

### Funciones para medidas de error
ecm<-function(d){
   error<-sum(d^2)
return(error)
} # fin function
eabs<-function(d){
   error<-sum(abs(d))/length(d)
return(error)
} # fin function
cv<-function(pb){
   c.v<-sd(pb)/mean(pb)
return(c.v)
} # fin function

### Discrepancias
error<-function(p,d){
     e.c.m<-ecm(d)
     e.abs<-eabs(d)
     c.v<-cv(p)
     e<-cbind(e.c.m,e.abs,c.v)
return(e)
} # fin function
```

```
#### PROBANDO ESQUEMAS DE MUESTREO Y
### SU IMPACTO
#### EN la probabilidad p

### Comencemos probando CON EL MUESTREO
### aleatorio
muest.alet<-function(m,dat){
 n<-ceiling(dim(dat)[1]*m)
 planta<-rep(0,n)
 s<-sample(dim(dat)[1],n)
 planta<-dat[s,]
return(planta)
} # fin function

muest.simples<-function(data,porc){
   muest<-muest.alet(porc,data[,1:3])
   datos<-muest
   datos<-data.frame(datos)
   row.names(datos)<-NULL
return(datos)
} # fin function muest.simples
r.inla<-function(datos,mesh){
   formula <- datos$Y ~ 1 +  f(spatial, model=spde)
   datos$spatial <- mesh$idx$loc
  ## Create the SPDE/GMRF model,
  ## (kappa^2-Delta)(tau x) = W:
   spde = inla.spde.create(mesh, model="matern",
   param=list(alpha=2))
```

```
  ### Se crean todas las combinaciones lineales posibles
   all.lc <- c()
   for(i in 1:dim(datos)[1])
    {lc <- inla.make.lincomb("(Intercept)"=1,
     spatial=c(rep(NA, i-1), 1))
     names(lc) <- paste('lc', formatC(i, flag='0',
     width=3), sep='')
     all.lc <- c(all.lc, lc)}

    r<- inla(formula, family="binomial", Ntrials=1,
       data = datos, lincomb=all.lc,
       control.compute=list(return.marginals=TRUE,
       dic=TRUE, cpo=TRUE),control.predictor=
       list(compute=TRUE),
       control.inla=list(lincomb.derived.only=TRUE))
return(r)
} # fin fuction r.inla

probabilidad<-function(r.in,mesh.v,mesh){
 proj<-inla.mesh.projector(mesh,dims=c(110,110))
 pdata.a<-inla.mesh.project(proj,
 r.in$summary.fixed['(Intercept)',"mean"]+
 r.in$summary.random$spatial[,"mean"])

 e.a <-exp(pdata.a)
 p.a <-e.a/(1+e.a)
 prob.a <-cbind(p.a[mesh.v$idx$loc])
 prob.a <-data.frame(prob.a)
 names(prob.a) <-c("p")
return(prob.a)
} # fin function que calcula las probabilidades en los
  # puntos del proyección del mesh
```

```
## Triangulacion para cada muestra aleatoria
porc<-0.25
datos<-muest.simples(data,porc)
mesh <- inla.mesh.create(cbind(datos$y,datos$x,0),
        boundary=boundary,
        refine=list(max.edge=150))

## Obtención del modelo ajustado con INLA usando
## los datos de cada muestra aleatoria generada
r.muestra<-r.inla(datos,mesh)

### Calculo del rango y tau para muestra aleatoria
kappa.m.a<-inla.tmarginal(function(x) exp(x)^0.5,
r.muestra$marginals.hyperpar$"K.0 for spatial-basisK")
kappa.m.1<-inla.emarginal(function(x) x, kappa.m.a)

rho.m.a<- sqrt(8*1)/kappa.m.1

tau.marg.a<-inla.tmarginal(function(x) exp(x),
r.muestra$marginals.hyperpar$"T.0 for spatial-basisT")
tau.m.1<-inla.emarginal(function(x) x, tau.marg.a)

# Calculo de probabilidades en funcion a la
muestra aleatoria
prob.m<-probabilidad(r.muestra,mesh.v,mesh)
```

```
##### Probabilidades para Datos Originales
pdata = r$summary.fixed['(Intercept)',"mean"]+
        r$summary.random$spatial[,"mean"]

e<-exp(pdata)
p<-e/(1+e)
prob<-cbind(p[mesh.v$idx$loc])
prob<-data.frame(prob)
names(prob)<-c("p")

prob$p[is.na(prob.m$p)]<-0
p<-prob$p[prob$p>0]

prob.m$p[is.na(prob.m$p)]<-0
# Elimina los puntos que son NA en
p.a<-prob.m$p[prob.m$p>0]
# cada muestreo aleatorio

### Diferencia entre las probabilidades
### calculadas a partir de los meshs: data
### original y data de cada muestra creada
### de acuerdo al muestreo seleccionado

dif<-p-p.a

e.1<-error(p.a,dif)
#discrepancias muestras aleatorias simples

### Media posterior del efecto espacial junto
#### la estimación de las probabilidades posteriores
#### específicamente para muestras aleatorias del 25%
pdata.a<-inla.mesh.project(proj,
r.m.alet$summary.random$spatial[,"mean"])
```

```
map.plot(pdata.a, proj)

trellis.focus("panel",1,1, highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="MediaEfEspM.aletoria25Porc.pdf")

###### Probab. posteriores para M.aleatoria del 25%
pdata.a<-r.m.alet$summary.fixed['(Intercept)',"mean"]+
                    r.m.alet$summary.random$spatial[,"mean"]

e<-exp(pdata.a)
pdata.a<-e/(1+e)
pdata.a<-inla.mesh.project(proj,pdata.a)

map.plot(pdata.a, proj, at=seq(0,0.12,0.01))


trellis.focus("panel",1,1, highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="MediaProbM.aletoria25Porc.pdf")

### Cuartiles para Muestras aleatorias del 25%
### Cuartil 1
pdata.a<-r.m.alet$summary.fixed['(Intercept)',"0.025quant"]+
r.m.alet$summary.random$spatial[,"0.025quant"]


e<-exp(pdata.a)
pdata.a<-e/(1+e)
pdata.a<-inla.mesh.project(proj,pdata.a)
```

```
map.plot(pdata.a,proj,at=seq(0,0.025,0.002))

trellis.focus("panel",1,1,highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="Q1ProbM.aletoria25Porc.pdf")

### Cuartil 3
pdata.a<-r.m.alet$summary.fixed['(Intercept)',
"0.975quant"]+r.m.alet$summary.random$spatial[,"0.975quant"]

e<-exp(pdata.a)
pdata.a<-e/(1+e)
pdata.a<-inla.mesh.project(proj,pdata.a)

map.plot(pdata.a,proj,at=seq(0,0.60,0.05))

trellis.focus("panel",1,1,highlight=FALSE)
trellis.unfocus()
dev.print(pdf, file="Q3ProbMaleatoria25P.pdf")

#### APLICACION DE Muestreo SISTEMATICO
muest.sist<-function(salto,viv){
  dat.sist<-viv[((1:273)%%(salto+1))==salto,]
  dat.sist<-data.frame(cbind(dat.sist,
  seq(salto,(273-(salto-1)),by = salto+1)))
return(dat.sist)
} # fin function

sist.1<-muest.sist(1,viv)  # muestra del 50%
sist.2<-muest.sist(3,viv)  # aprox. muestra del 25%
sist.3<-muest.sist(4,viv)  # aprox. muestra del 20%
sist.4<-muest.sist(10,viv) # aprox. muestra del 9%
```

```
conf.muestra.sist<-function(sist,cxy){
dat<-d<-NULL
 for (j in 1:dim(sist)[1]){
   i<-1
   loc.sist<-sist[j,41]
   while (i<=40){
     inc<-2*i
     id.xy<-c(cxy[loc.sist,inc-1],cxy[loc.sist,inc])
     d<-c(sist[j,i],id.xy)
     dat<-rbind(dat,d)
     i<-i+1
  } #fin while
} # fin for j
return(dat)
} # fin function

dat.sist.1<-conf.muestra.sist(sist.1,cxy)
dat.sist.2<-conf.muestra.sist(sist.2,cxy)
dat.sist.3<-conf.muestra.sist(sist.3,cxy)
dat.sist.4<-conf.muestra.sist(sist.4,cxy)

datos<-data.frame(dat.sist.1) # muestra del 50%
datos<-data.frame(dat.sist.2) # muestra del 25%
datos<-data.frame(dat.sist.3) # muestra del 20%
datos<-data.frame(dat.sist.4) # muestra del 9%

row.names(datos)<-NULL
names(datos)<-c("Y","y","x")
```

```
### En esta sección se usará la primera
### muestra sistemática para
### generar mapas de predicción

formula <- datos$Y ~ 1 +  f(spatial, model=spde)

mesh <- inla.mesh.create(cbind(datos$y,datos$x,0),
        boundary=boundary,
        refine=list(max.edge=150))

datos$spatial <- mesh$idx$loc

## Create the SPDE/GMRF model, (kappa^2-Delta)(tau x) = W:
spde = inla.spde.create(mesh, model="matern",
param=list(alpha=2))

### Se crean todas lsa combinaciones
### lineales posibles
all.lc <- c()
for(i in 1:dim(datos)[1])
{
  lc <- inla.make.lincomb("(Intercept)"=1,
  spatial=c(rep(NA, i-1), 1))
  names(lc) <- paste('lc',
  formatC(i, flag='0', width=3), sep='')
  all.lc <- c(all.lc, lc)
}
```

```
#### APLICACION DE Muestreo SISTEMATICO
muest.sist<-function(salto,viv){
  dat.sist<-viv[((1:273)%%(salto+1))==salto,]
  dat.sist<-data.frame(cbind(dat.sist,
  seq(salto,(273-(salto-1)),by = salto+1)))
return(dat.sist)
} # fin function

sist.1<-muest.sist(1,viv)  # muestra del 50%
sist.2<-muest.sist(3,viv)  # aprox. muestra del 25%
sist.3<-muest.sist(4,viv)  # aprox. muestra del 20%
sist.4<-muest.sist(10,viv) # aprox. muestra del 9%

conf.muestra.sist<-function(sist,cxy){
dat<-d<-NULL
 for (j in 1:dim(sist)[1]){
   i<-1
   loc.sist<-sist[j,41]
   while (i<=40){
     inc<-2*i
     id.xy<-c(cxy[loc.sist,inc-1],cxy[loc.sist,inc])
     d<-c(sist[j,i],id.xy)
     dat<-rbind(dat,d)
     i<-i+1
  } #fin while
} # fin for j
return(dat)
} # fin function
dat.sist.1<-conf.muestra.sist(sist.1,cxy)
dat.sist.2<-conf.muestra.sist(sist.2,cxy)
dat.sist.3<-conf.muestra.sist(sist.3,cxy)
dat.sist.4<-conf.muestra.sist(sist.4,cxy)
```

```
datos<-data.frame(dat.sist.1) # muestra del 50%
datos<-data.frame(dat.sist.2) # muestra del 25%
datos<-data.frame(dat.sist.3) # muestra del 20%
datos<-data.frame(dat.sist.4) # muestra del 9%

row.names(datos)<-NULL
names(datos)<-c("Y","y","x")

### En esta sección se usará la primera
### muestra sistemática para
### generar mapas de predicción

formula <- datos$Y ~ 1 + f(spatial, model=spde)

mesh <- inla.mesh.create(cbind(datos$y,datos$x,0),
        boundary=boundary,
        refine=list(max.edge=150))

datos$spatial <- mesh$idx$loc

## Create the SPDE/GMRF model, (kappa^2-Delta)(tau x) = W:
spde = inla.spde.create(mesh, model="matern",
param=list(alpha=2))
```

```
### Se crean todas lsa combinaciones
### lineales posibles
all.lc <- c()
for(i in 1:dim(datos)[1])
{
  lc <- inla.make.lincomb("(Intercept)"=1,
  spatial=c(rep(NA, i-1), 1))
  names(lc) <- paste('lc',
  formatC(i, flag='0', width=3), sep='')
  all.lc <- c(all.lc, lc)
}

e.a <-exp(pdata.a)
p.a <-e.a/(1+e.a)
prob.a <-cbind(p.a[mesh.v$idx$loc],datos$y,datos$x)
prob.a <-data.frame(prob.a)
names(prob.a) <-c("p","y","x")

prob.a$p[is.na(prob.a$p)]<-0
prob$p[is.na(prob.a$p)]<-0

### Diferencia entre las probabilidades calculadas
### a partir de los meshs: data original y data
### de cada muestra creada de acuerdo
### al muestreo seleccionado

dif<-prob$p-prob.a$p
```

```
### calculo de los errores de predicción
### para muestras sistemáticas
e.s.1<-error(prob.a$p,dif)
#discrepancias muestra sistemática 50%
e.s.2<-error(prob.a$p,dif)
#discrepancias muestra sistemática 25%
e.s.3<-error(prob.a$p,dif)
#discrepancias muestra sistemática 20%
e.s.4<-error(prob.a$p,dif)
#discrepancias muestra sistemática 9%
### AQUI termina el análisis para la
### MUESTRA sistemática

#PROCESO DE CALIBRACION#
#### Determinación de probabilidades
###  en función a coord. x
franja.x1<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]<=500&dato[i,3]>=0)
    {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
franja.x2<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>500&dato[i,3]<=1000)
    {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
```

```
franja.x3<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>1000&dato[i,3]<=1500)
    {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function

franja.x4<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>1500&dato[i,3]<=2000)
   {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function

franja.x5<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>2000&dato[i,3]<=2500)
   {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
```

```
franja.x6<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>2500&dato[i,3]<=3000)
   {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
franja.x7<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>3000&dato[i,3]<=3500)
   {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
franja.x8<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>3500&dato[i,3]<=4000)
   {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
franja.x9<-function(dato){
dat.franja<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,2]>=0&dato[i,3]>4000&dato[i,3]<=4500)
   {dat.franja<-rbind(dat.franja,dato[i,])}
  } # fin for
return(dat.franja)
} # fin function
```

```
dato.franja.x1<-franja.x1(prob)
dato.franja.x2<-franja.x2(prob)
dato.franja.x3<-franja.x3(prob)
dato.franja.x4<-franja.x4(prob)
dato.franja.x5<-franja.x5(prob)
dato.franja.x6<-franja.x6(prob)
dato.franja.x7<-franja.x7(prob)
dato.franja.x8<-franja.x8(prob)
dato.franja.x9<-franja.x9(prob)

### Evaluar cómo son las probabilidades
### en cada franja y
### determinar esquemas de
### muestreo adecuados
par(mfrow=c(2,2))
hist(dato.franja.x1$p,main ="(x>=0, x<=500)",xlab="p")
hist(dato.franja.x2$p,main ="(x>500, x<=1000)",xlab="p")
hist(dato.franja.x3$p,main ="(x>1000, x<=1500)",xlab="p")
hist(dato.franja.x4$p,main ="(x>1500, x<=2000)",xlab="p")
dev.print(pdf, file="Hist.FranjaXProbabVivero1.pdf")
par(mfrow=c(3,2))
hist(dato.franja.x5$p,main ="(x>2000, x<=2500)",xlab="p")
hist(dato.franja.x6$p,main ="(x>2500, x<=3000)",xlab="p")
hist(dato.franja.x7$p,main ="(x>3000, x<=3500)",xlab="p")
hist(dato.franja.x8$p,main ="(x>3500, x<=4000)",xlab="p")
hist(dato.franja.x9$p,main ="(x>4000, x<=4500)",xlab="p")
dev.print(pdf, file="Hist.FranjaXProbabVivero2.pdf")

quantile(dato.franja.x1$p)
quantile(dato.franja.x2$p)
quantile(dato.franja.x3$p)
quantile(dato.franja.x4$p)
```

```
quantile(dato.franja.x5$p)
quantile(dato.franja.x6$p)
quantile(dato.franja.x7$p)
quantile(dato.franja.x8$p)
quantile(dato.franja.x9$p)

mean.p<-c(mean(dato.franja.x1$p),mean(dato.franja.x2$p),
mean(dato.franja.x3$p),mean(dato.franja.x4$p),
          mean(dato.franja.x5$p),mean(dato.franja.x6$p),
          mean(dato.franja.x7$p),mean(dato.franja.x8$p),
          mean(dato.franja.x9$p))
sd.p<-c(sd(dato.franja.x1$p),sd(dato.franja.x2$p),
sd(dato.franja.x3$p),sd(dato.franja.x4$p),
         sd(dato.franja.x5$p),sd(dato.franja.x6$p),
         sd(dato.franja.x7$p),sd(dato.franja.x8$p),
         sd(dato.franja.x9$p))

###### TERMNINA CALIBRACION

### Metodo aleatorio estratificado PROPUESTO

franja.1<-function(dato){
dat.bloq<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,3]>=0&dato[i,3]<=500&dato[i,2]>=0)
    {dat.bloq<-rbind(dat.bloq,dato[i,])}
  } # fin for
return(dat.bloq)
} # fin function
```

```
franja.9<-function(dato){
dat.bloq<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,3]>4000&dato[i,3]<=4500&dato[i,2]>=0)
    {dat.bloq<-rbind(dat.bloq,dato[i,])}
  } # fin for
return(dat.bloq)
} # fin function
franja.i<-function(dato){ ## franja interior
dat.bloq<-NULL
  for (i in 1:dim(dato)[1]){
   if (dato[i,3]>500&dato[i,3]<=4000&dato[i,2]>=0)
    {dat.bloq<-rbind(dat.bloq,dato[i,])}
  } # fin for
return(dat.bloq)
} # fin function
### NOTA: la data que entra a todas estas
### funciones de bloque.letra
#### debe ser la data sin coordenadas
### (x,y) transformadas a (0,1)
dat.franja.1<-franja.1(data[,1:3])
dat.franja.9<-franja.9(data[,1:3])
dat.franja.i<-franja.i(data[,1:3])

### DEBO crear una SOLA muestra aleatoria a
### partir de CADA FRANJA
### considerada en el muestreo
### aleatorio estratificado
### Muestra sistemática x Estrato
mues.aleat.bloq<-function(dat.bloq,salto){
salto<-round(salto)
dat.sist<-dat.bloq[((1:dim(dat.bloq)[1])%%(salto+1))==salto,]
return(dat.sist)}
```

```
### Obtención de la muestra SISTEMATICA dentro
### en CADA FRANJA
muest.franja.1<-mues.aleat.bloq(dat.franja.1,1)
muest.franja.9<-mues.aleat.bloq(dat.franja.9,1)
muest.franja.i<-mues.aleat.bloq(dat.franja.i,1)
### Se crea la DATA Completa para el MUESTREO
### SISTEMATICO ESTRATIFICADO
### propuesto que se usará en la
### predicción SPDE
data.franjas<-NULL
data.franjas<-rbind(data.franjas,muest.franja.1)
data.franjas<-rbind(data.franjas,muest.franja.9)
data.franjas<-rbind(data.franjas,muest.franja.i)
### Se asigna la muestra aleatoria estratificada
###creada a la variable data
datos<-data.frame(data.franjas)
row.names(datos)<-NULL
### Obtiene una MUESTRA ALEATORIA x Estrato
muest.alet.est<-function(m,dat){
 n<-ceiling(dim(dat)[1]*m)
 planta<-rep(0,n)
 s<-sample(dim(dat)[1],n)
 planta<-dat[s,]
return(planta)
} # fin function
### Obtención de la muestra aleatoria
### dentro en CADA FRANJA
m<-0.20
m<-0.25
m<-0.35
m.a.est.1<-muest.alet.est(m,dat.franja.1)
m.a.est.9<-muest.alet.est(m,dat.franja.9)
```

```
m.a.est.i<-muest.alet.est(0.10,dat.franja.i)
#m.a.est.1<-m.a.est.1[1:413,]
#m.a.est.9<-m.a.est.9[1:400,]

### Se crea la DATA Completa para el MUESTREO
### ALEATORIO ESTRATIFICADO
### propuesto que se usará en la
### predicción SPDE
data.franjas<-NULL
data.franjas<-rbind(data.franjas,m.a.est.1)
data.franjas<-rbind(data.franjas,m.a.est.9)
data.franjas<-rbind(data.franjas,m.a.est.i)

### Se asigna la muestra aleatoria
### estratificada
### creada a la variable data
datos<-data.frame(data.franjas)
row.names(datos)<-NULL

### Hacer predicción a partir de la muestra
### ALEATORIA ESTRATIFICA
formula <- datos$Y ~ 1 +  f(spatial, model=spde)

mesh <- inla.mesh.create(cbind(datos$y,datos$x,0),
        boundary=boundary,
        refine=list(max.edge=150))

datos$spatial <- mesh$idx$loc

## Create the SPDE/GMRF model,
## (kappa^2-Delta)(tau x) = W:
spde = inla.spde.create(mesh, model="matern",
param=list(alpha=2))
```

```
### Se crean todas lsa combinaciones
### lineales posibles
all.lc <- c()
for(i in 1:dim(datos)[1])
{
  lc <- inla.make.lincomb("(Intercept)"=1,
  spatial=c(rep(NA, i-1), 1))
  names(lc) <- paste('lc',
  formatC(i, flag='0', width=3), sep='')
  all.lc <- c(all.lc, lc)
}

r.m.est <- inla(formula, family="binomial", Ntrials=1,
       data = datos, lincomb=all.lc,control.compute=
       list(return.marginals=TRUE,
       dic=TRUE, cpo=TRUE),control.predictor=
       list(compute=TRUE),
       control.inla=list(lincomb.derived.only=TRUE))

### Calculo del rango y tau para muestra aleatoria
### estratificada
kappa.m.a<-inla.tmarginal(function(x) exp(x)^0.5,
r.m.est$marginals.hyperpar$"K.0 for spatial-basisK")
kappa.m.1<-inla.emarginal(function(x) x, kappa.m.a)

rho.m.a<- sqrt(8*1)/kappa.m.1

tau.marg.a<-inla.tmarginal(function(x) exp(x),
r.m.est$marginals.hyperpar$"T.0 for spatial-basisT")
tau.m.1<-inla.emarginal(function(x) x, tau.marg.a)
```

```
##### Probabilidades para Datos Originales
pdata = r$summary.fixed['(Intercept)',"mean"]+
        r$summary.random$spatial[,"mean"]

e<-exp(pdata)
p<-e/(1+e)
prob<-cbind(p[mesh.v$idx$loc])
prob<-data.frame(prob)
names(prob)<-c("p")

## calculo probabilidades a partir del
## mesh de cada muestreo
proj<-inla.mesh.projector(mesh,dims=c(110,110))

pdata.a<-inla.mesh.project(proj,
r.m.est$summary.fixed['(Intercept)',
"mean"]+r.m.est$summary.random$spatial[,"mean"])

e.a <-exp(pdata.a)
p.a <-e.a/(1+e.a)
prob.a <-cbind(p.a[mesh.v$idx$loc])
prob.a <-data.frame(prob.a)
names(prob.a) <-c("p")

prob$p[is.na(prob.a$p)]<-0
p<-prob$p[prob$p>0]

prob.a$p[is.na(prob.a$p)]<-0
# Elimine los puntos que son NA en
# la Muestra FORMADA
p.a<-prob.a$p[prob.a$p>0]
```

```
### AQUI termina el calculo de probabilidades
dif<-p-p.a

# calculo de los errores de predicción
e.est.1<-error(p.a,dif) #discrepancias muestra
#estratificada-aleatoria

### Aqui termina Muestra ALEATORIA
###ESTRATIFICADA PROPUESTA
map.plot(pdata.a,proj)
```