uc3m | Universidad Carlos III de Madrid

Ciencias Actuariales y Financieras
2017-2019

Trabajo Fin de Máster

"Modelo espacial Bayesiano para la estimación de la invalidez en España con la metodología INLA"

Belén Mirás Mazás

Tutores

José Miguel Rodríguez Pardo

Jesús R. S. del Potro

Madrid, 4 de Julio 2019

RESUMEN

La insuficiencia de datos en áreas pequeñas y la existencia de dependencias espaciales no medidas a través de las modelizaciones convencionales de las aseguradoras (como los modelos lineales generalizados) nos llevan a recurrir a otro tipo de metodologías como la inferencia Bayesiana. Las estrategias de modelado espacial bayesianas se pueden aplicar a través de la Aproximación de Laplace Anidada Integrada (INLA en sus siglas en inglés) que permite capturar los efectos aleatorios espaciales estructurados (de influencias entre vecinos), no estructurados (heterogéneos de área) y efectos fijos de variables cómo la edad. Demostramos la efectividad de este tipo de metodología a través de un modelado espacial de la invalidez en España con el que se consigue llegar al mapeo de los riesgos relativos de cada área y localizar determinados *clusters* de riesgo. Otra posible utilidad de este tipo de modelización con INLA es la obtención de variables que capturan la dependencia espacial, que incluyéndolas en un modelo GLM pueden llegar a eliminar este efecto e incluso mejorar su ajuste.

Palabras Clave: *Inferencia Bayesiana, Aproximación de Laplace integrada Anidada,Tasa de incapacidad, Estructura espacial, Datos nivel área.*

ÍNDICE DE CONTENIDO

ÍNDICE DE GRÁFICOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. INTRODUCCIÓN

Los métodos de modelización espacial tradicionales basados en las covarianzas resultan, en muchos casos, ineficientes computacionalmente e inapropiados. Por un lado, el investigador es incapaz de cuantificar la incertidumbre correspondiente a los parámetros. Además, en situaciones complejas ante modelos con muchos parámetros y pocos datos, las estimaciones, mediante los enfoques tradicionales basados en la máxima verosimilitud, resultan problemáticas o imposibles.

Habitualmente, en la modelización estadística general, se asume independencia entre los datos observados. Cuando usamos este tipo de métodos estadísticos, ignorando las correlaciones espaciales entre los datos, podríamos estar subestimando el error estándar de los parámetros de covarianza y sobreestimando la significación estadística. Por ello, es importante otorgar estructuras espaciales a los datos observados objeto de modelización.

En los últimos años, la inferencia bayesiana ha estado en el centro del desarrollo de las estadísticas espaciales. Los modelos jerárquicos bayesianos, que incluyen tanto efectos fijos como efectos aleatorios, se han vuelto especialmente populares en muchos campos diferentes cómo las áreas financieras, ingenierías, salud, epidemiología, etc. La inferencia sobre este tipo de modelos rara vez está disponible en forma cerrada lo que implica acudir a métodos de simulación, como las cadenas de Markov-Monte Carlo, para su ajuste. No obstante, el problema de estos métodos es que suelen ser muy costosos computacionalmente. Frente a los modelos de simulación aparecen nuevas metodologías cómo la Aproximación de Laplace Anidada (INLA en sus siglas en inglés) ofreciendo un enfoque general para calcular las distribuciones marginales posteriores de todos los parámetros involucrados en el modelo.

Se ha demostrado que el enfoque INLA para la inferencia bayesiana aproximada y, concretamente, para la tipología de modelos gaussianos latentes (por ejemplo, en modelos mixtos lineales, modelos espaciales, modelos de supervivencia, modelos de mapeo de enfermedades, etc.) proporciona estimaciones rápidas y precisas de las marginales posteriores. Debido a sus ventajas en cuanto a rapidez y precisión, en los últimos años ha aumentado considerablemente el número de usuarios que han

encontrado en INLA la posibilidad de ajustar modelos que de otra manera no podrían ajustar.

El enfoque Bayesiano es especialmente interesante en el caso de los modelos jerárquicos espaciales, ya que permite que tanto las observaciones como los parámetros del modelo sean variables aleatorias, dando como resultado estimaciones más realistas y precisas de la incertidumbre. Además, permite incorporar información a priori, útil en la discriminación de efectos espaciales de autocorrelación presentes en los efectos lineales no espaciales. En los Modelos Lineales Generalizados (GLM) se ignoran las correlaciones espaciales, lo cual podría tener graves efectos en las inferencias ya que, por ejemplo, si existe correlación espacial positiva se pueden obtener errores estándar en los coeficientes muy pequeños, originando que los efectos se lleguen a juzgar cómo significativos cuando en realidad no lo son. Frente a esto, los modelos jerárquicos Bayesianos espaciales constituyen una metodología capaz de captar este tipo de fuentes de variabilidad no observadas.

Otra de las ventajas frente a los GLM es la robustez de los resultados ante la escasez de datos y la ausencia de variables explicativas ya que el investigador podrá incorporar el conocimiento científico previo sobre los sucesos a través de las distribuciones a priori. El proceso inferencial combina el estado de la información (es decir, antes de observar cualquier dato nuevo) y los datos disponibles, para derivar la distribución posterior.

En este estudio, examinaremos la modelización espacial bayesiana con el objetivo de mapear los riesgos relativos de invalidez en España en escalas geográficas pequeñas y así, identificar aquellas áreas o clústeres dónde el riesgo es mayor o menor que el promedio nacional. La posibilidad de incorporar variables aleatorias espaciales, tanto de efectos estructurados (derivados de las relaciones con las áreas colindantes) cómo efectos heterogéneos específicos de cada área, nos permite clasificar los riesgos según d la localización geográfica de cada área y, en definitiva, modelizar la variabilidad espacial.

Además de la geolocalización del riesgo de invalidez, realizaremos un ejercicio de mejora de ajuste de un Modelo de Regresión Lineal Generalizado introduciendo la variable de riesgo por área que se obtiene del modelo espacial Bayesiano.

El trabajo se divide en las siguientes partes: el Capítulo 2 describe los Modelos Gaussianos Latentes, tipología dentro de la que se engloban los modelos que utilizaremos en el estudio. El Capítulo 3 explica los principales elementos de la inferencia Bayesiana. El Capítulo 4 describe la metodología que se usará para la estimación de las distribuciones a posteriori de los modelos espaciales bayesianos propuestos. El Capítulo 5 contiene los principales elementos del paquete R-INLA, herramienta computacional que se utiliza para el estudio.

El Capítulo 6 se corresponde con el estudio del riesgo de invalidez a través de varios modelos. Por último, en los Capítulos 7 y 8 se exponen las principales conclusiones del análisis y las fuentes de discusión.

2. MODELOS GAUSSIANOS LATENTES

2.1.Introducción

Los modelos Gaussianos Latentes (MGL) engloban un amplio conjunto de modelos estadísticos dónde un campo gaussiano latente, observado indirectamente a través de los datos, se utiliza para modelar, por ejemplo, la dependencia del tiempo y del espacio y el efecto suavizado de covariables.

En los MGL se asume que la variable y_i pertenece a una familia de distribuciones concreta, dónde alguno de los parámetros Φ de la familia está *linkado* a un predictor linear η_i con componentes aditivos, a través de una función de enlace $g(\cdot)$: $g(\Phi_i) = \eta_{i.}$.

Por lo tanto, predictor contará los efectos de varias covariables de una forma aditiva:

$$\eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ij}) + \sum_{k=1}^{\eta_\beta} \beta_k z_{ki} + \varepsilon_i \qquad (1)$$

donde $\{f^{(j)}(\cdot)\}$ son funciones desconocidas de las covariables *u* usadas, por ejemplo, para modelizar dependencia temporal y/o espacial; el $\{\beta_k\}$ representa el efecto lineal de las covariables z y el $\{\varepsilon_i\}$ los términos desestructurados.

Este tipo de modelos también se pueden escribir utilizando una estructura jerárquica, dónde el primer escenario está formado por la función de verosimilitud con propiedades de independencia condicional (propiedad de Markov) dado el campo latente x= (η, α, f, β) y los posibles hiperparámetros Θ_1, dónde cada punto se conecta a un elemento del campo latente x_i.

$$y/x, \Theta_1 \sim \pi(y/x, \Theta_1) = \prod_{i=1}^{n_d} \pi(y_i/x_i, \Theta_1) \qquad (2)$$

El segundo escenario lo forma la distribución condicionada del campo latente x dados unos posibles hiperparámetros Θ_2,

$$x, \Theta_2 \sim \pi\,(x, \Theta_2) = \mathrm{N}(x; \mu(\Theta_2), Q^{-1}(\Theta_2)) \quad (3)$$

dónde $\mathrm{N}(\cdot; \mu, Q^{-1})$ denota una distribución multivariante gaussiana con un vector de medias μ y una matriz de precisión $Q(\Theta_2)$. En muchas aplicaciones, el campo gaussiano latente tiene propiedades de independencia condicional, que se traduce en una matriz de precisión dispersa $Q(\Theta_2)$, que permitirá algoritmos computacionales mucho más eficientes. Las distribuciones multivariantes Gaussianas con matriz de precisión dispersa[1] se conocen como *Gaussian Markov Random Field* (GMRF, Rue and Held, 2005)

El modelo jerárquico se completa con una distribución a priori adecuada para todos los hiperparámetros $\Theta = (\Theta_1, \Theta_2)$

$$\Theta \sim \pi(\Theta) \quad (4)$$

A veces, no es fácil identificar si un modelo en particular se puede escribir en la forma jerárquica explicada. En algunos casos, incluso se podría reescribir el modelo para que pueda encajar en esta clase.

Algunos ejemplos de modelos que se engloban dentro del conjunto de los MGL son los modelos mixtos, los modelos espaciales y los modelos espacio-temporales.

2.2.Modelos Mixtos

Los modelos Mixtos son una extensión de los modelos lineales (generalizados) que permiten modelizar la variabilidad y la presencia de observaciones correlacionadas. Este tipo de modelos contemplan, por un lado, efectos fijos que afectan a todas las observaciones de la misma forma y efectos aleatorios. Por lo tanto, una de las características principales de estos modelos es la existencia de variabilidad heterogénea.

El uso de efectos, tanto fijos cómo aleatorios, en un mismo modelo se puede pensar jerárquicamente lo que conlleva una relación muy cercana entre los modelos mixtos y

[1] Las matrices de dispersión son matrices en las que la mayoría de elementos son ceros.

los llamados modelos lineales jerárquicos. La jerarquía surge porque podemos pensar en un nivel para sujetos, y otro nivel para agrupaciones de esos sujetos.

Este tipo de modelos son de gran utilidad en estudios físicos, biológicos, sociales, etc. donde se realizan mediciones repetidas en las mismas unidades estadísticas, en mediciones en grupos de unidades estadísticas relacionadas.

2.3.Modelos espaciales

Legendre y Legendre (1998) definen la dependencia espacial como "la propiedad de las variables aleatorias que tomar valores, en pares de ubicaciones separadas por cierta distancia, que son más similares (autocorrelación positiva) o menos similares (autocorrelación negtiva) de lo esperado para la asociación aleatoria de pares de observaciones".

La dimensión espacial juega un papel clave en muchos fenómenos sociales. Por un lado, las cosas se distribuyen de manera desigual a través del espacio, creando diferenciación espacial y segregación. Además, hay un bucle de retroalimentación entre la organización de una sociedad y la configuración de los espacios. En la mayoría de casos, el modelado espacial requiere la combinación de conocimiento y habilidades de diversos campos como la estadística, matemáticas, informática o la física ya que proporcionan perspectivas metodológicas estimulantes al investigador interesado en la organización del espacio y la evolución de sus estructuras. Así, los desarrollos en todos estos campos permiten, cada vez más, modelar las relaciones espaciales de forma explícita, en lugar de tener que asumirla o eliminarla.

Los modelos espaciales describen la dependencia espacial existente en una variable respuesta o en los parámetros debido a las relaciones de vecindad.

El primer paso consiste en identificar los objetos elementales que determinan el nivel al que se recogerá la información. Estos objetos podrán ser celdas o píxeles, hogares, entidades espaciales o aglomeraciones, entre otros.

A continuación se puede formalizar un modelo que sirva para probar hipótesis o establecer escenarios que permitan realizar estudios prospectivos y ayudar a la toma de decisiones.

Cuándo los lugares han sido caracterizados, se ha determinado la intensidad y los tipos de diferenciaciones espaciales, y las similitudes y diferencias han salido a la luz, el siguiente paso es encontrar la relación entre las características de la organización espacial y los intercambios que estos lugares mantienen entre ellos, así como las influencias que tienen unos sobre otros que se pueden originar desde las disparidades del espacio.

La interacción entre las áreas puede ser el objeto de la investigación, tratando de explicar por qué algunos flujos son más o menos significativos entre ciertos pares de ubicaciones.

2.4.Modelos espacio-temporales

En la literatura estadística existe una amplia gama de modelos para describir la distribución geográfica de determinados eventos, así como su evolución en el tiempo. Con el paso de los años, las técnicas estadísticas para la modelización de tendencias y dependencias espacio-temporales de sucesos tales como enfermedades, mortalidad, supervivencia, etc. se han hecho cada vez más sofisticadas y nos permiten entender las dinámicas de estos procesos en espacio y tiempo.

En principio, cualquier proceso espacio-tiempo Gaussiano η_{it} podría ser usado. No obstante, el requerimiento de una estructura Gauss-Markoviana que permitan cálculos matriciales rápidos y el alcance actual de los modelos implementados en R-INLA imponen algunas restricciones.

Una formulación genérica de un modelo espacio-temporal Gaussiano latente para el predictor η_{it} que engloba muchos de los modelos que pueden ser ajustados con R-INLA sería:

$$\eta_{st} = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{sit}) + \sum_{k=1}^{\eta_\beta} \beta_k z_{kst} + z_t + z_s + z_{st} + \varepsilon_s \qquad (5)$$

Dónde denotamos como z_{kit}, k=1,…,K las covariables dadas que pueden depender del espacio y tiempo o sólo de alguno de ellos. Los coeficientes α y β_k son los que conocemos cómo efectos fijos, mientras que la función f(·) y los procesos z_t, z_i y z_{it} son los que conocemos como efectos aleatorios.

En un modelo espacio-temporal Bayesiano, los efectos aleatorios estructurados y no estructurados se utilizan para modelar la autocorrelación espacial inherente en los datos, los efectos del tiempo correlacionados y no correlacionados modelan la estructura dependiente del tiempo de los datos, las covariables que varían en el tiempo modelan el extra de incertidumbre en los datos debido a factores de confusión medidos, y los efectos de interacción espacio-tiempo modelan la variación residual espacio-temporal que no se tiene en cuenta por los efectos aleatorios del espacio y el tiempo para producir estimaciones en tiempo basadas en modelos confiables a nivel geográfico. Las distribuciones a posteriori de los parámetros en los modelos espacio-temporales jerárquicos bayesianos en este estudio se simularán con el software R-INLA.

Este método se puede aplicar a gran número de causas raras de los resultados de mortalidad para examinar la variación geográfica, la temporal o la conjunta con estimaciones de áreas pequeñas.

3. INFERENCIA BAYESIANA

La inferencia bayesiana es un potente método de inferencia estadística que permite modelar cualquier variable aleatoria proporcionando resultados precisos en situaciones de especial incertidumbre e imprecisión: cuándo no se conoce con certeza la verdad o falsedad de una hipótesis, cuando las bases de datos disponibles son limitadas o con modelos con elevada variabilidad.

El trabajo se centra en la metodología de aproximación de Laplace Integrada (INLA) a través de inferencia bayesiana.

Es importante introducir los principales componentes y notación sobre la inferencia bayesiana, que nos servirá de base para entender la metodología seguida en el resto del estudio.

Cada vez es más frecuente el uso de métodos Bayesianos para el mapeo de enfermedades y circunstancias relacionadas con la salud o mortalidalidad y supervivencia, especialmente en áreas con una cantidad de datos pequeña.

La inferencia bayesiana combina la información a priori de los parámetros, conocida por el investigador a través de estudios previos, o bien reflexiones y juicios racionalmente conformados, para derivar, a través de la regla de Bayes, la distribución que resume el comportamiento de los parámetros condicionada a los datos observados: la distribución a posteriori.

En el caso que nos ocupa, necesitaremos información acerca del monto de personas en determinadas áreas geográficas y el número de invalideces observadas. Estos dos conjuntos de datos permitirían el cálculo de estimaciones convencionales de tasas máximo verosímiles. Sin embargo, tratándose de áreas pequeñas podríamos llegar a estimaciones que resultasen extremas por lo que resulta lógico el uso de otro tipo de información como la proximidad geográfica, similitudes económicas o demográficas entre áreas, etc. Así, se configura lo que se conoce como información a priori que podrá ser representada por distribuciones de probabilidad.

El producto entre la información a priori y la verosimilitud de los datos nos permitirá obtener la distribución a posteriori para las tasas desconocidas.

Por lo tanto, la distribución a posteriori será la distribución de los parámetros **Θ** condicionada a un conjunto de datos observados **y** que se obtendrá a través del teorema de Bayes:

$$\boldsymbol{\pi}(\Theta/\boldsymbol{y}) = \frac{\boldsymbol{\pi}\left(\boldsymbol{y}/\Theta\right)\boldsymbol{\pi}(\Theta)}{\boldsymbol{\pi}(\boldsymbol{y})} \tag{6}$$

Siendo $\boldsymbol{\pi}(\Theta)$ la distribución a priori de los parámetros, $\boldsymbol{\pi}\left(\boldsymbol{y}/\Theta\right)$ la distribución de probabilidad de **x** y $\boldsymbol{\pi}(\boldsymbol{y})$ la distribución marginal de **y**.

En muchos casos, la obtención de la distribución marginal $\boldsymbol{\pi}(\boldsymbol{y})$ resulta de elevada dificultad siguiendo la siguiente forma:

$$\boldsymbol{\pi}(\boldsymbol{y}) = \int \boldsymbol{\pi}\left(\boldsymbol{y}/\Theta\right)\boldsymbol{\pi}(\Theta)\ \boldsymbol{d}\Theta \tag{7}$$

En la práctica, la distribución a posteriori se estima sin calcular la distribución marginal como

$$\boldsymbol{\pi}(\Theta/\boldsymbol{y}) \propto \boldsymbol{\pi}\left(\boldsymbol{y}/\Theta\right)\boldsymbol{\pi}(\Theta) \tag{8}$$

Es posible que $\boldsymbol{\pi}(\Theta/\boldsymbol{y})$ no tenga una forma cerrada por lo que se recurre a métodos tales como la simulación Morkov-Chain Monte Carlo (MCMC) para aproximar dicha distribución. No obstante, el principal problema de los métodos de simulación, en el marco de la inferencia bayesiana, reside en la gran carga computacional. La complejidad de los modelos con estructuras espacio-temporales y de bases de datos cada vez más grandes y provenientes de diversas fuentes podría conllevar horas o incluso días para el cálculo de distribuciones a posteriori precisas.

La metodología INLA (*Integrated Nested Laplace Approximation*) sustituye las simulaciones MCMC con aproximaciones de las distribuciones marginales posteriores precisas y deterministas. La calidad de estas aproximaciones es extremadamente alta, tanto que incluso ejecuciones muy largas de MCMC no han podido detectar ningún error en ellas. INLA presenta dos ventajas principales frente a las técnicas MCMC. La primera y más importante es computacional. Usando INLA los resultados pueden obtenerse en segundos o minutos incluso tratándose de modelos con un enorme campo latente, dónde un algoritmo MCMC podría tardar días. La segunda es que INLA trata

los modelos Gaussianos Latentes de forma unificada, lo que permite una mayor automatización del proceso de inferencia. El núcleo computacional se adapta automáticamente a cualquier tipo de campo latente, de modo que no importa si trabajamos, por ejemplo, con modelos espaciales o modelos espacio-temporales.

En el siguiente capítulo profundizaremos en todos los aspectos de esta metodología.

4. APROXIMACIÓN DE LAPLACE INTEGRADA ANIDADA (INLA)

4.1.El principio de aproximación de Laplace

Dado que la metodología INLA parte del principio de aproximación de Laplace, antes de nada, es interesante recordar en qué consiste y cómo es su cálculo en la práctica.

La aproximación de Laplace resulta útil para obtener el valor de una integral mediante la expansión de Taylor de segundo orden para una función f(x). Es decir, se busca

$$I = \int_a^b f(x)dx \tag{8}$$

dónde a y b pueden ser límites infinitos.

Primero se busca la moda x_0 para luego hacer la expansión de Taylor de segundo orden de f(x) centrada en este punto de la forma:

$$\log\big(f(x)\big) \approx \log\big(f(x_0)\big) + \frac{1}{2}A(x - x_0)^2 \tag{9}$$

Dónde

$$A = -\frac{d^2}{dw^2} logf(w); \quad w = w_0 \tag{10}$$

y tomando la exponencial:

$$f(x) \approx f(x_0)\exp\left\{\frac{1}{2}A(x - x_0)^2\right\} \tag{11}$$

4.2.La metodología INLA

H. Rue, S. Martino and N. Chopin (2009) proponen un nuevo enfoque para la aproximación de las distribuciones marginales posteriores de modelos Gaussianos Latentes basándose en la aproximación de Laplace.

El método realiza cálculos numéricos directos de un amplio conjunto de este tipo de modelos con la ventaja, frente a los métodos de simulación Markov Chain Monte Carlo,

de una menor carga computacional. Los modelos que abarca serán de la siguiente forma:

$$\Theta \sim \pi(\Theta) \tag{12}$$

$$(x/\Theta) \sim N(0, Q(\Theta)^{-1}) \tag{13}$$

$$\eta_i = \sum_j c_{ij} x_i \tag{14}$$

$$(y_i/x\,, \Theta) \sim \pi(y_i/\eta_i\,, \Theta) \tag{15}$$

Dónde Θ son los hiperparámetros, Q(Θ) es la matriz de precisión, x será el conjunto de variables gaussianas (vector de efectos latentes), η el predictor lineal, c_{ij} valores de covarianza conocidos e y el vector de datos observados.

H. Rue, S. Martino and N. Chopin escriben la distribución a posteriori de los efectos latentes como

$$\pi(x, \Theta/y) = \frac{\pi(y/x\,, \Theta)\pi(x, \Theta)}{\pi(y)} \propto \pi(y/x\,, \Theta)\pi(x, \Theta) \tag{16}$$

Dónde π (y) representa la verosimilitud marginal del modelo, una constante normalizada que normalmente se ignora debido a la dificultad para su cálculo.

La función π(y/x,Θ) también representa la verosimilitud y se puede escribir de la siguiente forma:

$$\pi(y/x\,, \Theta) = \prod_{i \epsilon I} \pi(y_i/x_i\,, \Theta) \tag{17}$$

Dado que se asume que **x** se distribuye como un GMRF[2] (Gaussian Markov Ramdon Field), la distribución a poteriori de los efectos latentes quedaría:

$$\pi(x/\Theta) \propto |Q(\Theta)|^{\frac{1}{2}} \exp\{-\frac{1}{2}x^T Q(\Theta)x\} \tag{18}$$

y tomando estas consideraciones, la distribución conjunta de los efectos latentes y los hiperparámetros se puede escribir como,

$$\pi(x,\Theta/y) \propto |Q(\Theta)|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}x^T Q(\Theta)x\right\} \prod_{i\epsilon I} \pi(y_i/x_i,\Theta) = \tag{19}$$

$$|Q(\Theta)|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}x^T Q(\Theta)x + \sum_{i\in I} \log(\pi(y_i/x_i,\Theta))\right\}$$

Sin embargo, INLA no busca estimar la distribución a posteriori conjunta del modelo, sino que se centra en el cálculo de las distribuciones marginales a posteriori de los efectos latentes e hiperparámetros.

Las distribuciones marginales posteriores de las variables latentes $\boldsymbol{\pi(x_i/y)}$ se calculan mediante integración numérica

$$\boldsymbol{\pi(x_i/y)} = \int \boldsymbol{\pi(x_i,\Theta/y)}\ \boldsymbol{\pi(\Theta/y)}\ \boldsymbol{d\Theta} \tag{20}$$

De una forma similar, se calcula la distribución marginal posterior de cada uno de los hiperparámetros $\Theta_{\boldsymbol{j}}$, J= 1,…,m

$$\boldsymbol{\pi}(\Theta_j/\boldsymbol{y}) = \int \boldsymbol{\pi}(\Theta/\boldsymbol{y})\, \boldsymbol{d}\Theta_{-\boldsymbol{k}} \tag{21}$$

Siendo Θ_{-k} el vector de hiperparametros de Θ sin Θ_k.

[2] *Campo Aleatorio de Markov Gaussiano (GMRF)*: η es un GMRF si tiene una densidad normal multivariada y además cumple la independencia condicional (*Propiedad de Markov*). La ventaja de tener un GMRF se debe a las matrices de precisión son dispersas lo cual redunda en algoritmos computacionales eficientes.

Por lo tanto, necesitamos calcular la distribución condicionada de los hiperparámetros $\boldsymbol{\pi}(\Theta/\boldsymbol{y})$ y $\boldsymbol{\pi}(\boldsymbol{x_i}, \Theta_{\boldsymbol{k}}/\boldsymbol{y})$**.** El enfoque INLA explota los supuestos del modelo para llegar a una aproximación numérica de las distribuciones posteriores de interés, basándose en la **aproximación de Laplace** (Tierney and Kadane, 1986). Por un lado, la distribución marginal posterior de los hiperparámetros Θ se aproximará siguiendo la fórmula:

$$\boldsymbol{\pi}(\Theta/\boldsymbol{y}) \approx \frac{\boldsymbol{\pi}(\boldsymbol{x}/\Theta)\,\boldsymbol{\pi}(\Theta)\boldsymbol{\pi}(\boldsymbol{y}/\boldsymbol{x})}{\boldsymbol{\pi}(\boldsymbol{x}/\Theta,\boldsymbol{y})} = \frac{\boldsymbol{\pi}(\boldsymbol{x},\Theta,\boldsymbol{y})}{\boldsymbol{\pi}_{\boldsymbol{G}}(\boldsymbol{x}/\Theta,\boldsymbol{y})}\Big|_{x=x^*(\Theta)} = \overline{\boldsymbol{\pi}}(\Theta/y) \tag{22}$$

Siendo $\boldsymbol{\pi}_{\boldsymbol{G}}(\boldsymbol{x}/\Theta\,,\boldsymbol{y})$ la aproximación gaussiana de la distribución condicionada de todos los efectos latentes y x*(Θ) su moda.

El siguiente paso es obtener $\boldsymbol{\pi}(\boldsymbol{x_i}, \Theta/\boldsymbol{y})$**.** H. Rue, S. Martino and N. Chopin (2009) proponen tres formas: aproximaciones gaussianas, aproximación de Laplace y la aproximación de Laplace aproximada.

La fórmula más rápida y sencilla para llegar a una aproximación de $\boldsymbol{\pi}(\boldsymbol{x_i}, \Theta/\boldsymbol{y})$ es mediante la aproximación gaussiana, no obstante, a pesar de proporcionar resultados bastante razonables pueden tener errores de localización y/o errores debido a la falta de asimetría.

La aproximación de Laplace es más precisa, pero requiere una carga computacional elevada.

El algoritmo más eficiente sería la aproximación de Laplace simplificada que usa la aproximación de Taylor de tercer grado en numerador y denominador.

Operacionalmente, INLA sigue el siguiente proceso:

Primero, explora la distribución marginal posterior para los hiperparámetros $\overline{\pi}(\Theta/y)$ con el fin de hallar su modo.

A continuación, se realiza un *grid search*[3] que producirá un conjunto de puntos relevantes para los hiperparámetros $\{\Theta_k\}$ y sus correspondientes pesos W_Θ para aproximar su distribución.

Las distribuciones marginales posteriores de cada uno de estos hiperparámetros $\bar{\pi}(\Theta_k/y)$ se pueden obtener por interpolación de los valores calculados y corrigiendo la posible asimetría, por ejemplo, utilizando *log-splines*.

La distribución condicionada posterior $\boldsymbol{\pi}(\boldsymbol{x_i}, \Theta_{\boldsymbol{k}}/\boldsymbol{y})$ para cada hiperparámetro se evalúa en una cuadrícula de valores seleccionados para x_i y las marginales posteriores $\boldsymbol{\pi}(\boldsymbol{x_i}/\boldsymbol{y})$ se obtienen mediante integración numérica:

$$\boldsymbol{\pi}(\boldsymbol{x_i}/\boldsymbol{y}) \approx \sum_{\boldsymbol{k=1}}^{\boldsymbol{K}} \boldsymbol{\pi}(\boldsymbol{x_i}, \Theta_{\boldsymbol{k}}/\boldsymbol{y})\, \boldsymbol{\pi}(\Theta_{\boldsymbol{k}}/\boldsymbol{y}) \boldsymbol{W}_\Theta \tag{23}$$

Las distribuciones posteriores marginales obtenidas a través de este proceso se podrán utilizar para calcular estadísticos de interés tales como las medias posteriores, varianzas, cuantiles, etc.

En definitiva, podemos decir que INLA explota completamente todas las características principales de los modelos Gaussianos latentes. El campo Gaussiano latente y el usual "buen comportamiento" de la función de verosimilitud justifican la adecuación de la aproximación de Laplace. Además, la poca cantidad de hiperparámetros facilitan la resolución de la ecuación anterior por integración numérica computacionalmente.

[3] El *grid search*, o búsqueda en cuadrícula, es el proceso de realizar el ajuste de hiperparámetros para determinar los valores óptimos para un modelo determinado.

5. PAQUETE R-INLA

El enfoque INLA descrito se ha implementado en un paquete de R llamado R-INLA. Está disponible para Linux, Mac y Windows y se puede obtener a través de un repositorio específico en "*http://www.r-inla.org*" Además, la misma web proporciona documentación y muchos ejemplos.

La sintaxis del paquete R-INLA se basa en la función GLM de R. Una llamada básica comienza con la construcción de la estructura aditiva del predictor linear descrito de la siguiente forma:

```
formula = y ~ 1 + z + f(x1, model1="...") + f(x2, model2="...")
```

Donde, ***y*** es la variable respuesta, ***z*** son los efectos fijos y las funciones ***f(.)*** representan los componentes Gaussianos con efectos aleatorios del modelo. El modelo, por defecto, será *iid*. No obstante, la interface está preparada para permitir la especificación de diferentes modelos y distribuciones a priori de los parámetros. La lista de todos los modelos disponibles se obtiene introduciendo *names(inla.models()$latent):*

Tabla 1: Distribuciones INLA.

names <-vnames(inla.models()$latent)					
linear	Iid	mec	Meb	rgeneric	rw1
rw2	crw2	seasonal	besag	besag2	bym
bym2	besagproper	besagproper2	Fgn	fgn2	ar1
ar1c	Ar	ou	generic	generic0	generic1
generic2	generic3	spde	spde2	spde3	iid1d
iid2d	iid3d	iid4d	iid5d	2diid	z
rw2d	rw2diid	slm	matern2d	Copy	clinear
sigm	Revsigm	log1exp	logdist		

Fuente: Elaboración propia

Explicaremos los más comunes:

- El modelo ***"iid"*** define un ruido Gaussiano aleatorio independiente a priori para la covariable ***ind.***Esto es:

$$\mathbf{x} \sim N(0, T^{-1}I_{k)},$$

dónde T es el parámetro de precisión e I_k la matriz de identidad de dimensión k x k.

- El **"besag"** define una distribución a priori CAR. Es decir,

  $$\mathbf{x} \sim N(0, [TR]^{-}),$$

  donde T es el parámetro de precisión y **R** es la matriz espacial adyacente de dependencias entre vecinos. Por defecto, este modelo impone la restricción de suma cero $\sum_{i=1}^{k} xi = 0$ (*constr=TRUE*).

Una vez definida la fórmula, se podrá introducir el algoritmo de INLA con la función *inla ()*:

```
inla(formula, family="...", data, ...)
```

Dónde *formula* será la designada más arriba, *data* será el data frame que contiene todas las variables de la fórmula y *family*[4] especifica la verosimilitud del modelo. Además, la función *inla* incluye muchas otras opciones.

Nótese que INLA, por defecto, estima la distribución marginal posterior de los hiperparámetros mediante un algoritmo de integración[5] que se ha mostrado adecuado para lograr estimaciones razonablemente precisas. Sin embargo, si el interés se establece principalmente en los hiperparámetros, un método alternativo sería utilizar el comando *inla.hyper* después de la ejecución de *inla*.

5.1.Las distribuciones a priori en R-INLA

Las distribuciones a priori juegan un papel determinante en los análisis Bayesianos. En caso de no ser asignada, R-INLA usará una por defecto. Por lo tanto, es importante conocer cuáles son las opciones por defecto, así como las demás alternativas que se deberán establecer explícitamente. Las priori de los hiperparámetros de la función de máxima verosimilitud deben definirse en el argumento *hyper* dentro *control.family* en la llamada de *inla*. Las priori de los hiperparámetros de los efectos latentes se establecen en el argumento *hyper* dentro de la función *f(.)*.

[4] Introduciendo *names(inla.models()$likelihood)* R devuelve una lista de todas las distribuciones.
[5] Martins (2012)

Introduciendo la siguiente instrucción obtenemos las diferentes distribuciones a priori que se pueden implementar en *inla*:

Tabla 2: Distribuciones a priori

names(inla.models()$prior)	
Normal	Gaussian
wishart1d	wishart2d
wishart3d	wishart4d
wishart5d	Loggamma
minuslogsqrtruncnormal	Logtnormal
Logtgaussian	Flat
Logflat	Logiflat
Mvnorm	pc.ar
Dirichlet	None
Invalid	Betacorrelation
Logitbeta	pc.prec
pc.dof	pc.cor0
pc.cor1	pc.fgnh
pc.spde.GA	pc.matern
pc.range	pc.gamma
pc.mgamma	pc.gammacount
Pc	ref.ar
Pom	Jeffreystdf
expression:	table:

Fuente: Elaboración propia.

En la siguiente tabla se incluye un resumen de estas distribuciones a priori con información sobre su parametrización, útil para definir la distribución. Por ejemplo, la normal utilizando la media y la precisión (en lugar de la varianza o desviación típica).

Tabla 3: Distribuciones a priori y parámetros

Priori	Descripción	Parámetros
normal	Priori Gaussiana	mean, Precision
Gaussian	Priori Gaussiana	mean, Precision
loggamma	log-Gamma	shape, rate
logtnormal	Priori Gausiana truncada (positiva)	mean, Precision
logtgaussian	Priori Gausiana truncada (positiva)	mean, Precision
Flat	Priori flat en Θ	
Logflat	Priori flat en exp(Θ)	
Logiflat	Priori flat en exp(-Θ)	
Mvnorm	Priori Normal multivariante	
Dirichlet	Priori Dirichlet	A
betacorrelation	Priori Beta para correlación	a,b
Logibeta	Priori beta, escala log	a,b
jeffreystdf	Priori Jeffreys	
Table	Priori definida por el usuario	
expression	Priori definida por el usuario	

Fuente: Elaboración propia

Con las expresiones *table* o *expression* el usuario podrá definir sus propias distribuciones a priori.

Tabla 4: Resumen R-INLA

Los 4 bloques del programa R-INLA
1. Organización de los **datos**.
2. La notación de la ***fórmula*** (similar a las funciones de GLM).
3. La llamada al **modelo *inla***.
4. Extracción de la **información posterior**.

Fuente: Elaboración propia

El output será un objeto de la clase *inla.* Esta es una lista que contendrá, al menos, los argumentos que se numeran en la Tabla 5.

Tabla 5: Output función "inla"

Elementos	Descripción
`summary.fixed`	Matriz que contiene media, desviación y cuantiles de los efectos fijos del modelo.
`marginals.fixed`	Lista con las densidades marginales posteriores de los efectos fijos.
`summary.random`	Liste de matrices que contienen media, desviación y cuantiles de los efectos aleatorios.
`marginals.random`	Lista con las densidades marginales posteriores de los efectos aleatorios.
`summary.hyperpar`	Matriz que contiene media, desviación y cuantiles de los hiperparámetros del modelo.
`marginals.hyperpar`	Lista con las densidades marginales posteriores de los hiperparámetros.
`summary.linear.predictor`	Matriz que contiene media, desviación y cuantiles de los predictores lineales en el modelo.
`marginals.linear.predictor`	Lista con las densidades marginales posteriores de los predictores lineales.
`summary.fitted.values`	Matriz que contiene media, desviación y cuantiles de los valores ajustados g^{-1}(η) obtenidos al transformar los predictores lineales por el inverso de la función enlace.
`marginals.fitted.values`	Lista de las densidades marginales de los valores ajustados g^{-1}(η) obtenidos al transformar los predictores lineales por el inverso de la función enlace.
`dic`	El criterio de información de desviación y el número efectivo de parámetros.
`cpo`	Valores de la ordenada predictiva condicional (cpo) y valores de la transformada integral de probabilidad (pit).
`waic`	Valores del criterio de Watanabe-Akaike y el número efectivo de parámetros estimados.

Fuente: Elaboración propia información R studio.

A partir de los resultados principales proporcionados por el software R-INLA, también podemos obtener otros resultados de interés que nos permiten comparar y escoger entre diferentes modelos tales como el Criterio de Información de Deviance (DIC) o verosimilitudes marginales así como diversas medidas predictivas.

Además, la biblioteca INLA también incluye un conjunto de funciones que procesan las densidades marginales a posteriori obtenidas. Estas funciones permiten calcular cuantiles, percentiles, expectativas de la función del parámetro original, muestreo, etc.

5.2.Selección de modelos

La medida principal para la comparación y selección de los modelos es Criterio de Información de Desviación (DIC en sus siglas en inglés). El DIC es una medida de

complejidad y ajuste que se utiliza para comparar modelos jerárquicos complejos basándose en la desviación:

$$\mathbf{D(\Theta,x,y) = -2log(\pi(y/\Theta))} \tag{24}$$

y se define cómo

$$\mathbf{DI}C = \overline{D} + P_D\mathbf{(\Theta)} \tag{25}$$

dónde $\overline{D}$ es la media posterior de la desviación y $P_D(\Theta)$ es el número efectivo de parámetros. Valores pequeños del DIC indican un mejor *trade-off* entre complejidad y ajuste del modelo.

Debemos ser cautos en el uso del DIC, ya que asume la media posteriori como una buena medida del centro de la distribución a posteriori. Sin embargo, se han detectado problemas en los que la distribución a posteriori es asimétrica o unimodal.

La densidad predictiva (cpo o *conditional predictive ordinate*) y la transformación integral de probabilidad (pit) son otras cantidades que ofrece R INLA para evaluar el poder predictivo del modelo o para detectar observaciones sorprendentes.

Además, como la aproximación simplificada de Laplace podría no ser lo suficientemente precisa para el cálculo de medidas predictivas, la función *inla* genera un vector (*failure*) que contiene valores de 0 y 1 para cada observación: el valor 0 indica que el cálculo del cpo y el pit para cada observación se ha realizado sin problemas mientras que un valor mayor que 0 indica la ocurrencia de algún problema informático y que las medidas deberán ser recalculadas.

6. MODELIZACIÓN ESPACIAL BAYESIANA PARA LA INVALIDEZ EN ESPAÑA CON R-INLA

6.1.Datos

Los datos seleccionados para el estudio se corresponden con la base de datos de invalidez permanente absoluta (IPA) de la cartera de una importante multinacional aseguradora en España. Filtramos los datos por pólizas vigentes entre 2016-2017 de asegurados (hombres y mujeres) con edades comprendidas entre los 35 y 65 años.

Las variables que se utilizarán a lo largo del estudio serán el código postal asignado a cada póliza, la edad y el sexo.

Para crear el archivo de información entre vecinos que permitirá definir el modelo espacial, se necesitan una serie de plantillas espaciales o "shape files" con información sobre código postal, latitudes, longitudes y área. Los "shape files" permiten crear los polígonos a través de la unión de coordenadas. Estos, contienen varios archivos indispensables que se leen de forma conjunta: las entidades geométricas de los objetos (.shp), información sobre cada uno de los objetos (.obj), índice de las entidades geométricas (.shx) y el sistema de coordenadas de referencia (.pri). Los "shape files" pueden obtenerse a través de fuentes oficiales como el Instituto Nacional de Estadística Español.

Con el fin de eliminar el efecto de las fronteras administrativas de una división del mapa de España en municipios, provincias o Comunidades Autónomas, se empleará el método de rasterización del mapa. Un ráster es una estructura de datos que divide una región (en nuestro caso España) en rectángulos llamados celdas o píxeles, que almacenan valores. La estructura también se conoce como "cuadrícula" y, a menudo, se contrasta con los datos vectoriales que se utilizan para representar puntos, líneas y polígonos.

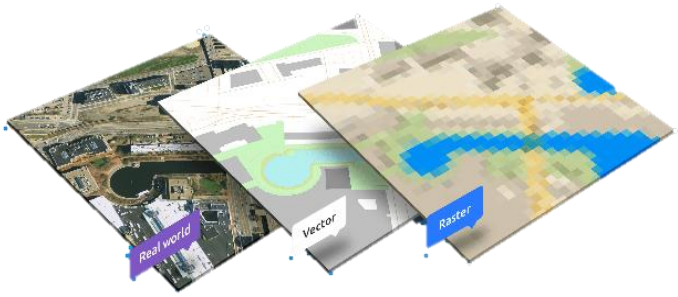

Figura 1. Mundo Real vs Vectores vs Ráster. Fuente :https://www.maptiler.com/blog/2019/02/what-are-vector-tiles-and-why-you-should-care.html

En el caso que nos ocupa, hemos dividido el mapa peninsular en 441 polígonos iguales y estas serán las áreas geográficas de estudio cuyos datos agruparemos para poder realizar la modelización espacial.

6.2 Metodología

El análisis se divide en tres grandes bloques.

Primero, se ajusta un modelo Lineal Generalizado de tipo logístico para explicar la invalidez absoluta a través de la variable edad actuarial. La razón por la que se escoge este tipo de modelización es por tratarse de una de las herramientas estadísticas más utilizadas entre las compañías aseguradas para la valoración de los riesgos y la fijación de las primas de sus pólizas.

Este modelo nos permitirá, por un lado, extraer predicciones globales sobre la invalidez (que no tienen en cuenta la localización geográfica ni la variabilidad espacial), que necesitaremos agrupar para crear una variable a incorporar en los modelos espaciales de INLA. Por otro lado, nos sirven para la comprobación, a través del contraste de I de Moran detallado más adelante, de la existencia de dependencia espacial y así, dar justificación a la necesidad de una modelización espacial de los fenómenos de estudio. El buen ajuste del modelo GLM se validará a través de la curva de ROC.

En el segundo bloque se construyen, a través de la metodología INLA explicada, dos modelos espaciales bayesianos para el riesgo de invalidez en España: el primero de ellos incluirá sólo las variables aleatorias de efectos espaciales estructurados y no

estructurados y en el segundo incorporaremos la variable edad como un efecto fijo. Se analizan los principales resultados: medias, desviación típica, cuantiles, distribuciones a posteriori de los componentes de los modelos (efectos fijos, los efectos aleatorios y los hiperparámetros).

Se compara el ajuste de los modelos a través de sus estadísticos DIC y escogemos el que muestre un mejor *trade-off* entre complejidad y ajuste del modelo (el valor DIC más bajo). La capacidad predictiva de cada uno de los modelos se compara a través del estadístico de logscore.

Se extrae un índice de riesgos relativos de invalidez por cada área que indicará las zonas donde el riesgo de invalidez supera a la media global. Esto permitirá dibujar mapas en los que se pueden apreciar a simple vista determinados *clusters* espaciales dónde los riesgos relativos de cada área pueden estar por encima o por debajo del riesgo global de invalidez para la cartera descrita.

Por último, en el tercer bloque recuperamos el primer GLM y le incorporamos como variable explicativa el índice de riesgo espacial que hemos obtenido a través de los modelos espaciales anteriores. Comprobamos de nuevo la significación de las variables, la mejora en el ajuste y la eliminación de la dependencia espacial en los resultados.

6.3 Resultados

6.3.1 Estimación de la invalidez: Modelo Lineal Generalizado

Una de las variables input necesarias para la estimación de la invalidez con el modelo espacial bayesiano a través de la metodología INLA es la tasa de invalidez esperada global agregada.

Por lo tanto, iniciamos el estudio construyendo un modelo lineal generalizado del tipo logístico que proporcione una estimación de la invalidez dependiente de la edad, para luego agregar tal estimación a nivel área.

Generamos un modelo de la familia Binomial con función *link* "*probit"* que tomará la siguiente forma:

$$\Phi^{-1}(\pi_i) = \eta_i = \beta_0 + \beta_1 fedad\_actuar_{1i} + NumExp \quad (26)$$

Siendo Φ^{-1} la función inversa de la función de distribución Normal estándar, la variable "fedad_actuar" la edad actuarial de cada asegurado y "NumExp" la variable *offset*[6] que se incorpora como medida de exposición al riesgo.

Los parámetros se estiman a través de la función *glm* de R, empleando el método de máxima verosimilitud.

En el siguiente cuadro se muestran los resultados del modelo:

Tabla 6: Resultados modelo glm

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0,1034	-0,0466	-0,0332	-0,025	41,3040

Coefficients:

	Estimate	std.Error	z value	Pr(>|z|)	
(Intercept)	-13,6273	0,4821	-28,2700	<2e-16	***
fedad_actuar	0,1138	0,0090	12,5900	<2e-16	***

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fuente: Elaboración propia con software R studio

El contraste de la hipótesis de nulidad de los parámetros se realiza mediante el test de Wald. En R, dicho contraste se expresa aprozimando el estadístico de Wald a un valor *z* con su correspondiente *p-value*.

[6] Variable que tomará el valor 1 si el asegurado a mantenido su póliza en la compañía durante todo el año, o menor que 1 (en la parte proporcional) de no ser así.

Tabla 7. Contraste de Hipótesis de Wald.

H_0: $\beta_i = 0$
H_α: $\beta_{iv} \neq 0$

Fuente: Elaboración propia.

En el modelo ajustado se observa que individualmente, el coeficiente de la variable edad actuarial es estadísticamente muy significativo con un nivel de significación por debajo del 1%.

Para la validación del modelo construiremos la curva de ROC (*Receiver Operating Characteristic*) La construcción de la curva ROC se realiza a través de los llamados Verdaderos Positivos (VPR) y los Falsos positivos (FPR). Por un lado, los VPR miden hasta qué punto el modelo clasifica correctamente los casos positivos y los FPR cuántos de los resultados positivos son erróneos de entre todos los casos negativos. La curva ROC representará los intercambios entre los verdaderos positivos y los falsos positivos. Por lo tanto, a través de la curva ROC obtenemos cual es la probabilidad de obtener un resultado positivo cuando el individuo verifica el suceso estudiado, para los falsos positivos en distintos tipos de corte.

El área debajo de la curva nos da la capacidad para discriminar entre ocurrencia y no ocurrencia del suceso analizado: si el área = 0,5 significa que existe la misma probabilidad de clasificar un evento en cualquiera de las dos alternativas posibles.

En el **¡Error! No se encuentra el origen de la referencia.** se muestra la curva ROC resultante del modelo. El área debajo de la curva es de 0.75, por lo tanto, la probabilidad de que la prueba se clasifique correctamente es bastante alta.

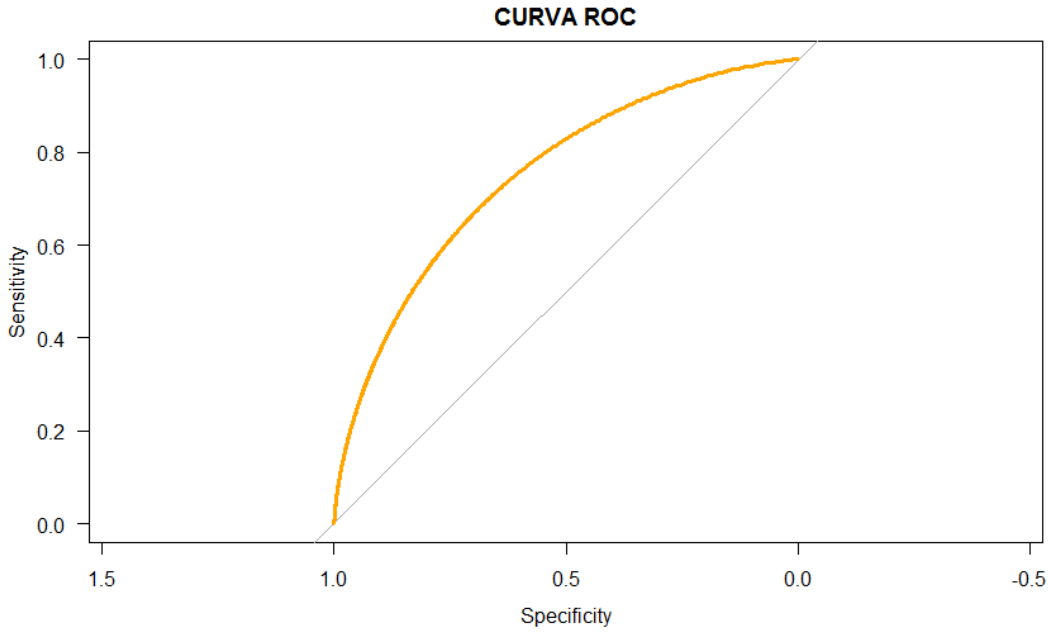

Gráfico 1: Curva ROC. Elaboración propia R-studio

En definitiva, se trata de un buen modelo para estimar y predecir la invalidez en la cartera.

Generaremos la predicción de la probabilidad de invalidez para cada individuo y utilizaremos esta variable como input de nuestro modelo espacial de INLA. En el **¡Error! No se encuentra el origen de la referencia.** se aprecia un buen ajuste de las predicciones del modelo (línea roja) con la frecuencia real observada (línea azul). Las barras naranjas se corresponden con la exposición en cada tramo de edad.

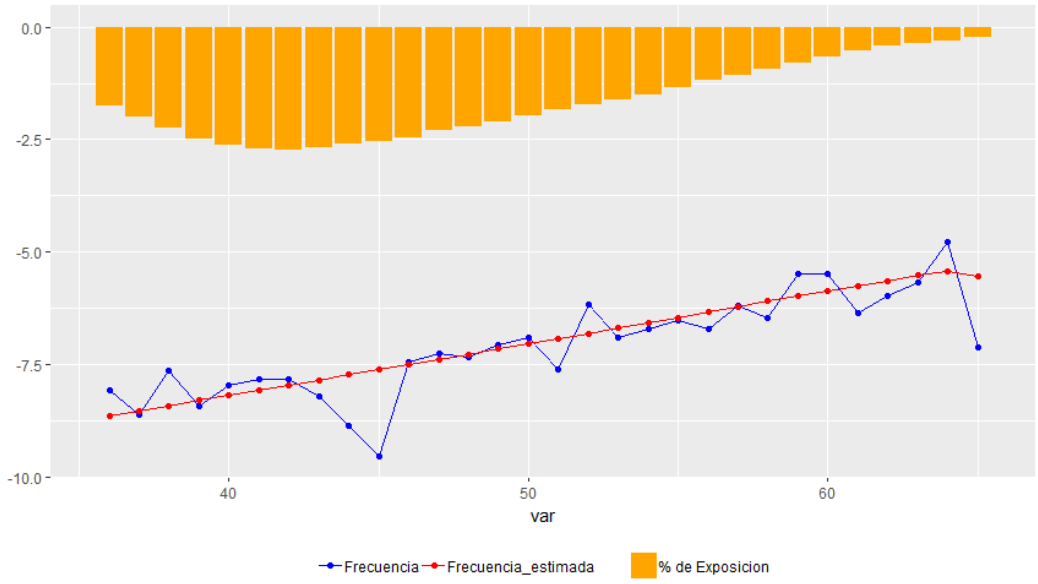

Gráfico 2: Real vs Predicción. Elaboración propia R-Studio.

La dependencia espacial supone la existencia de correlación entre observaciones próximas entre sí.

Como hemos mencionado en la primera parte del trabajo, los modelos lineales generalizados ignoran la existencia de correlaciones espaciales en los datos lo que nos puede llevar a resultados inconsistentes. Existen herramientas para la comprobación de la existencia de este tipo de variabilidad dentro de modelos como los GLM.

En este caso, comprobaremos la existencia de este tipo de dependencia en los residuos del modelo a través del test I de Moran.

La I de Moran es un coeficiente de correlación que mide la dependencia espacial general de un conjunto de datos.

La autocorrelación espacial es multidireccional y multidimensional, por lo que es útil para encontrar patrones en conjuntos de datos complejos. El análisis se realiza, por un lado, a través del gráfico de Moran (Gráfico 3) en el que se muestra el grado de dependencia espacial de una determinada variable. En este caso estamos realizando el test con los residuos del primer modelo GLM (no espacial) generado. En el eje de abscisas tenemos los valores del residuo medio dentro de cada polígono y en el eje de ordenadas tenemos los valores de los residuos medios de los vecinos de cada uno de esos polígonos. De tal forma que una correlación positiva indica que aquellos valores

inferiores del residuo se encuentran rodeados de áreas con residuos de cuantías homogéneas, y valores altos se encuentran rodeados de polígonos con valores altos.

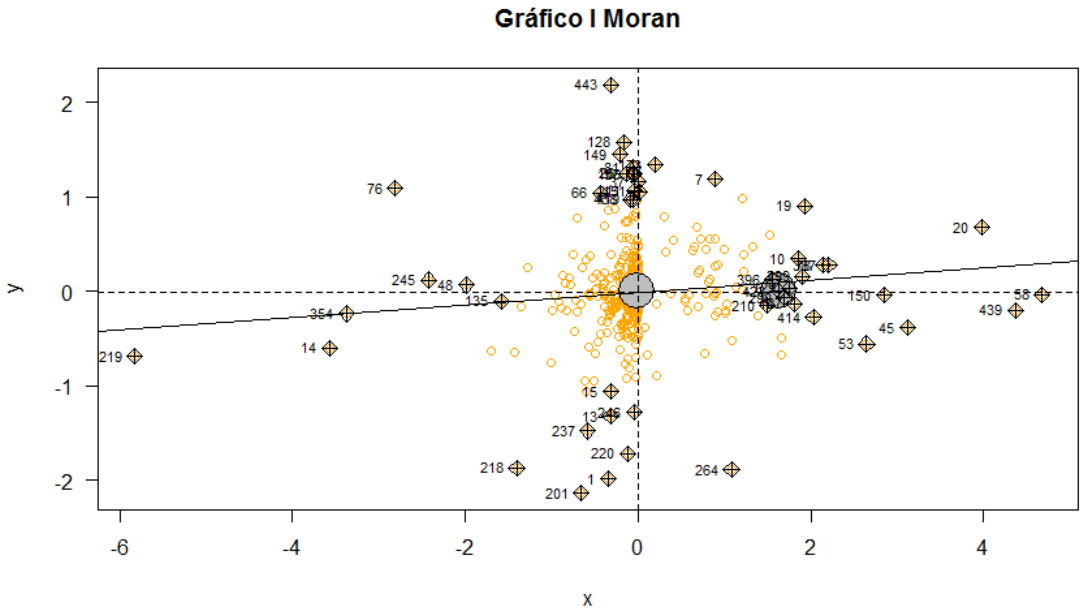

Gráfico 3: Gráfico test I de Moran. Elaboración propia software R-Studio.

Además, con el test de Moran se contrasta la hipótesis nula de distribución espacial aleatoria, comparando los valores de cada localización con los de sus vecinos. Si el valor p es estadísticamente significativo y el valor z (Ecuaciones 27 y 28) es positivo se puede rechazar la hipótesis nula, lo que significa que la distribución espacial de valores altos y bajos está más agrupada espacialmente de lo que se esperaría si los procesos espaciales que subyacen fuesen aleatorios.

H_0: No existe correlación espacial en la variable de interés.
H_α: Existe correlación espacial en la variable de interés.

El estadístico z se obtiene de la siguiente forma: Primero, se calcula la I de Moran siguiendo la fórmula

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \tag{27}$$

dónde z_i es la desviación de la observación i, $w_{i,j}$ es el peso espacial entre i y j, n es el número total de observaciones y S_0 es la suma de todos los pesos espaciales. Una vez calculado I, obtenemos el valor z de la siguiente forma:

$$z_I = \frac{I - E[I]}{\sqrt{V[I]}} \tag{28}$$

Los resultados de la Tabla 8 confirman la existencia de suficiente evidencia estadísticamente significativa para afirmar que la variable explicada posee correlación espacial a un nivel de significación del 5%. El p valor es inferior a 0,05 y, por lo tanto, rechazamos la hipótesis nula.

La correlación espacial del residuo del modelo, invalida las estimaciones del mismo y declara la inconsistencia e inexactitud de los estimadores calibrados por lo que llegamos a la necesidad de introducir en el modelo algún indicador de dependencia espacial.

Tabla 8: Test I de Moran

Estadístico I Moran	p-value
0,0650	0,0313

Elaboración propia. Software R-Studio

Alguna posible justificación de la existencia de dependencia espacial podría ser la presencia de factores geográficos muy distintos por ubicación. Además, la geolocalización podría estar ocultando otras variables relevantes como por ejemplo diferencias económicas Norte-Sur.

6.3.2 Modelo espacial bayesiano y metodología INLA para la invalidez

A continuación, ajustaremos dos modelos lineales generalizados con efectos espaciales a través de la metodología INLA.

Para ello, tomaremos los datos agregados por cada partición del área. Antes de la definición del modelo es necesario un archivo adicional con la información de los vecinos de cada partición geográfica. Se trata de un archivo con el formato ".graph", adecuado para ser leído por la función *inla*. En el **¡Error! No se encuentra el origen de la referencia.** se muestra una imagen de este archivo, dónde cada punto indica los vecinos de cada área.

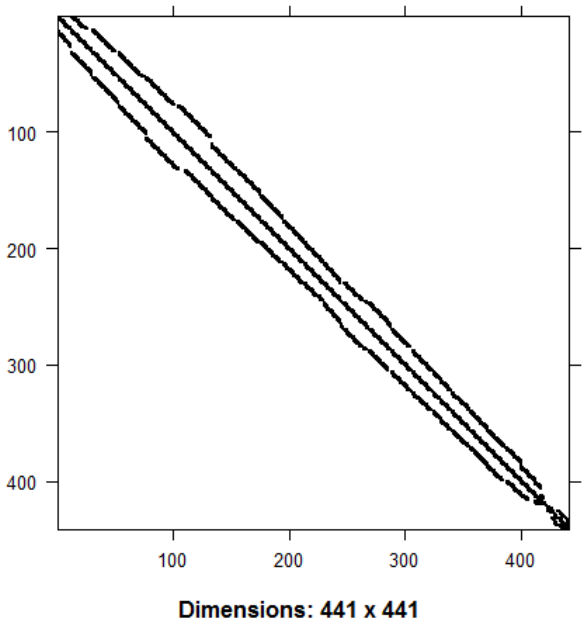

Gráfico 4: Matriz adyacente. Elaboración propia R-studio.

A continuación, se especifica el modelo que se ajustará.

Primero, asignaremos la familia de la función de probabilidad de nuestros datos empíricos y_i. Como para cada i-ésima área el número de invalidez permanente absoluta de individuos de entre 35 y 65 años presenta una gran cantidad de ceros (ver **¡Error! No se encuentra el origen de la referencia.**), se modela cómo una *zero inflated Poisson* de la siguiente forma:

$y_i=0$	con proabilidad p_i
$y_i \sim Poisson(\lambda_i)$	con probabilidad 1 - p_i

Cuya función de densidad consta de dos componentes que se describen :

$$P(y_i = y) = \begin{cases} p_i + (1 + p_i)e^{-\lambda_i}, & para\ y = 0 \\ (1 - p_i)\dfrac{e^{-\lambda_i}\lambda^y}{y!}, & para\ y > 0 \end{cases} \tag{29}$$

Denotaremos lo anterior cómo $y_i \sim$ ZIP (p_i , λ_i). El parámetro λ_i se define en términos del ratio ρ_i y el número de muertes esperadas por cada área E_i calculado ajustando el modelo glm del apartado anterior. Por lo tanto, $\lambda_i = \rho_i E_i$.

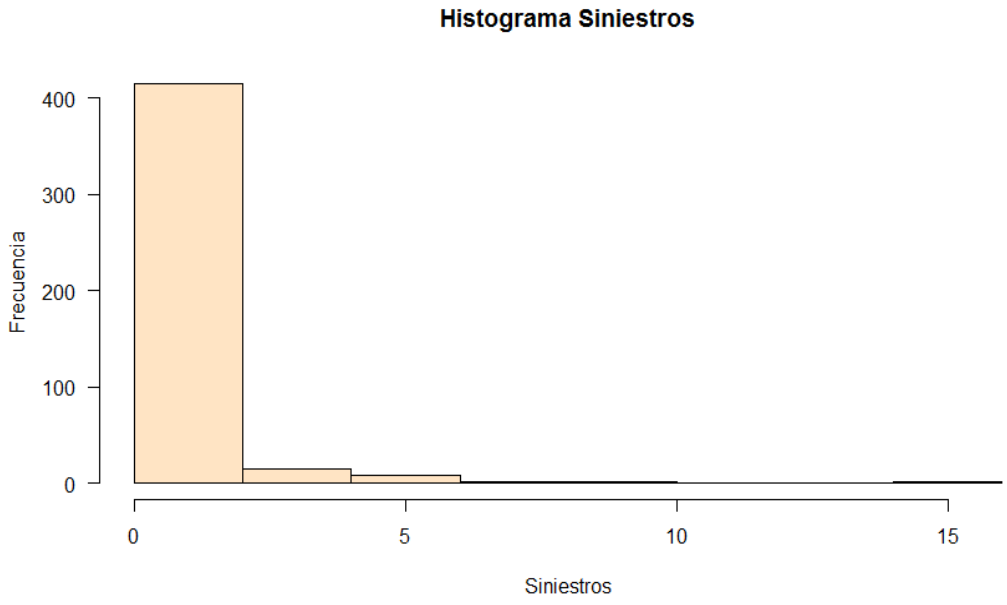

Gráfico 5: Histograma variable y_i .Elaboración propia R-studio.

El primer modelo que ajustaremos a través de la metodología INLA se define

$$\eta_i = \alpha + u_i + v_i \tag{30}$$

dónde α es el intercepto que cuantifica el ratio de invalidez medio global del conjunto de las 441 áreas; $u_i = f_1(i)$ y $v_i=f_2(i)$ son los dos efectos latentes espaciales; i={1,…,n} el indicador de cada área (que se corresponde con la variable ID de la base de datos).

La variable u_i será el componente espacial estructurado que representa el efecto de los vecinos de primer orden sobre cada área, capturando así las similitudes entre las áreas.

Para los efectos aleatorios espaciales estructurados asumimos la especificación Besag-York-Mollie (Besag, 1991) que asigna una distribución a priori autorregresiva Normal intrínseca (iCAR):

$$u_i \mid u_{j\neq i} \sim Normal(m_i, s_i^2)$$

$$m_i = \frac{\sum_{i\in \mathcal{N}(i)} u_j}{\#\mathcal{N}(i)}$$

$$s_i^2 = \frac{\tau_u^2}{\#\mathcal{N}(i)},$$

(31)

Dónde #ℕ(i) es el número de áreas que comparten frontera con el área i-ésima (lo obtenemos a través del archivo .graph).

El parámetro v_i representa los residuos desestructurados. Se trata del efecto específico de cada área independientemente del resto de áreas y, por tanto, captura la heterogeneidad entre áreas. En este caso, se asigna una distribución a priori normal i.i.d con media cero y precisión desconocida τ_v.

El modelo anterior puede especificarse en R-INLA con la opción f (ID, model="bym")[7], y los efectos espaciales se parametrizan cómo $\xi = u_i + v_i$. La importancia de los parámetros aleatorios dependerá de sus desviaciones típicas.

Así, el conjunto de parámetros estimados con R-INLA serán $\Theta = \{\alpha, \xi\}$ y el conjunto de hiperparámetros de precisión que capturarán la magnitud del efecto aleatorio $\varphi = \{\tau_u^2, \tau_v^2\}$ a los que asignamos la siguiente distribución a priori[8]:

[7] Otra alternativa es especificar los dos componentes con dos funciones separadas: por un lado, los efectos estructurados usando f(ID, model="besag") y para los desestructurados f(ID2, model="iid").

$$\tau_u \sim \text{loggama}(1, 0.001)$$

$$\tau_v \sim \text{loggama}(1, 0.001)$$

La Tabla 9 muestra el resumen de las estimaciones posteriores para el parámetro del único efecto fijo del modelo, el intercepto (α). La media posterior global del conjunto de todas las áreas se calcula como la exponencial del intercepto. El resultado es un 0,9% de ratio de invalidez medio de la cartera: 1- exp(0,008)=0,009.

Tabla 9: Resultados efectos fijos

	Mean	Sd	0.025quant	0.5quant	0.975quant	mode
α (Intercept)	0,008	0,058	-0,109	0,009	0,121	0,010

Fuente: Elaboración propia R-studio

El **¡Error! No se encuentra el origen de la referencia.** representa la distribución del efecto aleatorio espacial total ξ para el área 1 (línea negra) y el área 440 (línea naranja).

[8] Por defecto, INLA asigna precisiones distribuciones a priori loggama(1,0.001) para el intercepto, los efectos fijos y los efectos aleatorios del modelo (Rue,H. and Martino,S.(2009))

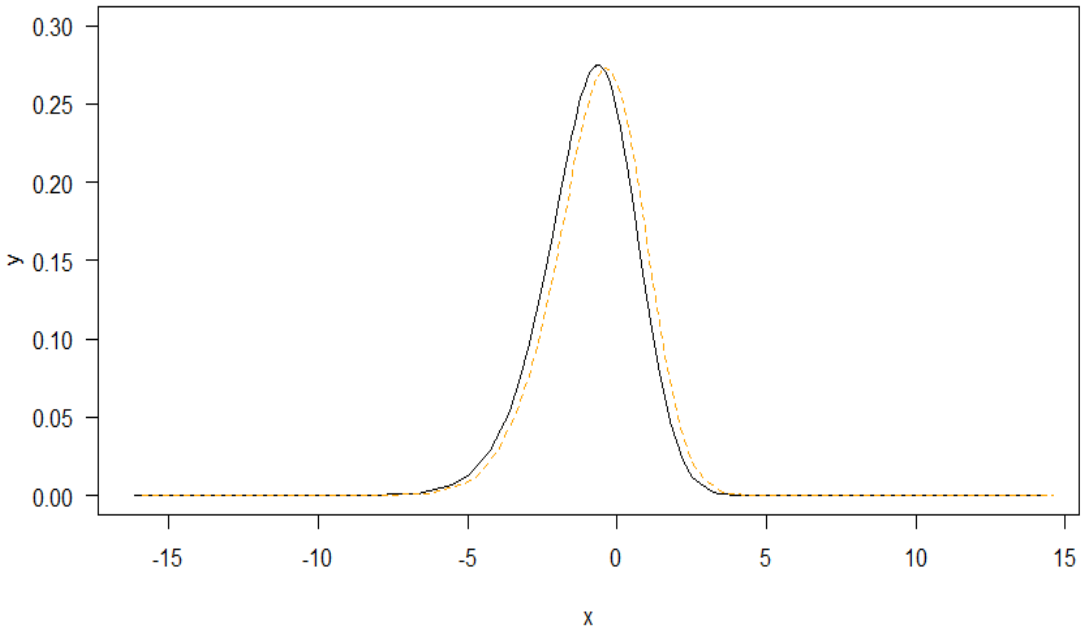

Gráfico 6: Efecto Aleatorio (u+v). Elaboración propia R-Studio.

La Figura 2 muestra el mapa de la media de los riesgos relativos posteriores específicos de cada área con respecto al global de España. Las distribuciones a posteriori se obtienen fácilmente con una transformación exponencial de los componentes de ξ. Los polígonos amarillos y rosas se corresponden con mayores niveles de riesgo relativo. Por otro lado, la Figura 3 expone la distribución de probabilidad posterior específica de cada área, es decir, la probabilidad de que el riesgo relativo de cada área sea mayor que el riesgo global dados los datos empíricos.

En ambas figuras se aprecian determinados clusters de riesgos relativos de invalidez. Las áreas con mayor riesgo coinciden con las Comunidades de Madrid, Pais Vasco y Cataluña; las zonas más industrializadas y de mayor actividad económica. Otras áreas en las que el riesgo es superior a la media son las Islas Baleares y Andalucía.

En el Gráfico 7 se observa la distribución de los riesgos relativos a posteriori: se aprecia una forma asimétrica y concentración de la masa en torno a la media global (RR=1).

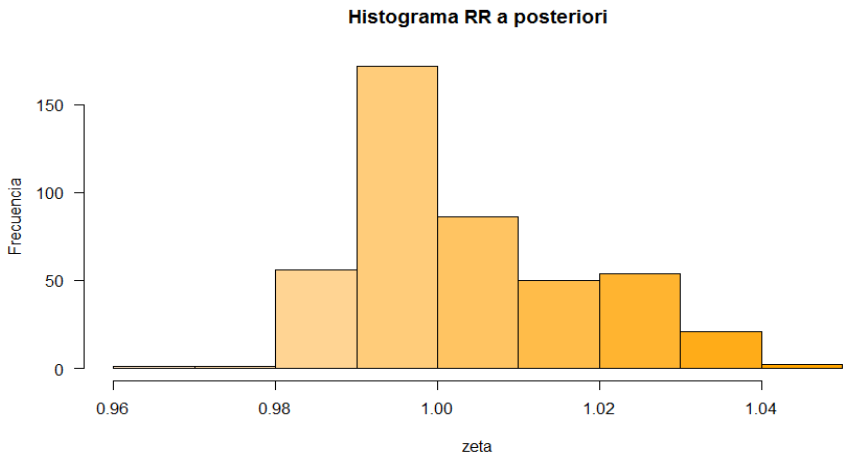

Gráfico 7: Distribución Riesgos Relativos a posteriori. Elaboración propia. R-Studio.

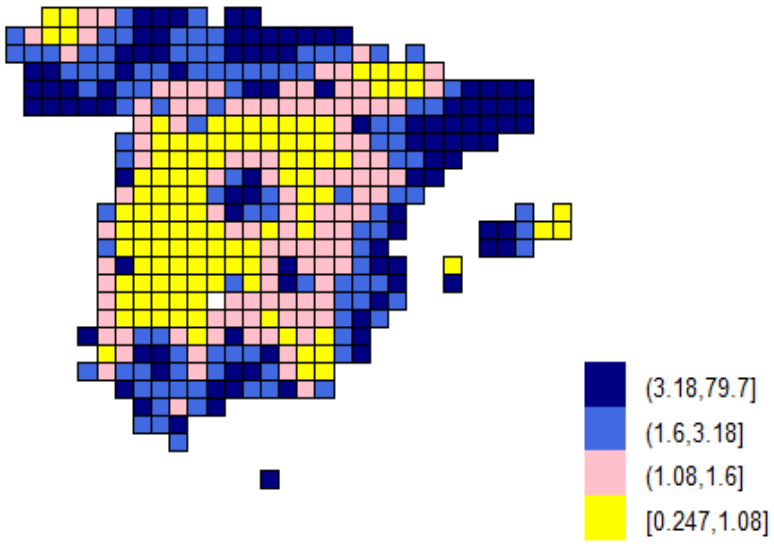

Figura 2: Mapa de Riesgos Relativos a posteriori: Distribución de los riesgos relativos de mortalidad de las áreas. Elaboración propia R-Studio.

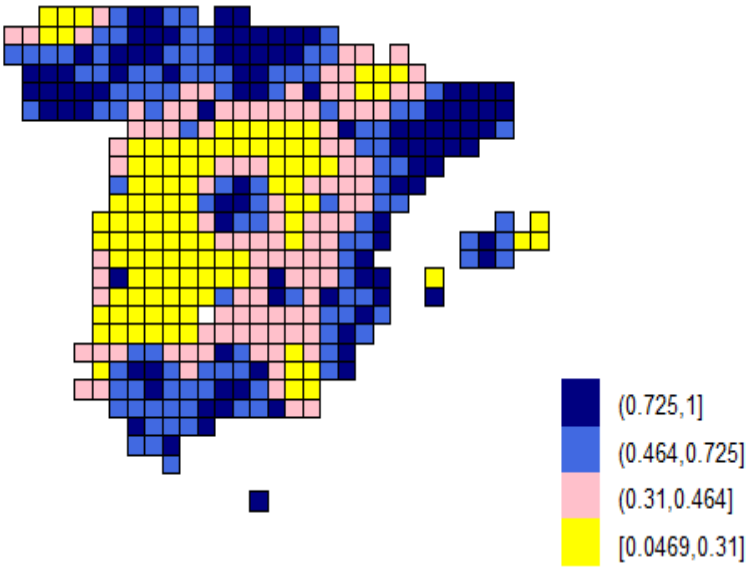

Figura 3: Distribución de la probabilidad posterior específica de área p(ui+vi>1|y). Elaboración propia R-Studio.

Es interesante evaluar también la proporción de variación explicada por el componente estructurado espacial, es decir, aquel que captura los efectos de los vecinos sobre el área. La cantidad σ_u^2 es la varianza de la especificación autorregresiva condicional (CAR), mientras que σ_v^2 es la varianza del componente desestructurado. Estos, no son directamente comparables pero se puede obtener una estimación de la varianza posterior marginal del efecto estructurado de la siguiente forma

$$s_u^2 = \frac{\sum_{i=1}^{n}(u_i - \bar{u})^2}{n-1} \tag{32}$$

dónde $\bar{u}$ es la media de u, para luego compararla con la varianza posterior marginal del efecto desestructurado proporcionado por σ_v^2

$$frac_{\ spatial} = \frac{s_u^2}{(s_u^2 + \sigma_v^2)} \tag{33}$$

En nuestro caso, la proporción de variabilidad espacial estructurada respecto a la heterogeneidad es de un 90%. Esto significa que hay una fuerte relación espacial de la incapacidad en España traducida en una dependencia espacial significativa con los vecinos.

Siguiendo el mismo procedimiento, construiremos de nuevo el modelo incorporando, además de los efectos espaciales, un efecto fijo: la edad actuarial. Esto nos permitirá medir su impacto sobre el riesgo de invalidez y compararlo con el modelo espacial.

Este segundo modelo será de la siguiente forma:

$$\eta_i = \alpha + \beta_1 x_{1i} + u_i + v_i \tag{34}$$

dónde x_{1i} se corresponde con la variable de edad actuarial y β_1 el efecto fijo asociado.

En la Tabla 2 presentamos los resultados de los efectos fijos de este nuevo modelo. Si los exponenciamos, se interpretan como riesgos relativos: incrementos de la edad de un año suponen incrementos del ratio de riesgo de invalidez en un 4%=1-exp (0.042) y con un intervalo de confianza al 95% este efecto se situará entre el 2% y el 6%.

Tabla 10: Resultados efectos fijos

	mean	Sd	0.025quant	0.5quant	0.975quant	mode
α (Intercpt)	-6.197	0.512	-7.211	-6.193	-5.204	-6.186
β1 (age)	0.042	0.009	0.025	0.042	0.059	0.042

Fuente: Elaboración propia. R Studio

Nuevamente, estimamos los riesgos relativos a posteriori para cada área con respecto al global y los pintamos en un mapa (Figura 4).

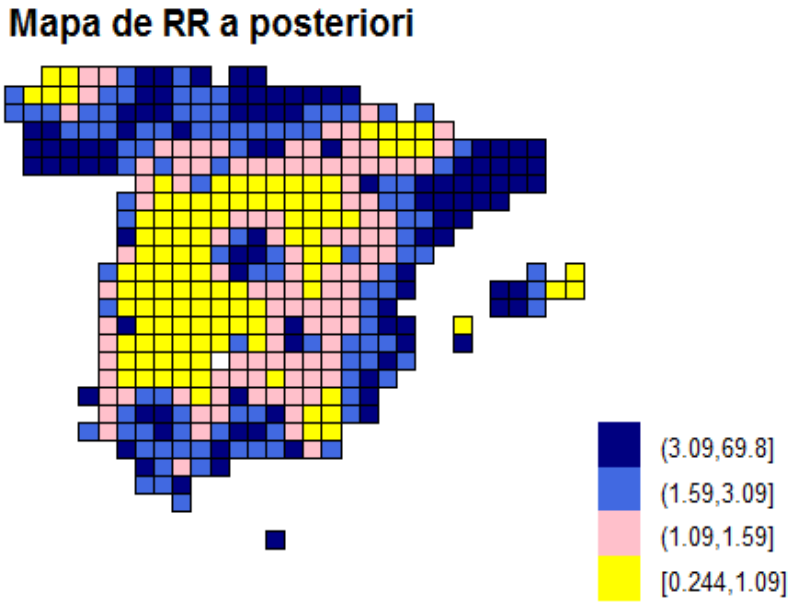

Figura 4: Mapa Riesgos Relativos a posteriori Modelo 2. Elaboración propia. Software R-studio.

A simple vista, la distribución de los riesgos en el mapa no cambia mucho con respecto al primer modelo.

En cuanto al porcentaje de variación explicado por la parte estructurada se mantiene en el 90%.

Tal y como hemos visto en el Capítulo 5, una posible herramienta para evaluar y comparar el ajuste de diferentes modelos estimados con INLA es el Criterio de Información de desviación (DIC).

La Tabla 11 expone los componentes DIC para los dos modelos. El segundo modelo, que incluye el efecto lineal de la edad y el efecto aleatorio espacial, muestra el DIC más pequeño lo que sugiere que, a pesar de la complejidad agregada, es el más adecuado.

Tabla 11: Resumen DIC

Modelo	D	p_D	DIC
Modelo 1	1983,19	113,91	1869,28
Modelo 2	1851,02	89,77	1851,02

Fuente: Elaboración propia. Software R-studio

Para comparar los modelos en cuanto a su capacidad predictiva calculamos el score logarítmico de cada uno de ellos. Para ello, necesitaremos los resultados de los CPO (conditional predictive ordinate) de los modelos, que se calculan directamente usando la función *inla*.

Un valor menor del score logarítmico indica una mejor calidad predictiva del modelo. Por lo tanto, como se puede comprobar en la, el modelo 2 es el más apropiado para realizar predicciones.

Tabla 12: Resultado LogScore

Modelo	LogScore= - mean(log(CPO))
Modelo 1	0.094
Modelo 2	0.090

Fuente: Elaboración propia. R-Studio

6.3.3. Ampliación Modelo Lineal Generalzado

El GLM es la principal herramienta de las las compañías aseguradoras para modelos de predicción y *pricing*. Es importante que los resultados proporcionados bajo esta metodología de modelización sean lo más robustos posibles. Uno de los problemas que vimos acerca de los GLM era la dificultad para recoger las correlaciones espaciales que pueden presentar los datos observados. Con el objetivo de comprobar si podemos eliminar la dependencia espacial en los residuos del modelo GLM (probada en el primer bloque de este capítulo con el test I de Moran) a la vez que mejoramos el ajuste, tomaremos la variable de riesgo relativo estimada con el primer modelo INLA (que captura las relaciones geográficas de los datos observados a través de las variables de efectos aleatorios) y la incluimos en el modelo como otra variable explicativa de la probabilidad de invalidez., junto con la edad actuarial.

El nuevo modelo será de la siguiente forma:

$$\Phi^{-1}(\pi_i) = \eta_i = \beta_0 + \beta_1 fedad\_actuar_{1i} + \beta_2\, RR + NumExp \tag{35}$$

siendo RR la variable de riesgo relativo que captura los efectos espaciales obtenida a través de la modelización INLA.

En la Tabla 13 se puede comprobar que ambas variables son muy significativas, con un p - valor por debajo de 0,01. Tal y cómo habíamos visto en el primer GLM, la variable edad actuarial guarda una relación positiva con el riesgo de invalidez. La variable RR también tiene impacto significativamente positivo.

Tabla 13: Resultados GLM 2

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0,1350	-0,0459	-0,0304	-0,019	4,3253

Coefficients:

	Estimate	std.Error	z value	Pr(>|z|)	
(Intercept)	-6,0754	0,1571	-38,6660	<2e-16	***
fedad_actuar	0,0338	0,0028	12,0600	<2e-16	***
RR	0,1030	0,0156	6,5940	4.28e-11	***

Fuente: Elaboración propia. Software R-Studio

Los resultados de la curva ROC demuestran un mejor ajuste de este segundo modelo frente al primero. En este caso, el área debajo de la curva es de 0.82, es decir, se ha incrementado la probabilidad de acierto en 7 puntos.

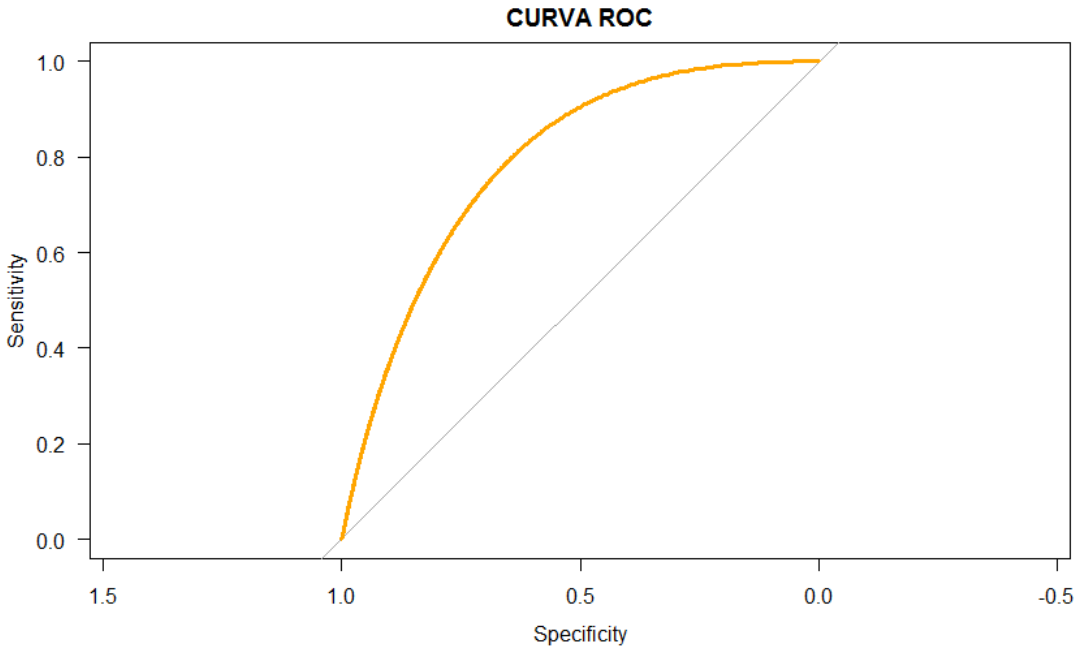

Gráfico 8: Curva ROC GLM 2. Elaboración propia software R-studio.

Por último, repetimos el test de I de Moran sobre el nuevo modelo. En la Tabla 14 se observa un p – valor por encima del 5%, lo que indica que no podemos rechazar la hipótesis nula y que, por lo tanto, los valores podrían ser el resultado de procesos puramente aleatorios. En definitiva, no podemos afirmar que exista dependencia espacial en nuestros datos.

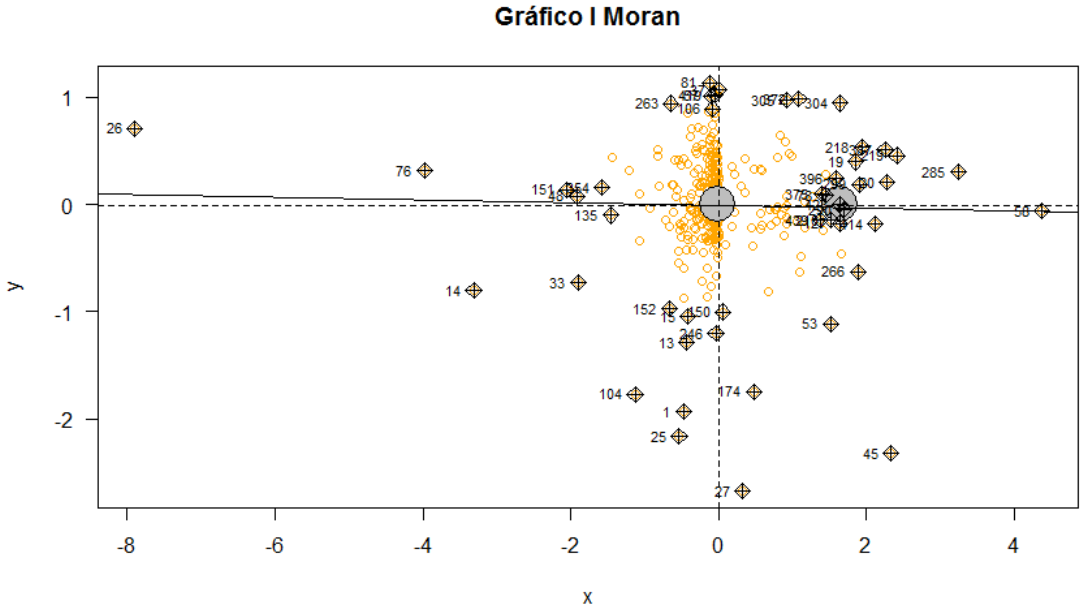

Gráfico 9: Grafico test I de Moran GLM 2. Elaboración propia con R-Studio.

Tabla 14: Resultados test I de Moran

Estadístico I Moran	p-value
-0,0128	0,6165

Fuente: Elaboración propia. R Studio

Con la incorporación de una nueva variable en el modelo GLM que captura la dimensión espacial de nuestros datos hemos logrado, además de una mejorara significativa del ajuste, recoger la estructura de interrelaciones espaciales lo que nos llevará a resultados más consistentes y robustos.

7. CONCLUSIONES

El objetivo de este estudio no es tanto encontrar el mejor modelo para ajustar el riesgo a partir de un conjunto de datos, sino proponer una nueva metodología para la modelización de fenómenos asociados a procesos espaciales. Contrastar esta metodología con datos reales, en nuestro caso de invalidez en España, permite reconocer la importancia de tener en cuenta las dependencias espaciales.

Hemos trabajado con dos tipos de modelización para el estudio de la invalidez. Primero, y por ser uno de los métodos de modelización más utilizados por las aseguradoras en España, se presenta un modelo de regresión lineal generalizado con una variable explicativa (la edad actuarial). La construcción de este modelo cumple principalmente dos objetivos: por un lado, nos permite obtener tasas globales esperadas de invalidez por edad que nos servirán como input para la posterior construcción de los modelos bayesianos. Además, con los resultados proporcionados por el test de I de Moran se puede comprobar como un modelo aparentemente bien ajustado, es inadecuado debido a la existencia de cierta dependencia espacial en sus resultados que no se está capturando de ningún modo.

Para solventar este problema y tener en cuenta los efectos de la estructura geográfica, trabajamos con un modelo bayesiano de tipo espacial al que se le pueden incluir tanto efectos fijos como efectos aleatorios, siendo estos últimos los que modelicen los efectos de espacio.

La metodología INLA permite extraer aproximaciones rápidas y precisas de las distribuciones marginales a posteriori además de otras muchas medidas estadísticas cómo las medias a posteriori, desviaciones a posteriori, cuartiles, etc. Una de sus ventajas es que se puede realizar de manera sencilla con un paquete de R que incorpora numerosas funciones de interés.

Otra de las ventajas de esta metodología y de los análisis Bayesianos es la posibilidad de incorporar distribuciones previas para la estimación de los parámetros, lo que incorpora mucha información y permite estimaciones más coherentes cuando tratamos con bases de datos reducidas, por ejemplo en nuestro caso, datos de pequeñas áreas.

Construimos dos modelos: el primero tiene en cuenta sólo los efectos aleatorios espaciales estructurados (efectos de influencia de los vecinos) y desestructurados

(efectos heterogéneos de las áreas) como factores explicativos, y al segundo le incorporamos, además de los efectos aleatorios espaciales, la edad actuarial como un efecto fijo. Mapeando los riesgos relativos derivados de ambos modelos comprobamos que los niveles de riesgo relativo de invalidez más altos coinciden con las zonas más industrializadas, lo que da robustez al estudio. Además, el segundo, a pesar de la complejidad añadida, muestra un DIC más bajo por lo que concluimos que su ajuste es más adecuado.

En definitiva, con estos modelos se demuestra cómo la metodología INLA supone una poderosa herramienta para la inferencia de modelos en los que existen dependencias espaciales.

Para completar el estudio, combinamos los dos tipos de modelización vistos. Se crea una variable de riesgo relativo, derivada del modelo espacial bayesiano, que captura las relaciones espaciales y la incorporamos a la base de datos principal con el fin de que funcione como un indicador de la dependencia espacial. Se incorpora esta nueva variable al modelo GLM inicial obteniendo cómo resultado ambas variables significativas y un mejor ajuste del modelo. Además unos mejores resultados en cuanto al ajuste del modelo, realizando el test I de Moran se comprueba la eliminación de la correlación espacial que se había probado al principio.

Debido a su reciente creación, la metodología INLA está menos establecida que los métodos de simulación MCMC. En consecuencia, su desarrollo sigue en curso. Al mismo tiempo, es importante notar que la creciente popularidad de INLA está generando numerosos avances y nuevos complementos que se incluyen en el paquete de R-INLA y amplían y facilitan la inferencia de modelos bayesianos.

8. AMPLIACIONES

Los modelos propuestos sirven para demostrar la posibilidad y la importancia de emplear metodologías que tengan en cuenta la variabilidad espacial. Existen muchas formas de ampliar los modelos y conseguir resultados de gran interés de cara al estudio y estimación de los riesgos, en este caso, de invalidez. Ninguno de los propuestos considera más de una covariable observada del fenómeno. La inclusión de más covariables, o bien de datos concretos de cada póliza (como el sexo, capital, canal, etc.), o variables socioeconómicas a nivel de área (como la tasa de desempleo), podrían suponer una mejora en la interpretación del fenómeno y aportar mayor generalidad a la modelización.

Otra forma de ampliarlo sería considerando además de la dependencia espacial, también la dimensión temporal permitiendo recoger las tendencias y la interacción entre espacio y tiempo. Este tema podría abordarse a través de los modelos espacio-temporales explicados en el Capítulo 2.

En cuanto a la propia metodología, existe un amplio campo de opciones de mejora y desarrollo. Hemos presentado la metodología INLA centrándonos en modelos cuyos efectos latentes son un Campo Aleatorio Gaussiano Markoviano (GMRF). No obstante, existen estudios recientes (Gómez, V. y Rue, H. 2017) que buscan alternativas para ampliar el número de posibles modelos que pueden ser aproximados utilizando INLA. Estos autores demuestran como ajustar los valores de los parámetros mediante modelos condicionales ajustados con INLA y con algoritmos estándar MCMC como el Metropolis-Hastings.

BIBLIOGRAFÍA

Adin, A. (2017). Hierarquical and spline-based models in space time disease mapping. Universidad Pública de Navarra.

Besag, J., York, J., Mollie, A., (1991). Bayesian Image restoration, with two applications un spatial statistics. Annals of the Institute os Statistical Mathematics 43, 1-59

Blangiardo, M.m Cameletti, M.m (2012). Bayesian Spatio and Spatio-Temporal Models with R-INLA. Wiley.

Coromoto, N. (2013). Modelos Jerárquicos Bayesianos espaciales en epidemología agrícola . Universidad de Valencia.

Gómez, V. (2019). Bayesian inference with INLA and R-INLA.

Gómez, V. Rue, H. (2017). Markov Chain Monte Carlo with the Integrated Nested Laplace Approximation. Department of Mathematics, School of Industrial Engineering. Universidad de Castilla la Mancha.

Khana, D., Rossen, L., Hedegaard, H., Warner, M. (2018). A Bayesian Spatial and temporal modeling approach to mapping geographic variation in mortality rates for subnational areas with R-INLA. Journal of Data Sciencie 18, 147182.

Lindgren, F., Rue, H. Bayesian Spatial Modelling with R-INLA. Journal of Statistical Software.

Machiavelli, R., Torres-Saavedra P. (2018). Modelos Estadísticos avanzados aplicados a la investigación en salud y ambiente. Facultad de Ciencias Exactas y Naturales. Universidad Nacional de Buenos Aires.

Martino, S., Rue, H., (2010). Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a manual for the INLA program.

Martino, S., Rue, H. Case studies in Bayesian Computation using INLA.

Martins, G., Simpson, D., Lindgren, F., Rue, H., (2012). Bayesian computation with INLA: new features. Norwegian University of Science and Tchnology Report.

Opitz, T. (2017). Latent Gaussian modeling and INLA: A review with focus on space.time applications. Journal de la Societe Française de Statistique.

Rue, H., Martino, S., Chopin, N. (2009). Approximate Bayesian Inference for latent Gaussian models by using Integrated nested Laplace approximations. Journal of the Royal Statistical Society.

Rue, H., Held, L., (2005). Gaussian Markov Random Fields. Theory and Apliccations. Chapman & Hall.

ANEXO I: CÓDIGO R

```
################################################################################
#################### ESTIMACIÓN DE LA MORTALIDAD EN ESPAÑA ######################
################################################################################

########### Cargamos las librerías #############

require (INLA)
require (INLA)
require (raster)
require (mgcv)
library(rgdal)
library(lubridate)            # READING     : Trabajar con Horas
library(foreign)              # READING     : Leer DBFs Files
library(sp)                   # SPATIAL     : Merge Points and Polygons Over
library(rgdal)                # READING     : Leer ShapeFiles
library(dplyr)                # FILTROS     : SQL en R
library(maptools)             # READING     : Leer ShapeFiles
library(ggmap)                # SPATIAL     : Visualizar Spatial Data
library(tmap)                 # SPATIAL     : Visualizar y crear Mapas
library(spdep)                # SPATIAL     : Create Neighbourhoods y Diferentes Test
library(rgeos)                # SPATIAL     : Calcular Centroides
library(ggplot2)              # PLOT        : Graficos espacializados
library(mgcv)                 # REGRESION   : GAM Modelos GAM
library(rsatscan)             # SPATIAL     : Software SatScan Clustering Spacial
library(SpatialEpi)           # SPATIAL     : Clustering Spacial
library(pROC)                 # REGRESION   : Calculo de ROC
library(Matrix)               # READING     : Hacer Matrices Sparce
library(splines)              # REGRESION   : Hacer Splines
library(earth)                # REGRESION   : Metodología MARSplines
library(stats)                # REGRESION   : Autoregresive Regression Models
library(ROCR)                 # REGRESION   : Calculo ROC para Árboles
library(smerc)                # SPATIAL     : Spatial Clustering
library(rvest)
library(raster)
require(maptools)
require(plotrix)
library(data.table)


######## Asignamos ruta de trabajo y creamos variable ########

setwd("C:/Users/ebt0629/Desktop/TFM/Modelo INLA")
remove(list=ls())
my.dir<-paste(getwd(),"/",sep="")

######## Cargamos algunas funciones que vamos a necesitar########

## 1. Análisis univariante de impacto de variable respuesta en todas las varibales.
Histograma.
histogramas_binomial<-function(X,response,tr=8,up=1,ya=0.5,round=2,mult=1){
  yy<-ya;  xx<-2
  for (i in 1:ncol(X)) {
    print(i)
    X$response<-response
    ddd<-class(X[,i])

    if (!is.numeric(X[,i])){
      X[,i]<-as.numeric(as.factor(X[,i]))  }

    if (length(table(X[,i]))>1){

      q1<-as.numeric(quantile(X[,i],0.05))
      q2<-as.numeric(quantile(X[,i],0.95))
      min<-round(min(X[,i])-0.2,2);  max<-round(max(X[,i])+0.2,2)
      q<-c(min,seq(q1,q2,(q2-q1)/tr),max);  q<-unique(round(q,round))

      a<-as.character(cut(X[,i],breaks=q))
      cc<-colnames(X)[i] ; cc<-paste(cc,length(table(X[,i])),ddd);cc<-colnames(X)[i]
      b<-response
      bind<-cbind(X,a)
      ag<-aggregate(bind$response, by=list(Category=bind$a), FUN=sum)
      f<-as.data.frame(table(a))
```

```
      ag2<-cbind(ag,n=f$Freq,lap=(ag$x/f$Freq)*mult)

      mm<-as.numeric(as.character(substr(ag2$Category,2,regexpr(",",ag2$Category)-1)))

      ag2<-cbind(ag2,mm); ag2<-ag2[order(ag2[,5]),]; ag2$ID<-seq.int(nrow(ag2))


barplot(prop.table(ag2$n),cex.names=yy,cex.main=xx,names.arg=ag2$Category,beside=T,main=
cc,ylim=c(0,up),las=2,space=0.06)
      points(ag2$lap, col="red");     abline(lm(ag2$lap~ag2$ID))
    }}}

## 2.Curva ROC para valoración modelo GLM
ROC<-function(real,esperado){

  plot(roc(real, esperado, direction="<",smooth = TRUE),
       col="orange", lwd=3, main="CURVA ROC", las=1)

  c<-auc(real,esperado)
  fff<-as.data.frame(cbind(real,esperado))
  p8 <- ggplot(fff, aes(x = esperado, fill = as.factor(real))) +
    geom_density(position="identity", alpha=0.6) +
    scale_x_continuous(name = "Probability Expected") +
    scale_y_continuous(name = "Density") +
    coord_cartesian(xlim=c(0,1))+
    ggtitle("Model Performance") +
    theme_bw() +
    theme(plot.title = element_text(size = 14, family = "Tahoma", face = "bold"),
          text = element_text(size = 12, family = "Tahoma")) +
    scale_fill_brewer(palette="Accent")

  p8

  k<-list(p8,c)
  return(k)

}

## 3. Función para sacar plot de real vs predicción
AE_1<-function(bbdd_fff,predicted,var,inc=1){

  cols <- c("% de Exposicion" = "orange", "Frecuencia" =
"blue","Frecuencia_estimada"="red")
  q1<- bbdd_fff %>% mutate(fq_estimada = predicted) %>%
    group_by_(paste(var)) %>%
    summarise(Exposicion = n(),
              Frecuencia = log(sum(response)/sum(VIVO)),
              fq_estimada = log(sum(fq_estimada)/sum(VIVO)))

  q1<-as.data.table(q1)

  c1<-q1[,1]
  q1$c1<-q1[,1]

  ggplot(q1,aes(x=c1)) +
    geom_bar(aes(y=(Exposicion/sum(Exposicion))*inc, fill="% de Exposicion"), stat =
"identity")+
    geom_point(aes(y=Frecuencia, colour="Frecuencia"), group=1)+
    geom_line(aes(y=Frecuencia, colour="Frecuencia"), group=1)+
    geom_point(aes(y=fq_estimada, colour="Frecuencia_estimada"), group=1)+
    geom_line(aes(y=fq_estimada, colour="Frecuencia_estimada"), group=1)+
    xlab("var") + ylab("") +
    scale_colour_manual("", values=cols) +
    scale_fill_manual("",values=cols) +
    theme(legend.key=element_blank(),legend.title=element_blank(),
          legend.box="horizontal", legend.position = "bottom") +
    ggtitle("")

}

## 4. Test IMORAN de Autocorrelación espacial entre variables. Salida Gráfico
IMoran_P<-function(Geo,kk=8){

  Geo@data <- as.data.frame(sapply(Geo@data, function(x) as.numeric(as.character(x))))
```

```
  #Geo_nbr <- poly2nb(Geo, queen = FALSE)

  Geo@data$LONG_IND<-coordinates(Geo)[,1]
  Geo@data$LAT_IND<-coordinates(Geo)[,2]

  co<-cbind((Geo@data$LONG_IND),Geo@data$LAT_IND)
  Dist<-knearneigh(co, k=kk, longlat=TRUE)
  Geo_nbr<-knn2nb(Dist)


  Dist2<-nb2mat(Geo_nbr,zero.policy=TRUE)
  Dist3<-mat2listw(Dist2)

  Geo@data<-dplyr::select(Geo@data,-LONG_IND,-LAT_IND)

  for (i in 1:ncol(Geo@data)){

    jj<-moran.plot(Geo@data[,i], Dist3, zero.policy = TRUE,main=colnames(Geo@data[i]))
    jj

  }

}

## 5. Test IMORAN de Autocorrelación espacial entre variables. Salida Test
IMoran_T<-function(Geo,kk=8){

  Geo@data <- as.data.frame(sapply(Geo@data, function(x) as.numeric(as.character(x))))

  #Geo nbr <- poly2nb(Geo, queen = FALSE)

  Geo@data$LONG_IND<-coordinates(Geo)[,1]
  Geo@data$LAT_IND<-coordinates(Geo)[,2]

  co<-cbind((Geo@data$LONG_IND),Geo@data$LAT_IND)
  Dist<-knearneigh(co, k=kk, longlat=TRUE)
  Geo_nbr<-knn2nb(Dist)


  Dist2<-nb2mat(Geo_nbr,zero.policy=TRUE)
  Dist3<-mat2listw(Dist2)

  Geo@data<-dplyr::select(Geo@data,-LONG_IND,-LAT_IND)

  for (i in 1:ncol(Geo@data)){

    jj<-moran.test(Geo@data[,i], Dist3, zero.policy = TRUE)
    print(jj)

  }

}

######## Cargamos el fichero shape: Nivel Código Postal #########
Geo_CP <-readOGR(paste(my.dir, "0_Codigo_Postal_2016/Capa_CP_Poligonos.shp",sep=""))
Geo_CP@data$CP2<-(as.numeric(as.character(Geo_CP@data$CP)))
Geo_CP@data$CP2<-floor(Geo_CP@data$CP2)
Geo_CP   <- spTransform(Geo_CP, CRS("+proj=longlat +datum=WGS84"))


#########################################################################
############### PARTE 0: PREPARACIÓN DE LA BBDD ###################
#########################################################################

####### 1º Leemos los datos #########
db<-read.csv("IPA.csv",sep=";")
db$Siniestro<-db$Siniestro_IPA
db$fedad_actuar<-db$FEDAD_ACTUAR
db$Capital<-db$Capital_IPA
db$NumExp<-db$NumExpIPA


####### 2º Establecemos los filtros ##
db<-dplyr::filter(db,YEAR_PROC>=2016,YEAR_PROC<2019,fedad_actuar>35,fedad_actuar<66)
nrow(dplyr::filter(db,is.na(CPCLI) | CPCLI==0))/nrow(db)
```

```
db<-dplyr::filter(db,!is.na(CPCLI),CPCLI>=1000)
sum(db$Siniestro)
db$Codigo<-db$CPCLI

####### 3º Creamos la BBDD espacial ##

###### Geolocalizamos BBDD #######
#################################

Centroides<-gCentroid(Geo_CP, byid=TRUE)
Centroides<-as.data.frame(cbind(Centroides@coords,Geo_CP@data$CP2))
Centroides$Codigo<-Centroides$V3
db<-merge(db,Centroides, by="Codigo", all.x=TRUE)
db<-dplyr::filter(db,y>0)

###### Creamos BBDD espacial ######
#################################

db_sp<-db
coordinates(db_sp)<- c("x","y")
proj4string(db_sp) <- proj4string(Geo_CP)

###### Rasterizamos BBDD #########
#################################

raster <- raster(ncol = 50, nrow = 50)
extent(raster) <- extent(db_sp)
nuevo_raster<-rasterize(db_sp, raster, db_sp$Capital, fun = mean)
plot(nuevo_raster)

Geo<-rasterToPolygons(nuevo_raster)
Geo@data$ID<-1:nrow(Geo@data)
plot(Geo)

###### Pasamos ID a BBDD ##########
#################################

db_sp$ID <- over(x = db_sp, y = Geo)$ID
db<-db_sp@data

#### Creamos Matriz de Conexiones #
#################################

zzz <- poly2nb(Geo)
nb2INLA("Geo.graph", zzz)
Geo.adj <- paste(getwd(),"/Geo.graph",sep="")

#########################################################################
############ PARTE 1: CREAMOS EL MODELO GLM BÁSICO ################
#########################################################################

#Tratamos algunas de las variables
db$VIVO<-1
db$response<-db$Siniestro

#LLamamos al modelo
model_1_glm<-glm(Siniestro~fedad_actuar,offset=NumExp,family=binomial(link="logit"),
data=db)
summary(model_1_glm)

# Creamos variable de predicción de mortalidad según nuestro modelo
predicted<-fitted(model_1_glm,type="response")
db$predicted<-predicted

#Ajuste del Modelo
ROC(db$Siniestro,predicted) #Función para extraer el área y el gráfico de la curva ROC
AE_1(db,predicted,"fedad_actuar",inc=-50) #Comparamos frecuencia real con predicción en
base log
```

```
####################################
## CONTRASTE SOBRE DEPENDENCIA ESP ##
####################################

# Preparamos datos residuos
data<-aggregate(cbind(Siniestro,predicted,NumExp) ~ ID, db, sum)
data$residuo<-data$Siniestro-data$predicted
Merge_datos<-merge(Geo@data, data, by="ID", all.x=TRUE)
Merge_datos[is.na(Merge_datos)]<-0
Geo@data<-Merge_datos
hist(Geo@data$residuo)

# ¿Hay dependencia Espacial?
w<-poly2nb(Geo, queen=FALSE)
w3<-nb2listw(w,zero.policy=TRUE)

plt<-moran.plot(x = Geo@data$residuo, listw = w3, zero.policy = TRUE,
                main="Gráfico I Moran", col="orange", las=1, xlab="x",ylab="y")
tst<-moran.test(x = Geo@data$residuo, listw = w3, zero.policy = TRUE)
tst
# El test confirma la existencia de dependencia espacial en nuestros datos #

#######################################################################
############ PARTE 2: MODELIZACIÓN ESPACIAL CON INLA ###############
#######################################################################

# Sumarización de la información por área
data<-aggregate(cbind(Siniestro,predicted) ~ ID+fedad_actuar , db, sum)
data_2<-aggregate(cbind(Siniestro,predicted) ~ ID, db, sum)
data<- data.frame(y= data$Siniestro, E= data$predicted,
                  ID=as.numeric(data$ID),
                  ID.1=as.numeric(data$ID),
                  age=as.integer(data$fedad_actuar))

# Creamos la variable "graph"
Geo.graph<-inla.read.graph(paste(my.dir, "Geo.graph",sep=""))
image(inla.graph2matrix(Geo.graph),xlab="",ylab="", col="blue")

# Creamos la fórmula del modelo sólo con efecto espacial

formula_1<- y ~ 1 +f(ID,  model="bym", graph=Geo.graph, scale.model=TRUE,
                     hyper=list(prec.unstruct=list(prior="loggamma",param=c(1,0.001)),
                                prec.spatial=list(prior="loggamma",param=c(1,0.001))))

# Creamos la fórmula del modelo con efecto espacial y efecto fijo (edad)
formula_2<- y ~ 1 +age+f(ID,  model="bym", graph=Geo.graph, scale.model=TRUE,
                         hyper=list(prec.unstruct=list(prior="loggamma",param=c(1,0.001)),
                                    prec.spatial=list(prior="loggamma",param=c(1,0.001))))

# Histograma para decidir la distribución family de los modelos
hist(data_2$Siniestro, col="bisque", las=1, xlab="Siniestros",
     ylab="Frecuencia", main="Histograma Siniestros")

# Modelo INLA
model_1_inla <-
inla(formula_1,family="zeroinflatedpoisson1",data=data,control.predictor=list(compute=TR
UE),control.compute=list(dic=TRUE, waic=TRUE,cpo=TRUE))
summary(model_1_inla)

model_2_inla <-
inla(formula_2,family="zeroinflatedpoisson1",E=E,data=data,control.predictor=list(comput
e=TRUE),control.compute=list(dic=TRUE, waic=TRUE, cpo=TRUE))
summary(model_2_inla)


# Algunos resultados modelo sin efectos fijos

 # Media, desviación y cuantiles de los hiperparámetros
 round(model_1_inla$summary.hyperpar, 3)
 # Media, desviación y cuantiles de los efectos fijos (el intercepto)
 round(model_1_inla$summary.fixed, 3)
 # Media, desviación y cuantiles de los efectos aleatorios (espaciales)
 round(head(model_1_inla$summary.random$ID),3)
 #EL ratio esperado en España (exp.b0.mean) y el intervalo de confianza al 95%
(exp.b0.95CI ):
```

```
 # Intercepto: Media posterior global
 exp.b0.mean <- inla.emarginal(exp,model_1_inla$marginals.fixed[[1]])
 exp.b0.mean
 #Intervalo de confianza al 95%
 exp.b0.95CI <-
inla.qmarginal(c(0.025,0.975),inla.tmarginal(exp,model_1_inla$marginals.fixed[[1]]))
 exp.b0.95CI

 #Grafico de las distribuciones marginales de los efectos espaciales de algunos ID
 plot(model_1_inla$marginals.random$ID$index.1, type="l", main="Efecto Aleatorio (u+v)",
las=1, ylim=c(0, 0.3))
 lines(model_1_inla$marginals.random$ID$index.440, type="l", col="orange", lty=2)

 #Número de particiones del área
 N_polygons <- nrow(Geo)

 # Extraemos las predicciones
 zeta.esp.1 <-
data.frame(zeta=unlist(lapply(model_1_inla$marginals.random$ID[1:N_polygons],
function(x)inla.emarginal(exp, x))))
 head(zeta.esp)

 # Transformamos zeta en una variable categórica
 RR.cutoff.1 <- quantile(zeta.esp.1$zeta)
 RR.esp.1 <- cut(zeta.esp.1$zeta, breaks=RR.cutoff.1,include.lowest=TRUE)
 results.1 <- data.frame(ID=data_2$ID,RR.esp.1)
 data.esp.shp.1 <- attr(Geo, "data")
 attr(Geo, "data") <- merge(data.esp.shp.1,results.1, by="ID")

 # Mapa de riesgos relativos a posteriori
 library(lattice)
 trellis.par.set(axis.line=list(col=NA))
 spplot(obj=Geo, zcol="RR.esp.1", main="Mapa de RR a posteriori", )
 hist(zeta.esp.1$zeta, col=heat.colors(9), las=1, xlab="zeta", ylab="Frecuencia",
main="Histograma RR a posteriori")

 ### Mapa Distribución de probabilidad a posteriori específica de cada área.
 a <- 0

 prob.csi <- data.frame(cs=unlist(lapply(model_1_inla$marginals.random$ID[1:N_polygons],
function(x) {1 - inla.pmarginal(a, x)}))) #inla.pmarginal devuelve la función de
distribución
 prob.csi.cutoff <- quantile(prob.csi$cs)
 cat.prob.csi <- cut(unlist(prob.csi),breaks=prob.csi.cutoff, include.lowest=TRUE)
 maps.cat.prob.csi <- data.frame(ID=data_2$ID, cat.prob.csi=cat.prob.csi)

 data.esp.csi <- attr(Geo, "data")
 attr(Geo, "data") <- merge(data.esp.csi, maps.cat.prob.csi, by="ID")
 spplot(obj=Geo, zcol= "cat.prob.csi", col="blue")

 #Para calcular la proporción de variación espacial estructurada frente la
heterogeneidad#
 mat.marg <- matrix(NA, nrow=N_polygons, ncol=1000)
 m <- model_2_inla$marginals.random$ID
 for (i in 1:N_polygons)
 { # Remember that the first Nareas values of the random effects
   # are u+v, while u values are stored in the Nareas+1 to 2*Nareas elements.
   u <- m[[N_polygons+i]]
   mat.marg[i,] <- inla.rmarginal(1000, u) }
 var.u <- apply(mat.marg, 2, var)
 var.v <- inla.rmarginal(1000,inla.tmarginal(function(x) 1/x,
                                             model$marginals.hyper$
                                               "Precision for ID (iid component)"))
 perc.var.u <- mean(var.u/(var.u+var.v))
 perc.var.u


 # Algunos resultados modelo efectos fijos(edad)+efecto espacial
 # Media, desviación y cuantiles de los hiperparámetros
 round(model_2_inla$summary.hyperpar, 3)
 # Media, desviación y cuantiles de los efectos fijos (el intercepto)
 round(model_2_inla$summary.fixed, 3)
 # Media, desviación y cuantiles de los efectos aleatorios (espaciales)
 round(head(model_2_inla$summary.random$ID),3)
```

```
 #EL ratio esperado en España (exp.b0.mean) y el intervalo de confianza al 95%
(exp.b0.95CI ):
 # Intercepto: Media posterior global
 exp.b0.mean.2 <- inla.emarginal(exp,model_2_inla$marginals.fixed[[1]])
 exp.b0.mean.2
 #Intervalo de confianza al 95%
 exp.b0.95CI.2 <-
inla.qmarginal(c(0.025,0.975),inla.tmarginal(exp,model_2_inla$marginals.fixed[[1]]))
 exp.b0.95CI.2
 # Intercepto: Media posterior global
 exp.b1.mean.2 <- inla.emarginal(exp,model_2_inla$marginals.fixed[[2]])
 exp.b1.mean.2
 #Intervalo de confianza al 95%
 exp.b1.95CI.2 <-
inla.qmarginal(c(0.025,0.975),inla.tmarginal(exp,model_2_inla$marginals.fixed[[2]]))
 exp.b1.95CI.2

 #Grafico de las distribuciones marginales de los efectos espaciales de algunos ID
 plot(model_2_inla$marginals.random$ID$index.1, type="l", main="Efecto Aleatorio (u+v)",
ylim=c(0, 0.5))
 lines(model_2_inla$marginals.random$ID$index.188, type="l", col="blue", lty=2)

 #Número de particiones del área
 N_polygons <- nrow(Geo)

 # Extraemos las predicciones
 zeta.esp.2 <-
data.frame(zeta=unlist(lapply(model_2_inla$marginals.random$ID[1:N_polygons],
function(x)inla.emarginal(exp, x))))
 head(zeta.esp.2)

 # Transformamos zeta en una variable categórica
 RR.cutoff.2 <- quantile(zeta.esp.2$zeta)
 RR.esp.2 <- cut(zeta.esp.2$zeta, breaks=RR.cutoff.2,include.lowest=TRUE)
 results.2 <- data.frame(ID=data_2$ID,RR.esp.2)
 data.esp.shp.2 <- attr(Geo, "data")
 attr(Geo, "data") <- merge(data.esp.shp.2,results.2, by="ID")

 # Mapa de riesgos relativos a posteriori
 library(lattice)
 trellis.par.set(axis.line=list(col=NA))
 spplot(obj=Geo, zcol="RR.esp.2", main="Mapa de RR a posteriori")


# Comparamos los dos modelos
 dic_1<-model_1_inla$dic$dic
 dic_2<-model_2_inla$dic$dic
 dic 1
 dic_2

 log.score1 = -mean(log(model_1_inla$cpo$cpo))
 log.score2 = -mean(log(model_2_inla$cpo$cpo))

 log.score1
 log.score2

 hist(model_1_inla$cpo$pit,br=30)
 hist(model_2_inla$cpo$pit,br=30)

 ########################################################
 # PARTE 3: AJUSTE DEL GLM CON VARIABLE ESPACIAL #######
 ########################################################

 # Incorporamos la variable de riesgos relativos espaciales a la BBDD principal
 zeta<-zeta.esp.1$zeta
 zeta2<-cut(zeta.esp.1$zeta,
breaks=quantile(zeta.esp.1$zeta,c(0,0.1,0.4,0.8,0.9,1)),include.lowest=TRUE)
 ID<-1:length(zeta)

 tabla<-cbind(zeta,ID,zeta2)
 head(db)
 db_2<-merge(db, tabla, by="ID", all.x=TRUE)
 db_2[is.na(db_2)]<-0

 #LLamamos al modelo
```

```
 model_2_glm<-
glm(Siniestro~fedad_actuar+zeta2,offset=NumExp,family=binomial(link="probit"),
data=db_2)
 summary(model_2_glm)

 # Creamos variable de predicción de mortalidad según nuestro modelo
 predicted_2<-fitted(model_2_glm,type="response")
 db$predicted_2<-predicted_2

 #Ajuste del Modelo
 ROC(db_2$Siniestro,predicted_2) #Función para extraer el área y el gráfico de la curva
ROC
 AE_1(db_2,predicted_2,"fedad_actuar",inc=-100) #Comparamos frecuencia real con
predicción en base log

 # El ajuste del modelo es mejor que el primero: área de ROC mayor

 ####################################
 ## CONTRASTE SOBRE DEPENDENCIA ESP ##
 ####################################

 # Preparamos datos residuos
 data_fin<-aggregate(cbind(Siniestro,predicted_2,NumExp) ~ ID, db_2, sum)
 data_fin$residuo_2<-data_fin$Siniestro-data_fin$predicted_2
 Merge_datos_fin<-merge(Geo@data, data_fin, by="ID", all.x=TRUE)
 Merge_datos_fin[is.na(Merge_datos_fin)]<-0
 Geo@data<-Merge_datos_fin
 hist(Geo@data$residuo_2)

 # ¿Hay dependencia Espacial?
 w_2<-poly2nb(Geo, queen=FALSE)
 w3_2<-nb2listw(w,zero.policy=TRUE)

 plt_2<-moran.plot(x = Geo@data$residuo_2, listw = w3_2, zero.policy =
TRUE,main="Gráfico I Moran", col="orange", las=1, xlab="x", ylab="y")
 tst_2<-moran.test(x = Geo@data$residuo_2, listw = w3_2, zero.policy = TRUE)
 tst_2
 # El test confirma que hemos eliminado la dependencia espacial del modelo
```