

**PONTIFICIA UNIVERSIDAD
CATÓLICA DEL PERÚ**

Escuela de Posgrado



Modelamiento bayesiano espacial multivariado para datos
de áreas

Tesis para obtener el grado académico de Maestro en Estadística
que presenta:

Miguel Ángel López Esquivel

Asesora:

Zaida Jesús Quiroz Cornejo


Lima, 2023

Informe de Similitud

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Modelamiento bayesiano espacial multivariado para datos de áreas*, del autor Miguel Ángel López Esquivel, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 17%. Así lo consigna el reporte de similitud emitido por el software Turnitin el 21/07/2023.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

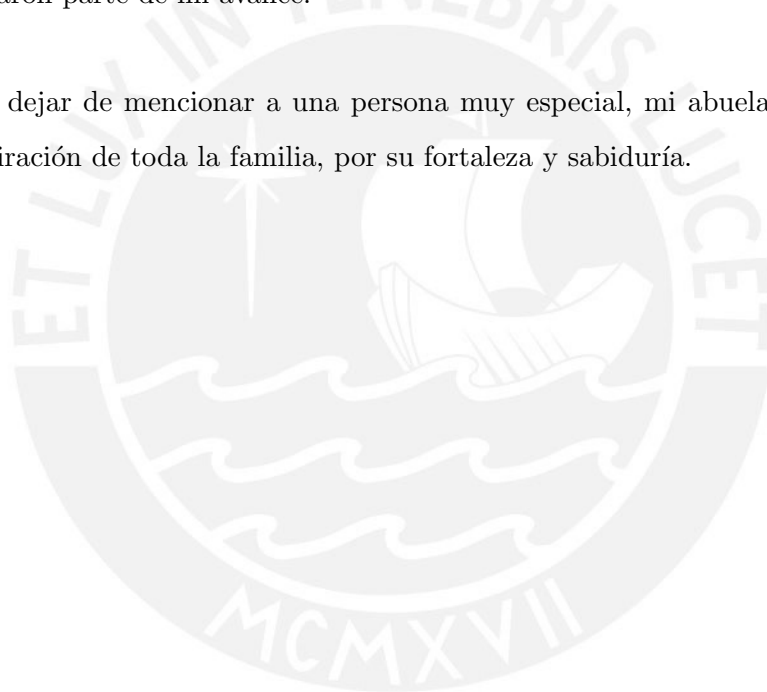
Lima, 21 de julio de 2023

Apellidos y nombres de la asesora: Quiroz Cornejo Zaida Jesús	
DNI: 43704124	Firma: 
ORCID: https://orcid.org/0000-0003-3821-0815	

Dedicatoria

Quiero dedicar este trabajo a mis padres por su abnegada dedicación que tuvieron conmigo en toda mi formación académica, y la confianza que depositaron en mí, sin su apoyo no hubiera llegado a conseguir muchas cosas, sin dejar de mencionar a mis hermanos que también formaron parte de mi avance.

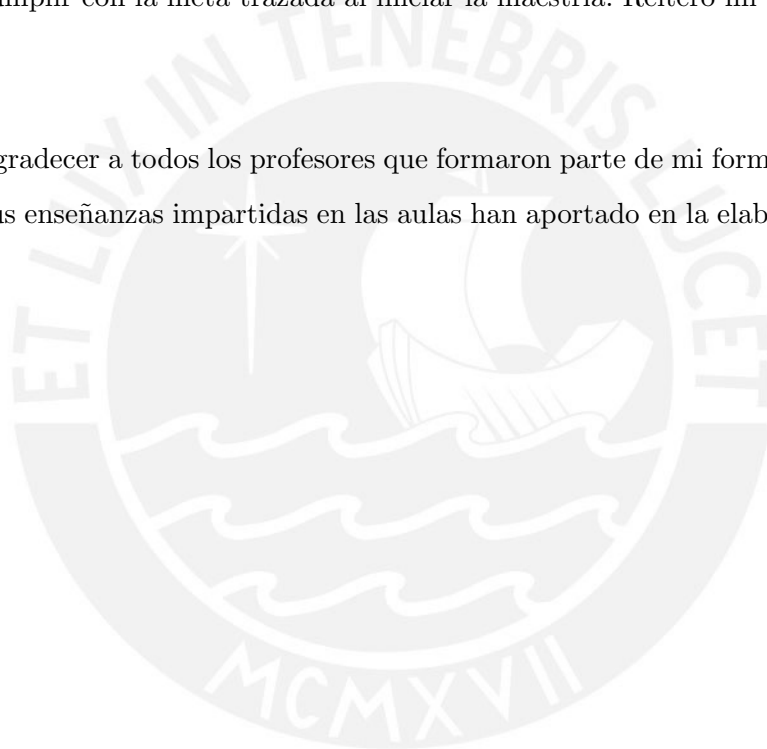
No quiero dejar de mencionar a una persona muy especial, mi abuela Olinda Guevara; ella es la inspiración de toda la familia, por su fortaleza y sabiduría.



Agradecimientos

El agradecimiento sincero a mi asesora Zaida Quiroz quien me brindo todo el apoyo y tiempo necesario, sin ello no hubiera podido concluir con la investigación planteada y de esta manera poder cumplir con la meta trazada al iniciar la maestría. Reitero mi agradecimiento total.

Por último agradecer a todos los profesores que formaron parte de mi formación académica, ya que con sus enseñanzas impartidas en las aulas han aportado en la elaboración de esta investigación.



Resumen

Las infecciones respiratorias son enfermedades que ingresan a nuestro tracto respiratorio afectando la faringe hasta a los pulmones y según la Organización mundial de salud es la causa más común de muertes en el mundo. En particular, en esta tesis se propone estudiar la relación entre la incidencia de infecciones respiratorias agudas (IRA) y la incidencia de neumonía en el Perú. Por un lado estas variables pueden estar correlacionadas, conforme aumenta el número de casos de una enfermedad también aumenta el de la otra. Por otro lado, si nos enfocamos en la incidencia de estas enfermedades a nivel provincial, esperamos que la incidencia de IRA sea similar en provincias vecinas, lo mismo esperamos que ocurra con la incidencia de neumonía. En este contexto, en esta tesis se propone estudiar la distribución espacial entre la incidencia de IRA y neumonía a nivel provincial en el Perú a través de un modelo espacial multivariado, el cual nos permite estudiar la distribución espacial de dos o más variables correlacionadas entre sí. En particular, se propone aplicar un modelo espacial multivariado con efectos aleatorios condicionales autoregresivos. Para conseguir implementar la inferencia bayesiana del modelo jerárquico espacial multivariado de forma eficiente se propone usar el método de integración aproximada anidada de Laplace (INLA).

Palabras-clave: CAR multivariado, INLA, IRA, neumonía.

Abstract

Respiratory infections are diseases that enter our respiratory tract affecting the pharynx to the lungs and according to the World Health Organization it is the most common cause of death in the world. In particular, this thesis proposes to study the relationship between the incidence of acute respiratory infections (ARI) and the incidence of pneumonia in Peru. These variables may be correlated, as the number of cases of one disease increases, the number of the other also increases. On the other hand, if we focus on the incidence of these diseases at the provincial level, we expect that the incidence of ARI to be similar in neighboring provinces, and the same is expected to occur with the incidence of pneumonia. In this context, this thesis proposes to study the spatial distribution between the incidence of ARI and pneumonia at the provincial level in Peru through a multivariate spatial model, which allows us to study the spatial distribution of two or more interrelated variables. In particular, it is proposed to apply a multivariate spatial model with conditional autoregressive random effects. In order to implement fast Bayesian inference of the multivariate spatial hierarchical model, it is proposed to use Integrated nested Laplace approximation (INLA) method.

Keywords: INLA, multivariate CAR, pneumonia, RTI.

Índice general

1. Introducción	1
1.1. Consideraciones preliminares	1
1.2. Objetivos	2
1.3. Organización del trabajo	3
2. Conceptos	4
2.1. Modelos lineales generalizados	4
2.1.1. Distribución de Poisson	4
2.1.2. Modelo lineal generalizado de Poisson	5
2.2. Datos de áreas univariado	6
2.2.1. Modelo condicional autorregresivo (CAR)	7
2.3. Inferencia bayesiana	9
2.3.1. Inferencia bayesiana para modelo CAR-Poisson	9
2.3.2. Aproximación de Laplace Anidada Integrada (INLA)	10
2.4. Criterios de selección de modelos	13
2.4.1. Criterio de información de devianza (DIC)	13
2.4.2. Criterio de información de Watanabe-Akaike (WAIC)	14
2.4.3. Logarithm pseudo marginal likelihood (LPML)	14
3. Modelos bayesianos espaciales multivariados	16
3.1. Modelo 1: MCAR propio sin offset	16
3.2. Modelo 2: MCAR propio con offset	18
3.3. Inferencia bayesiana usando INLA	18
4. Estudio de Simulación	20
4.1. Simulación con interceptos	21
4.2. Simulación con interceptos y covariables	24
4.3. Escenarios de simulaciones bajo diferentes ρ	27

5. Aplicación	29
6. Conclusiones	35
Bibliografía	36



Capítulo 1

Introducción

1.1. Consideraciones preliminares

El avance computacional actual ha hecho posible recolectar datos que están a nuestra disposición. Por ejemplo realizar un seguimiento de los efectos que causan las enfermedades permiten hacerle frente, y así tener un panorama más general. En la actualidad el análisis estadístico por intermedio de un entorno espacial facilita el estudio de situaciones críticas que afronta el sector salud de algunos países, en particular la estadística espacial permite conocer patrones de estos acontecimientos geográficos y nos permiten proponer soluciones viables que mejoren las políticas de estado. En particular, las infecciones respiratorias agudas, son enfermedades que ingresan a nuestro tracto respiratorio afectando la faringe hasta los pulmones y según la organización mundial de salud (OMS) es la causa más común de muertes en el mundo. En este contexto, la tesis propone estudiar las infecciones respiratorias agudas (IRA) sin neumonía, como son la tuberculosis, y la neumonía, enfermedades que presentan una prevalencia alta en el Perú (Valero et al., 2009).

Es natural, pensar que la prevalencia de cada una de estas enfermedades es similar en provincias vecinas, por ello se asume que puede haber evidencia de autocorrelación espacial. Esta información es crucial porque permitiría plantear soluciones focalizadas para determinadas provincias. Por otro lado, aunque se puede estudiar la distribución espacial de cada una de estas enfermedades en las provincias del país de forma separada, es importante también analizar ambas enfermedades de forma conjunta, pues es natural que ambas variables estén correlacionadas de forma directa, por ser enfermedades que están relacionadas al sistema respiratorio. El análisis de este tipo de correlación entre las enfermedades se realiza recurriendo a modelos multivariados y a la autocorrelación espacial a través de modelos espaciales para datos de áreas. En este contexto, el presente trabajo de investigación se circunscribe en plantear

modelos espaciales multivariados para estudiar estas dos enfermedades de forma conjunta. Estos modelos permitirán realizar estimaciones, predicciones y mapas de la distribución de estas enfermedades en las provincias del país. En particular, se propone aplicar un modelo espacial condicional autoregresivo (CAR) multivariado propuesto por (Palmí-Perales et al., 2021), el cual es una extensión de los modelos CAR univariados al análisis multivariado.

En general, se podrían estimar los parámetros de estos modelos usando inferencia clásica o inferencia bayesiana. Los métodos bayesianos consisten en actualizar la información obtenida de los datos mediante el conocido teorema de Bayes, incorporando la información del investigador sobre los parámetros de interés. Dadas las características de las variables respuesta, se puede asumir que cada variable respuesta sigue una distribución de Poisson, por lo tanto el modelo multivariado espacial pertenece a una familia de modelos gaussianos latentes, por lo cual es razonable usar inferencia bayesiana para la estimación de los parámetros. En particular, se podría usar métodos de Monte Carlo vía cadenas de Markov (MCMC), sin embargo, es conocido que este método es muy lento debido a la convergencia de las cadenas de Markov. Motivo por el cual en esta tesis, dada la complejidad de los modelos espaciales multivariados, se propone usar la integración anidada de la aproximación de Laplace (INLA) propuesta por (Rue et al., 2009), método de inferencia preciso que minimiza los tiempos de cómputo.

1.2. Objetivos

El objetivo general de la tesis es de estimar y aplicar a conjuntos de datos reales modelos bayesianos espaciales multivariados para datos de área. De manera específica:

- Revisar la literatura acerca de los diferentes propuestas de modelos bayesianos espaciales multivariados.
- Estudiar la estimación a través del INLA del modelo bayesiano espacial multivariado.
- Realizar estudios de simulación acerca de los modelos bayesianos espaciales multivariados sobre diferentes escenarios.
- Aplicar el modelo a conjunto de datos reales en el área de salud.

1.3. Organización del trabajo

En el capítulo 2 se revisarán los conceptos básicos de modelos lineales generalizados de Poisson; datos de áreas, el modelo condicional autorregresivo (CAR); inferencia bayesiana e INLA. En el capítulo 3 se realizará una descripción de los diferentes modelos espaciales multivariados de datos de área y la inferencia bayesiana a través del INLA para los modelos. En el capítulo 4 se mostrará la simulación de los modelos ajustados. En el capítulo 5, se mostrará la aplicación del modelos a un conjunto de datos reales, asociados a las enfermedades de infecciones respiratorias agudas sin neumonía y con neumonía en el Perú. Finalmente en el capítulo 6 se discutirá las conclusiones obtenidas de los métodos aplicados.



Capítulo 2

Conceptos

2.1. Modelos lineales generalizados

Los modelos lineales generalizados (MLG) fueron introducidos por (Nelder y Wedderburn, 1972), donde se plantea extender los modelos lineales mediante una variable respuesta que sigue una distribución de la familia exponencial, como la normal, binomial, Poisson y gamma. Estos modelos son una alternativa a la transformación de la variable respuesta Y_i , que se suele utilizar ante la falta de linealidad y de homocedasticidad (Dobson y Barnett, 2018).

En un modelo lineal generalizado asumimos que la variable respuesta Y_i pertenece a una familia exponencial; y la esperanza condicional de Y_i , μ_i , está asociada a un predictor lineal η_i , que depende de los coeficientes de regresión β y un vector de covariables X_i , a través de una *función de enlace* g , es decir,

$$g(\mu_i) = \eta_i = X_i^\top \beta.$$

Para todas las distribuciones consideradas de Y_i existe al menos una función de enlace (Lindsey, 2000). Los datos que trabajaremos son de conteo y enfocaremos más nuestro estudio de los modelos lineales generalizados (MLG) en el caso particular de la distribución de Poisson.

2.1.1. Distribución de Poisson

La distribución de Poisson fue propuesta a mediados del siglo XIX por Siméon-Denis Poisson (Poisson, 1837), en su trabajo sobre la probabilidad de sentencias en materia penal y civil. El trabajo se enfocó en el número de condenas injustas en un país determinado, centrándose en ciertas variables aleatorias que cuentan, entre otras cosas, el número de sucesos discretos conocidos también como llegadas que tienen lugar durante un intervalo unitario

establecido.

La distribución de Poisson es una distribución de probabilidad discreta que expresa probabilidades de ocurrencias en un período unitario; estos eventos ocurren con un tasa media conocida e independientemente del periodo transcurrido, puede ser aplicado a sistemas con un número grande de posibles resultados, utilizado en intervalos unitarios especificados como el tiempo, la distancia, área o volumen (Hu, 2008). Por ejemplo, llegada de clientes a un cajero, número de salmones en un determinadas zonas de un río.

Si el número esperado de ocurrencias en este intervalo unitario es λ , entonces la probabilidad de que haya exactamente y ocurrencias es igual a:

$$f_{Y_i}(y) = \frac{e^{-\lambda} \cdot \lambda^y}{y!}; y = 0, 1, 2, 3, \dots,$$

donde y es un entero no negativo y denota el número de ocurrencias de un evento; λ es un número real positivo, igual al número esperado de ocurrencias que suceden durante el intervalo unitario fijado. A esta distribución lo denotaremos como *Poisson*(λ).

2.1.2. Modelo lineal generalizado de Poisson

Los modelos lineales asumen supuestos de independencia, linealidad y distribución normal. Estos supuestos no siempre se cumplen por ello los modelos lineales generalizados extienden este modelo lineal cuando existe asociación no lineal entre la variable respuesta y las covariables, o cuando la distribución de los datos no es normal y pertenece a la familia exponencial. En particular, la distribución Poisson pertenece a la familia exponencial y este modelo se denotará MLG Poisson. Se utiliza para modelar datos que resultan de un conteo, es decir que toman valores enteros no negativos (Dobson y Barnett, 2018). La variable respuesta Y_i para una observación i sigue la distribución de Poisson donde su parámetro λ_i es el promedio de ocurrencias sobre un intervalo unitario, y será denotada por $E(Y_i) = \lambda_i$; tal que $Y_i \sim Poisson(\lambda_i)$. Para enlazar esta variable con un vector de covariables X_i y un predictor lineal $\eta_i = X_i^\top \beta$ resulta conveniente introducir una transformación logarítmica, es decir,

$$\eta_i = \log(\lambda_i) = X_i^\top \beta,$$

donde β es un vector de coeficientes de regresión.

2.2. Datos de áreas univariado

Los datos de área se definen en una malla (lattice) de un dominio $D = \{s_i : i = 1, \dots, n\} \subseteq \mathbb{R}^2$ que contiene una colección contable de áreas, donde la variable respuesta se observa en cada área y es definida por:

$$\{Y(s_i) : i = 1, \dots, n\}.$$

Cada área tendrá un número de vecinos que sirve para establecer una estructura espacial que permita plantear modelos espaciales (Perales, 2020).

El vecindario puede ser considerado como: áreas con una frontera en común, o tomando la distancia entre los centroides de cada área, o también como un número de vecinos más cercanos. A partir de ello se define una matriz de vecindad \mathbf{W} , siendo esta una matriz de proximidad, estructura o adyacencia. Específicamente, esta matriz brinda un mecanismo para introducir la estructura de autocorrelación espacial en el modelo espacial. Esta matriz presenta entradas w_{ij} referentes a las áreas i y j ; donde $w_{ij} = 1$ en el caso de que i y j sean vecinos y 0 en caso contrario. Los elementos de \mathbf{W} pueden ser considerados como ponderaciones o pesos, cuanto mayor sea el peso significaría que i y j son áreas más próximas. Esta relación entre vecinos puede ser representada a través de un grafo. Por ejemplo, la Figura 2.1 muestra el grafo asociado a las regiones vecinas en el Perú, donde dos regiones son vecinas están unidas por una arista.

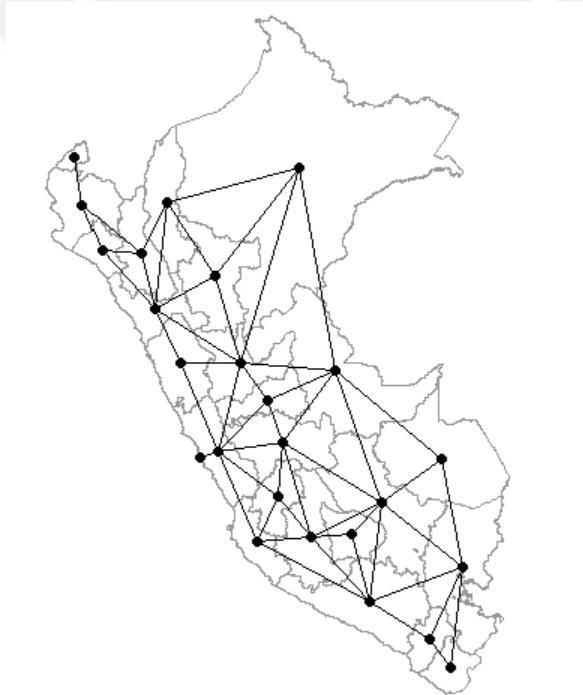


Figura 2.1: Grafo de vecinos a nivel regional en el Perú.

2.2.1. Modelo condicional autorregresivo (CAR)

El modelo autorregresivo condicional (CAR) fue introducido por (Besag, 1974), siendo este una extensión de los modelos autoregresivos clásicos que son ampliamente utilizados en análisis de series de tiempo y se basaban en la suposición de que el valor actual de la variable es una combinación lineal de las variables en los tiempos pasados. En un modelo CAR, se asume que cada variable en un área dependerá de otras áreas vecinas. Los modelos CAR se han implementado para modelar la dependencia espacial para datos de área, y es comúnmente descrita por un efecto espacial aleatorio u_i en modelos jerárquicos. Los modelos CAR son empleados como a prioris del efecto aleatorio en un enfoque bayesiano (Song, 2004).

El modelo CAR es construido a partir de las distribuciones condicionales completas de los efectos aleatorios espaciales u_i de la siguiente forma:

$$u_i | u_j; j \neq i \sim N(\sum_j b_{ij} u_j, \tau_i^2), \quad i, j = 1, \dots, n, \quad (2.1)$$

donde los b_{ij} son constantes conocidas y τ_i^2 es un parámetro de escala. A través del lema de Brook (Brook, 1964), se puede obtener una fdp conjunta a partir de las distribuciones condicionales completas. De esta forma, según (Besag, 1974), las distribuciones condicionales completas definidas en la ecuación 2.1 generan la siguiente función de densidad (fdp) conjunta:

$$p(\mathbf{u}) = p(u_1, \dots, u_n) \propto \exp \left\{ -\frac{1}{2} \mathbf{u}^\top \mathbf{D}^{-1} (\mathbf{I} - \mathbf{B}) \mathbf{u} \right\}, \quad (2.2)$$

donde

$$\mathbf{B} = \begin{pmatrix} b_{11} & b_{12} & b_{13} & \dots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \dots & b_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ b_{n1} & b_{n2} & b_{n3} & \dots & b_{nn} \end{pmatrix}$$

y

$$\mathbf{D} = \begin{pmatrix} \tau_1^2 & 0 & 0 & \dots & 0 \\ 0 & \tau_2^2 & 0 & \dots & 0 \\ 0 & 0 & \tau_3^2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \tau_n^2 \end{pmatrix}.$$

Para que el núcleo de esta fdp sea semejante al núcleo de una fdp normal, se requiere ciertas condiciones. A continuación se detallan las mismas según (Banerjee et al., 2003). Una de

estas condiciones es que $\mathbf{D}^{-1}(\mathbf{I} - \mathbf{B})$ sea simétrica. Así la estructura espacial del vecindario se introduce en la matriz \mathbf{B} , y es una función de la matriz de vecindad \mathbf{W} . Específicamente, se considera $b_{ij} = \frac{w_{ij}}{w_{i+}}$ y $\tau_i^2 = \frac{\tau^2}{w_{i+}}$ donde w_{ij} son las entradas de la matriz de vecindad \mathbf{W} y $w_{i+} = \sum_j w_{ij}$ es el número de vecinos de i . Por lo tanto:

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \quad \forall i, j,$$

garantizando la simetría de $\mathbf{D}^{-1}(\mathbf{I} - \mathbf{B})$. Luego se define la matriz de precisión \mathbf{Q} (inversa de la matriz de covarianza) la cual se puede expresar de la siguiente manera:

$$\mathbf{Q} = \mathbf{D}^{-1}(\mathbf{I} - \mathbf{B}) = \frac{1}{\tau^2}(\mathbf{D}_{\mathbf{W}} - \mathbf{W}),$$

donde $\mathbf{D}_{\mathbf{W}} = \text{diag}(w_{i+})$ y \mathbf{Q} es una matriz simétrica. Así las fdp completas ahora son definidas por $p(u_i | u_j, j \neq i) = N(\sum_j w_{ij}u_j/w_{i+}, \tau^2/w_{i+})$. Luego la fdp conjunta de \mathbf{u} se puede expresar de la forma siguiente:

$$p(\mathbf{u}) = p(u_1, \dots, u_n) \propto \exp \left\{ -\frac{1}{2\tau^2} \mathbf{u}^\top (\mathbf{D}_{\mathbf{W}} - \mathbf{W}) \mathbf{u} \right\}. \quad (2.3)$$

Como la matriz de precisión \mathbf{Q} es singular, no existe su inversa, dado lugar a la otra condición para que la ecuación 2.2 represente el núcleo de una fdp conjunta normal. Para asegurar la no singularidad se reemplaza \mathbf{W} por $\rho\mathbf{W}$, donde la medida de asociación espacial ρ estará limitada en el intervalo $[\lambda_1^{-1}; \lambda_n^{-1}]$; y $\lambda_1 \leq \dots \leq \lambda_n$ son los autovalores ordenados de $\mathbf{D}_{\mathbf{W}}^{-1/2} \mathbf{W} \mathbf{D}_{\mathbf{W}}^{-1/2}$. De esta manera $(\mathbf{D}_{\mathbf{W}} - \rho\mathbf{W})$ será definida positiva y existe su inversa (Song, 2004). Luego $(\mathbf{D}_{\mathbf{W}} - \mathbf{W})$ en la ecuación (2.3) será reescrita como $(\mathbf{D}_{\mathbf{W}} - \rho\mathbf{W})$, además si $|\rho| < 1$ también se asegura que $(\mathbf{D}_{\mathbf{W}} - \rho\mathbf{W})$ sea definida positiva, tomando todo esto en cuenta la fdp conjunta de \mathbf{u} se podrá definir como:

$$p(\mathbf{u}) = p(u_1, \dots, u_n) \propto \exp \left\{ -\frac{1}{2\tau^2} \mathbf{u}^\top (\mathbf{D}_{\mathbf{W}} - \rho\mathbf{W}) \mathbf{u} \right\}.$$

Esto es equivalente a decir que:

$$\mathbf{u} = (u_1, \dots, u_n)^\top \sim N\left(0, \left[\frac{1}{\tau^2} (\mathbf{D}_{\mathbf{W}} - \rho\mathbf{W}) \right]^{-1}\right), \quad \text{para } |\rho| < 1. \quad (2.4)$$

lo cual puede ser reescrito como:

$$\mathbf{u} = (u_1, \dots, u_n)^\top \sim N(0, \mathbf{Q}^{-1}). \quad (2.5)$$

Si $\rho = 0$ entonces u_i son independientes, mientras que si ρ tiende a 1 significaría mayor dependencia espacial entre áreas vecinas.

2.3. Inferencia bayesiana

En la teoría de probabilidades condicionales existe el llamado teorema de Bayes el cual fue desarrollado por el matemático británico Thomas Bayes y publicado de modo póstumo por Richard Price (Bayes, 1763). Este teorema nos sugiere como modificar el parámetro a evaluar con los datos actuales, utilizando toda la información disponible sobre el parámetro en cuestión (Pereira da Silva, 2016).

El teorema de Bayes nos dice que la probabilidad condicional del evento A_i de una partición del espacio muestral Ω , dado un evento B es:

$$P(A_i | B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B | A_i)P(A_i)}{P(B)} = \frac{P(B | A_i)P(A_i)}{\sum_{i=1}^n P(B | A_i)P(A_i)},$$

donde $\Omega = A_1 \cup \dots \cup A_n$ tal que $A_i \cap A_j = \emptyset; i \neq j$. Una vez observado un conjunto de datos, nuestro conocimiento sobre un parámetro al cual lo denominaremos por θ , puede con el teorema de Bayes ser actualizado mediante la distribución a priori de θ , resultando la llamada distribución a posteriori de θ , tal que:

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

donde $p(y) = \int p(\theta) \cdot p(y | \theta) d\theta$. Como $p(y)$ no depende de θ podemos también escribir

$$p(\theta | y) \propto p(\theta) \cdot p(y | \theta). \quad (2.6)$$

La distribución a priori, $p(\theta)$ representa el conocimiento previo que tenemos sobre un parámetro θ antes de realizar un experimento o tomar una muestra, mientras que la verosimilitud $p(y | \theta)$ es la distribución de los datos, si conociéramos θ y por último la distribución a posteriori $p(\theta | y)$, es la distribución de θ que resume la información previa sobre θ y la obtenida por los datos (Box y Tiao, 2011).

2.3.1. Inferencia bayesiana para modelo CAR-Poisson

El modelo espacial CAR-Poisson asume que la distribución condicional de Y_i es dada por:

$$Y_i | \beta, u, \rho, \tau^2 \sim \text{Poisson}(\mu_i), \quad i = 1, \dots, n;$$

donde la media μ_i es definida por:

$$\log(\mu_i) = X_i^\top \beta + u_i,$$

donde:

Y_i : es la variable respuesta en el área i

X_i : es un vector de covariables

β : es un vector de coeficientes de regresión

Para los efectos aleatorios espaciales u_i se asume el modelo CAR ecuaciones 2.4 y 2.5, siendo

$$\mathbf{u} \mid \rho, \tau^2 \sim N(0, \mathbf{Q}^{-1}), \quad (2.7)$$

con matriz de precisión $\mathbf{Q} = \frac{1}{\tau^2}(\mathbf{D}\mathbf{W} - \rho\mathbf{W})$. Se definen también las distribuciones a priori para los hiperparámetros $\boldsymbol{\theta} = (\beta, \tau^2, \rho)$, distribución normal para β , distribución gamma inversa para τ^2 y distribución uniforme para ρ :

$$\beta \sim N(\mu_\beta, \Sigma_\beta); \quad \tau^2 \sim GI(a, b); \quad \rho \sim U(0, 1). \quad (2.8)$$

Luego la función de verosimilitud es dada por,

$$L(\beta, \mathbf{u}, \rho, \tau^2) = L(\mathbf{u}, \boldsymbol{\theta}) = f(y \mid \mathbf{u}, \boldsymbol{\theta}) = \prod_{i=1}^n f_{Y_i}(y_i \mid \beta, \mathbf{u}, \rho, \tau^2). \quad (2.9)$$

La fdp conjunta a posteriori por intermedio de las ecuaciones 2.7, 2.8 y 2.9 estará dada por:

$$\begin{aligned} \pi(\beta, \mathbf{u}, \rho, \tau^2 \mid y) &\propto f(y \mid \mathbf{u}, \boldsymbol{\theta}) \cdot \pi(\mathbf{u}, \boldsymbol{\theta}) \\ \pi(\beta, \mathbf{u}, \rho, \tau^2 \mid y) &\propto f(y \mid \mathbf{u}, \boldsymbol{\theta}) \cdot \pi(\mathbf{u} \mid \boldsymbol{\theta}) \cdot \pi(\boldsymbol{\theta}) \\ \pi(\beta, \mathbf{u}, \rho, \tau^2 \mid y) &\propto f(y \mid \mathbf{u}, \boldsymbol{\theta}) \cdot \pi(\mathbf{u} \mid \boldsymbol{\theta}) \cdot \pi(\beta) \cdot \pi(\tau^2) \cdot \pi(\rho) \end{aligned}$$

Como la fdp conjunta a posteriori por lo general no tiene una forma conocida, se emplearán métodos computacionales para estimar los parámetros a posteriori, por ejemplo MCMC, (Gilks et al., 1995).

2.3.2. Aproximación de Laplace Anidada Integrada (INLA)

El método integrated nested Laplace approximation (INLA) propuesto por (Rue et al., 2009), es un algoritmo determinista que permite aproximar la distribución a posteriori con muy buena precisión y un tiempo de computo menor a otros métodos como los de la simulación de cadenas de Markov de Montecarlo (MCMC). Esta implementación está disponible en el paquete R-INLA, y puede desarrollarse para una gran familia de modelos

gaussianos latentes (Blangiardo et al., 2013) que definiremos con detalle más adelante. Sea $\mathbf{y} = (Y_1, Y_2, \dots, Y_n)^\top$ un vector de variables de respuesta y suponga que la distribución de Y_i presenta un parámetro μ_i , por lo que el predictor lineal se puede definir de la siguiente manera:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{m=1}^M \beta_m z_{mi} + \sum_{j=1}^J f_j(u_{ji}); \quad i : 1, \dots, n;$$

donde $g(\cdot)$ es una función de enlace, β_0 es un intercepto del modelo, $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \dots, \beta_M)$ son coeficientes de regresión que miden el efecto de las covariables $z = (z_1, z_2, z_3, \dots, z_M)$ sobre el parámetro μ_i de la variable respuesta Y_i , además $\mathbf{f} = \{f_1(\cdot), \dots, f_J(\cdot)\}$ es una colección de funciones definidas en términos de un subconjunto de covariables u_{ji} (covariable j para la observación i), las funciones $f_j(\cdot)$ pueden asumir diferentes formas al igual que los efectos aleatorios espaciales. Esta formulación puede acomodar una amplia gama de modelos, desde estándares como regresión lineal hasta modelos espaciales, gracias a las diferentes formas que pueden tomar las funciones desconocidas en \mathbf{f} . Los modelos gaussianos latentes son modelos bayesianos jerárquicos que tienen tres niveles: i) En el primer nivel definen la distribución del vector observado \mathbf{y} ; ii) en el segundo nivel asignan una a priori gaussiana a ciertos parámetros o efectos aleatorios, como por ejemplo de β_o ; $\{\beta_m\}$; $\{f_j(\cdot)\}$, todos estos términos definen un vector latente $\mathbf{x} = (\beta_o, \boldsymbol{\beta}, \mathbf{f})$; y iii) en el tercer nivel se asigna una distribución al resto de parámetros llamados hiperparámetros $\boldsymbol{\theta}$.

La fdp a posteriori de $\mathbf{x}, \boldsymbol{\theta}$ es dada por:

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) = \frac{\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \cdot \pi(\mathbf{x}, \boldsymbol{\theta})}{\pi(\mathbf{y})} \propto \pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \cdot \pi(\mathbf{x}, \boldsymbol{\theta}), \quad (2.10)$$

donde $\pi(\mathbf{y})$ representa la fdp marginal de \mathbf{y} . Mediante el INLA se obtiene una aproximación precisa de esta distribución, la cual mediante métodos convencionales sería complicado obtener (Outzen Berild et al., 2021).

En particular, $\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta})$ representa la verosimilitud, y asumiendo observaciones condicionalmente independientes dado \mathbf{x} y $\boldsymbol{\theta}$ se obtiene por:

$$\pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) = \prod_{i=1}^n \pi(\mathbf{y}_i | \mathbf{x}_i, \boldsymbol{\theta}). \quad (2.11)$$

Asumiendo una a priori normal multivariante sobre el vector latente \mathbf{x} , con media cero y una matriz de precisión $\mathbf{Q}(\boldsymbol{\theta})$, es decir $\mathbf{x} \sim N(0, \mathbf{Q}^{-1}(\boldsymbol{\theta}))$, su función de densidad está dada por:

$$\pi(\mathbf{x} | \boldsymbol{\theta}) = (2\pi)^{-n/2} |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta})\mathbf{x}\right], \quad (2.12)$$

donde $|\mathbf{Q}(\boldsymbol{\theta})|$ es el determinante de $\mathbf{Q}(\boldsymbol{\theta})$. En particular, cuando la matriz de precisión es dispersa se tiene que \mathbf{x} es un campo aleatorio markoviano gaussiano (Rue y Held, 2005), y esta propiedad permite que el INLA sea eficiente en términos computacionales. Luego la distribución a posteriori de \mathbf{x} y $\boldsymbol{\theta}$ se puede expresar por:

$$\begin{aligned} \pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) &\propto \pi(\boldsymbol{\theta}) \cdot \pi(\mathbf{x} | \boldsymbol{\theta}) \cdot \pi(\mathbf{y} | \mathbf{x}, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) \cdot (2\pi)^{-n/2} |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta})\mathbf{x}\right] \cdot \prod_{i=1}^n \pi(y_i | \mathbf{x}_i, \boldsymbol{\theta}) \\ &\propto \pi(\boldsymbol{\theta}) \cdot |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left[-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta})\mathbf{x}\right] \cdot \prod_{i=1}^n \exp(\log(\pi(y_i | \mathbf{x}_i, \boldsymbol{\theta}))). \end{aligned}$$

La ecuación puede ser reescrita:

$$\pi(\mathbf{x}, \boldsymbol{\theta} | \mathbf{y}) \propto \pi(\boldsymbol{\theta}) |\mathbf{Q}(\boldsymbol{\theta})|^{1/2} \exp\left\{-\frac{1}{2}\mathbf{x}^\top \mathbf{Q}(\boldsymbol{\theta})\mathbf{x} + \sum_{i=1}^n \log(\pi(y_i | x_i, \boldsymbol{\theta}))\right\}. \quad (2.13)$$

INLA genera aproximaciones numéricas de las marginales a posteriori para el vector latente \mathbf{x} y los hiperparámetros $\boldsymbol{\theta}$. Las marginales a posteriori para las componentes del vector latente se pueden expresar:

$$\pi(\mathbf{x}_i | \mathbf{y}) = \int \pi(\mathbf{x}_i | \boldsymbol{\theta}, \mathbf{y}) \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}. \quad (2.14)$$

Las marginales a posteriori para las componentes de los hiperparámetros son definidas como:

$$\pi(\boldsymbol{\theta}_k | \mathbf{y}) = \int \pi(\boldsymbol{\theta} | \mathbf{y}) d\boldsymbol{\theta}_{-k}, \quad (2.15)$$

donde $\boldsymbol{\theta}_{-k}$ denota todos los hiperparámetros excepto el componente k .

Para ello INLA aproxima $\pi(\boldsymbol{\theta} | \mathbf{y})$ por $\tilde{\pi}(\boldsymbol{\theta} | \mathbf{y})$. Específicamente, como

$$\begin{aligned} \pi(\boldsymbol{\theta} | \mathbf{y}) &= \frac{\pi(x, \boldsymbol{\theta} | \mathbf{y})}{\pi(x | \boldsymbol{\theta}, \mathbf{y})} \\ &= \frac{\pi(y | x, \boldsymbol{\theta}) \pi(x, \boldsymbol{\theta})}{\pi(y | x, \boldsymbol{\theta}) \pi(x | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})} \cdot \frac{1}{\pi(x | \boldsymbol{\theta}, \mathbf{y})} \\ &= \frac{\pi(y)}{\pi(y | x, \boldsymbol{\theta}) \pi(x | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})} \cdot \frac{1}{\pi(x | \boldsymbol{\theta}, \mathbf{y})} \\ &\propto \frac{\pi(y)}{\pi(y | x, \boldsymbol{\theta}) \pi(x | \boldsymbol{\theta}) \pi(\boldsymbol{\theta})} \cdot \frac{1}{\pi(x | \boldsymbol{\theta}, \mathbf{y})}, \end{aligned}$$

es aproximada por:

$$\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y}) \propto \frac{\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta})\pi(\mathbf{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\tilde{\pi}_G(\mathbf{x} \mid \boldsymbol{\theta}, \mathbf{y})} \Big|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\theta})}, \quad (2.16)$$

donde $\tilde{\pi}_G$ denota la aproximación gaussiana para la condicional completa de \mathbf{x} ; además $\mathbf{x}^*(\boldsymbol{\theta})$ es la moda a posteriori de $(\mathbf{x} \mid \boldsymbol{\theta}, y)$ es decir $\text{Arg máx}_{\mathbf{x}} \pi_G(\mathbf{x} \mid \boldsymbol{\theta}, y)$.

Las marginales a posteriori para los hiperparámetros $\tilde{\pi}(\boldsymbol{\theta}_i \mid \mathbf{y})$ se pueden derivar de $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ a través de la integración numérica. De forma similar se aproxima $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$ usando otra aproximación de Laplace. Para más detalles sobre el INLA ver (Rue et al., 2009).

2.4. Criterios de selección de modelos

Cuando se realiza una investigación donde intervienen modelos estadísticos por lo general suelen ajustarse varios modelos de acuerdo a las características de los datos, desde el más simple hasta uno más complejo. Sin embargo por el principio de parsimonia se requiere obtener modelos los mas simples posibles con el objetivo que el modelo llegue a explicar adecuadamente la variable respuesta planteada. Para comparaciones de los modelos se emplearán los siguientes criterios:

- Criterio de información de devianza (DIC)
- Criterio de información de Watanabe-Akaike (WAIC)
- logarithm pseudo marginal likelihood (LPML).

2.4.1. Criterio de información de devianza (DIC)

Este criterio es similar a un criterio empleado en inferencia clásica conocido con el nombre de Akaike information criterion (AIC), solo que en esta ocasión el DIC es aplicado en inferencia bayesiana, y es igual a la devianza estimada $D(\hat{\theta})$, más el doble del número efectivo de parámetros \hat{p}_D (Spiegelhalter et al., 2002). Para un conjunto de datos de tamaño n , se simulan los valores $\theta^1, \theta^2, \dots, \theta^S$ de la distribución a posteriori, entonces el DIC se estima de la siguiente forma:

$$\widehat{DIC} = D(\hat{\theta}) + 2\hat{p}_D,$$

donde:

- $\hat{\theta} = \frac{1}{S} \sum_{j=1}^S \theta^{(j)}$; (Media a posteriori de θ)
- $D(\hat{\theta}) = -2 \sum_{i=1}^n \log f(y_i \mid \hat{\theta})$; (Devianza estimada de θ)

- $\hat{p}_D = \bar{D} - D(\hat{\theta})$; (Número efectivo de parámetros)
- $\bar{D} = E[D(\theta) | Y] = \frac{1}{S} \sum_{j=1}^S D(\theta^{(j)}) = \frac{1}{S} \sum_{j=1}^S [-2 \sum_{i=1}^n \log f(y_i | \theta^{(j)})]$
(Media a posteriori de la devianza).

El DIC solo puede ser utilizado para comparar modelos que presentan la misma verosimilitud, y da resultados equivocados cuando la distribución a posteriori no es distribuida normalmente (Gelman et al., 2014).

2.4.2. Criterio de información de Watanabe-Akaike (WAIC)

El WAIC fue desarrollada por el matemático Sumio Watanabe, y se basa en la densidad predictiva a posteriori, y es donde radica su principal ventaja frente a otras medidas. Es particularmente útil para los modelos jerárquicos. El WAIC es una alternativa al DIC y desde el punto de vista teórico es un mejor indicador que el DIC (Watanabe y Opper, 2010). Para un conjunto de datos de tamaño n , se simulan los valores $\theta^1, \theta^2, \dots, \theta^S$ de la distribución a posteriori obteniéndose de esta manera la estimación de WAIC:

$$\widehat{WAIC} = -2 \sum_{i=1}^n \widehat{\log\{m_i\}} + 2\widehat{p_W},$$

donde:

- $\widehat{m}_i = \frac{1}{S} \sum_{j=1}^S \pi(y_i | x^{(j)}, \theta^{(j)})$; (Media a posteriori de $f(y_i | \theta)$)
- $\widehat{p_W} = \sum_{i=1}^n v_i$; (Tamaño efectivo del modelo)
- $v_i = V_{j=1}^S \log\{\pi(y_i | x^{(j)}, \theta^{(j)})\}$; (Varianza a posteriori de $\log[f(y_i | \theta)]$).

2.4.3. Logarithm pseudo marginal likelihood (LPML)

Este criterio parte de las coordenadas predictivas condicionales (CPO) y es una alternativa adicional a los criterios antes mencionados. Sea $y_{(-i)} = (y_1, \dots, y_{i-1}, y_{i+1}, \dots, y_n)$, la coordenada predictiva condicional para la observación i se define:

$$\begin{aligned} CPO_i &= \frac{\pi(y | y_{(-i)})}{\pi(y)} \\ &= \frac{\pi(y)}{\pi(y_{(-i)})}. \end{aligned}$$

Si y_i brinda información relevante en el modelo $\pi(y)$ y $\pi(y_{(-i)})$ deberían ser muy próximos, valores de y_i que son mal ajustados por el modelo (Held et al., 2010). La coordenada predictiva condicional se puede expresar de la siguiente forma:

$$\begin{aligned}
CPO_i &= \frac{\pi(y)}{\int \pi(y_{(-i)}, \theta) d\theta} = \frac{\pi(y)}{\int f(y_{(-i)} | \theta) \cdot \pi(\theta) d\theta} \\
&= \frac{\pi(y)}{\int f(y_{(-i)} | \theta) \cdot \pi(\theta) \cdot \frac{\pi(y_i | y_{(-i)}, \theta)}{\pi(y_i | y_{(-i)}, \theta)} d\theta} \\
&= \frac{\pi(y)}{\int \frac{\pi(y_i | y_{(-i)}, \theta) \cdot \pi(y_{(-i)} | \theta) \cdot \pi(\theta)}{\pi(y_i | y_{(-i)}, \theta)} d\theta} \\
&= \frac{\pi(y)}{\int \frac{\pi(y | \theta) \cdot \pi(\theta)}{\pi(y_i | y_{(-i)}, \theta)} d\theta}.
\end{aligned}$$

Por el teorema de Bayes se sabe que $\pi(y | \theta) \cdot \pi(\theta) = \pi(\theta | y) \cdot \pi(y)$ luego tendríamos:

$$\begin{aligned}
CPO_i &= \frac{\pi(y)}{\int \frac{\pi(\theta | y) \cdot \pi(y)}{\pi(y_i | y_{(-i)}, \theta)} d\theta} \\
&= \frac{1}{\int \frac{1}{\pi(y_i | y_{(-i)}, \theta)} \cdot \pi(\theta | y) d\theta}.
\end{aligned}$$

Por ser y_i independiente de $y_{(-i)}$ dado θ , se cumple que $\pi(y_i | y_{(-i)}, \theta) = f(y_i | \theta)$ por lo que:

$$\begin{aligned}
CPO_i &= \frac{1}{\int \frac{1}{f(y_i | \theta)} \cdot \pi(\theta | y) d\theta} \\
&= \frac{1}{E(\theta | y) \cdot \left(\frac{1}{f(y_i | \theta)}\right)} \\
&= \left[E(\theta | y) \cdot \left(\frac{1}{f(y_i | \theta)}\right)\right]^{-1}.
\end{aligned}$$

Para poder calcular CPO utilizaremos la aproximación de Monte Carlo, es decir:

$$\widehat{CPO}_i = \left[\frac{1}{S} \sum_{j=1}^S \frac{1}{f(y_i | \theta^{(j)})} \right]^{-1}.$$

El CPO es una medida de bondad del ajuste para cada observación, que puede resumirse para los n datos mediante un único valor denominado logaritmo de la pseudoverosimilitud marginal (LPML); ya que se requiere una medida global, de modo que la comparación entre modelos puede realizarse de la siguiente forma:

$$LPML = \sum_{i=1}^n \log(\widehat{CPO}_i).$$

A mayor LPML mejor será el ajuste del modelo. Para evaluar la proximidad el valor observado \mathbf{y} con la estimación de la media a posteriori, se calcula la raíz del error cuadrático medio de estimación (RMSE) cuyo calculo es del siguiente modo:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}.$$

Capítulo 3

Modelos bayesianos espaciales multivariados

Los casos de enfermedades respiratorias pueden ser estudiados a nivel provincial en el Perú. En esta tesis se propone estudiar la correlación entre estas enfermedades tomando en cuenta el patrón espacial de estas enfermedades. Por lo que se aplicará modelos multivariados de regresión de Poisson que incluyen efectos espaciales en el predictor lineal para analizar estos datos (Lawson, 2018). En particular, las distribuciones a priori condicionalmente autorregresivas (CAR) son empleadas para modelar los efectos aleatorios espaciales.

3.1. Modelo 1: MCAR propio sin offset

El modelo multivariado de regresión de Poisson condicionalmente autorregresivo (MCAR) que se presenta a continuación fue propuesto por (Gómez-Rubio et al., 2019) y en esta tesis nos permite estudiar la distribución espacial de d diferentes enfermedades y de esta manera identificar los patrones de alto riesgo. Como las enfermedades con causas similares pueden mostrar patrones similares, se podría realizar un análisis multivariado para obtener mejores estimaciones de estos patrones compartidos, con la finalidad de construir un modelo conjunto para d diferentes enfermedades.

Se define Y_{id} como el número de casos de una enfermedad d en el área i , luego se asume

$$Y_{id} \sim \text{Poisson}(\mu_{id}) \quad (3.1)$$

donde $i = 1, \dots, n$; $d = 1, \dots, K$; donde μ_{id} representa el número de casos promedio de la enfermedad d en la i -ésima área. Se usa una función de enlace logarítmica para enlazar la

media μ_{id} al predictor lineal η_{id} , de la siguiente manera:

$$\log(\mu_{id}) = \eta_{id} = X_{id}^\top \beta_d + \theta_{id}, \quad (3.2)$$

donde X_{id} es un vector de covariables, β_d es un vector de coeficientes de regresión y θ_{id} representa el efecto aleatorio multivariante espacial. La forma de modelar θ_{id} puede ser compleja, donde el efecto latente espacial multivariante puede verse como la combinación de la variabilidad entre enfermedades y la variabilidad espacial de cada enfermedad. Ambos se modelan con sus respectivas matrices de varianza y sus propios hiperparámetros (Martínez-Beneito y Botella-Rocamora, 2019).

Los efectos aleatorios espaciales se representan mediante la matriz Θ con las entradas θ_{id} correspondiente al valor de la variable respuesta de la enfermedad $d = 1, \dots, K$ en el área $i = 1, \dots, n$. Sea $\Theta_{i\bullet}$ la i -ésima fila de la matriz Θ y $\Theta_{\bullet d}$ la d -ésima columna de la matriz Θ . La matriz Θ se define a través del operador $vec(\cdot)$, tal que:

$$vec(\Theta) = (\Theta_{\bullet 1}^\top, \dots, \Theta_{\bullet K}^\top)^\top.$$

Luego para modelar la variabilidad dentro de cada enfermedad se puede elegir la distribución CAR propia multivariada propuesta por (Perales, 2020), la cual es una extensión del modelo CAR propio definido en la sección 2.2.1. Luego la distribución de Θ es definida:

$$vec(\Theta) \sim N(0, \mathbf{\Gamma}^{-1} \otimes (\mathbf{D} - \alpha \cdot \mathbf{W})^{-1}), \quad (3.3)$$

donde $\mathbf{Q} = \mathbf{\Gamma} \otimes (\mathbf{D} - \alpha \mathbf{W})$ es la matriz de precisión, siendo α un mismo parámetro de autocorrelación espacial para todas las enfermedades, \mathbf{D} es una matriz diagonal compuesta por el número de vecinos del área i , \mathbf{W} es una matriz de vecindad, compuesta por elementos w_{ij} igual a 1, si las áreas i y j son vecinas, y cero caso contrario, y por último $\mathbf{\Gamma}$ controla la variabilidad restante entre las enfermedades. Finalmente $\mathbf{\Gamma}^{-1}$ es definida como:

$$\mathbf{\Gamma}^{-1} = \begin{bmatrix} 1/\tau_1 & \rho_{12}/\sqrt{\tau_1\tau_2} & \cdots & \rho_{1K}/\sqrt{\tau_1\tau_K} \\ \rho_{12}/\sqrt{\tau_1\tau_2} & 1/\tau_2 & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1K}/\sqrt{\tau_1\tau_K} & \cdots & \cdots & 1/\tau_K \end{bmatrix}$$

donde ρ_{jl} es el coeficiente de correlación de las enfermedades j y l , y τ_d es la precisión marginal de la enfermedad d .

3.2. Modelo 2: MCAR propio con offset

La distribución multivariada de los casos observados de la enfermedad d en el área i , Y_{id} , siguen una distribución de Poisson, tal que:

$$Y_{id} \sim \text{Poisson}(\mu_{id} = E_{id}R_{id}); \quad i = 1, 2, \dots, n; \quad d = 1, \dots, K \quad (3.4)$$

donde E_{id} es llamado compensación (en inglés offset) que representa el número de casos esperados con la enfermedad d en el área i o casos posibles con la enfermedad d en el área i , y R_{id} es el riesgo relativo para la enfermedad d en el área i . Luego este riesgo relativo se modela en un segundo nivel como una suma de diferentes términos:

$$\log(R_{id}) = \eta_{id} = \log(E_{id}) + X_{id}^{\top} \beta_d + \theta_{id}, \quad (3.5)$$

donde X_{id} es un vector de covariables, β_d es un vector de coeficientes de regresión y θ_{id} representa el efecto aleatorio multivariante espacial CAR como se definió en la ecuación 3.3,

$$\text{vec}(\Theta) \sim N(0, \Gamma^{-1} \otimes (\mathbf{D} - \alpha \cdot \mathbf{W})^{-1}).$$

3.3. Inferencia bayesiana usando INLA

Bajo el enfoque bayesiano de forma general se pueden definir los modelos presentados como modelos gaussianos latentes, donde:

- **Vector Observado:** Las d variables respuesta se representan mediante la matriz \mathbf{Y} con las entradas $\mathbf{Y}_{\mathbf{id}}$ correspondientes al valor de la variable respuesta de la enfermedad $d = 1, \dots, K$ en el área $i = 1, \dots, n$. Sea $\mathbf{Y}_{\mathbf{i}\bullet}$ la i -ésima fila de la matriz \mathbf{Y} y $\mathbf{Y}_{\bullet\mathbf{d}}$ la d -ésima columna de la matriz \mathbf{Y} . Dado el vector latente \mathbf{x} y los hiperparámetros θ , $\mathbf{Y}_{\mathbf{id}}$ son condicionalmente independientes, con distribución definida en las ecuaciones luego la fdp conjunta de \mathbf{Y} es: $\pi(y | \mathbf{x}, \theta) = \prod_{d=1}^K \prod_{i=1}^n \pi(y_{id} | \mathbf{x}, \theta)$.
- **Campo Gaussiano latente:** Asumiendo una a priori gaussiana para los coeficientes de regresión β_d , entonces β , vector compuesto por los vectores β_d para $d = 1, \dots, K$, tiene distribución normal. También se asume una CAR propia multivariada para $\text{vec}(\Theta) \sim N(0, \Gamma^{-1} \otimes (\mathbf{D} - \alpha \cdot \mathbf{W})^{-1})$. Luego se asume una distribución normal multivariada sobre $\mathbf{x} = \{\beta, \text{vec}(\Theta)\}$,

$$\mathbf{x} | \theta \sim N(0, \mathbf{Q}^{-1}(\theta))$$

- Hiperparámetros: $\boldsymbol{\theta} = (\alpha, \Gamma)$

$$\pi(\boldsymbol{\theta}).$$

El conjunto de hiperparámetros a estimar está compuesto por K precisiones τ_d , $\frac{K(K-1)}{2}$ parámetros de correlación ρ_{jl} ($j, l = 1, \dots, K$), y un parámetro común de autocorrelación espacial, α , es decir un total de hiperparámetros θ de $\frac{K(K+1)}{2} + 1$. Por ello en vez de asignar a priori para las precisiones se asigna una a priori para la matriz Γ . En la implementación se considera que $\alpha \sim U(\alpha_{min}, \alpha_{max})$; es decir, una distribución uniforme siendo α_{min} y α_{max} los límites del dominio de α . Una distribución de Wishart se considera como una distribución a priori conjunta para la matriz

$$\mathbf{\Lambda} \sim \text{Wishart}_K(\mathbf{r}, \mathbf{R}^{-1}),$$

donde r es el número de grados de libertad y \mathbf{R}^{-1} es una matriz fija simétrica definida positiva de dimensión $K \times K$. Bajo las definiciones previas, la distribución a posteriori de \mathbf{x} y θ está dada por:

$$\pi(\mathbf{x}, \theta | y) \propto \pi(y | \mathbf{x}, \theta) \cdot \pi(\mathbf{x} | \theta) \cdot \pi(\theta).$$

Las estimaciones a posteriori no son fáciles de obtener, y por ello se hace uso del INLA, que nos facilita el ahorro en el tiempo de cálculo, proporcionándonos aproximaciones de las marginales a posteriori de las variables latentes e hiperparámetros. La distribución marginal de θ es aproximada por:

$$\pi(\theta | \mathbf{y}) \approx \frac{\pi(\mathbf{y} | \mathbf{x}, \theta) \pi(\mathbf{x}, \theta) \pi(\theta)}{\pi(\mathbf{x} | \theta, \mathbf{y})} \Bigg|_{\mathbf{x}=\mathbf{x}^*(\theta)} = \tilde{\pi}_{\mathbf{G}}(\theta | \mathbf{y}),$$

donde $\pi_{\mathbf{G}}$ es una aproximación gaussiana. De forma similar se aproxima $\pi(\mathbf{x}_i | \theta, \mathbf{y})$. Luego las marginales a posteriori para el vector latente se pueden aproximar por:

$$\pi(\mathbf{x}_i | \mathbf{y}) = \int \tilde{\pi}(\mathbf{x}_i | \theta, \mathbf{y}) \tilde{\pi}(\theta | \mathbf{y}) d\boldsymbol{\theta}.$$

y las marginales a posteriori para los hiperparámetros también se obtienen numéricamente a partir de:

$$\pi(\theta_k | \mathbf{y}) = \int \tilde{\pi}(\theta | \mathbf{y}) d\boldsymbol{\theta}_{-k},$$

donde $\boldsymbol{\theta}_{-k}$ denota todos los hiperparámetros excepto el componente k .

Capítulo 4

Estudio de Simulación

En este capítulo, se realiza un estudio de simulación con el objetivo de mostrar la correcta estimación de los parámetros del modelo MCAR utilizando la librería INLAMSM a través del R-INLA, que se encuentra disponible en www.r-inla.org. Los resultados obtenidos para esta simulación se obtuvieron mediante una computadora con las siguientes características, un procesador Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz, una memoria RAM de 8GB y un sistema operativo de 64 bits, procesador basado en x64. Para simular los datos de áreas en las $n = 196$ provincias del Perú, se generó un grafo para definir una matriz de vecindad (\mathbf{W}) en base a las provincias del Perú, como se muestra en la Figura 4.1, donde los cuadrados amarillos representan las provincias no vecinas (valores de w_{ij} igual a cero) y los cuadrados rojos las provincias vecinas (valores de w_{ij} igual a uno). A partir de esta matriz de vecindad es evidente que cada provincia solo tiene algunas provincias vecinas.

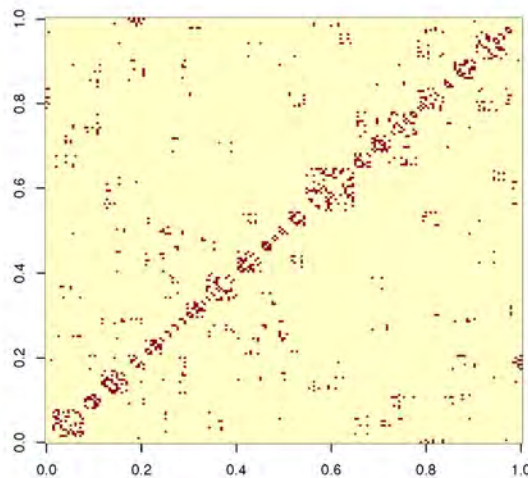


Figura 4.1: Matriz de vecindad asociada a las provincias del Perú.

4.1. Simulación con interceptos

Se simularon datos a partir del modelo definido en la sección 3.1 para $K = 2$, para generar datos de una distribución bivariada. Para ello se asignaron los siguientes valores a los parámetros: $\beta_{01} = 1$, $\beta_{02} = 5$, $\alpha=0.8$, $\tau = (\tau_1, \tau_2) = (0.3; 0.5)$ y $\rho=0.8$. Para simular datos espaciales de un modelo multivariado CAR se calculó la matriz de precisión $\Gamma \otimes (D - \alpha \cdot W)$, donde $\Gamma^{-1} = \begin{bmatrix} 1/\tau_1 & \rho/\sqrt{\tau_1\tau_2} \\ \rho/\sqrt{\tau_1\tau_2} & 1/\tau_2 \end{bmatrix}$ y $D = \text{diag}(w_{i+})$.

Luego se simularon los efectos espaciales θ_{id} , $d = 1, 2$ para cada variable respuesta a partir de $N(0, \Gamma^{-1} \otimes (D - \alpha \cdot W)^{-1})$. A continuación se calculó $\mu_{id} = \exp(\beta_0 + \theta_{id})$, y finalmente generamos la variable respuesta según una distribución de Poisson, $Y_{id} \sim \text{Poisson}(\mu_{id})$. La Figura 4.2 muestra los histogramas de ambas variables respuestas del modelo multivariado condicional autorregresivo (MCAR).

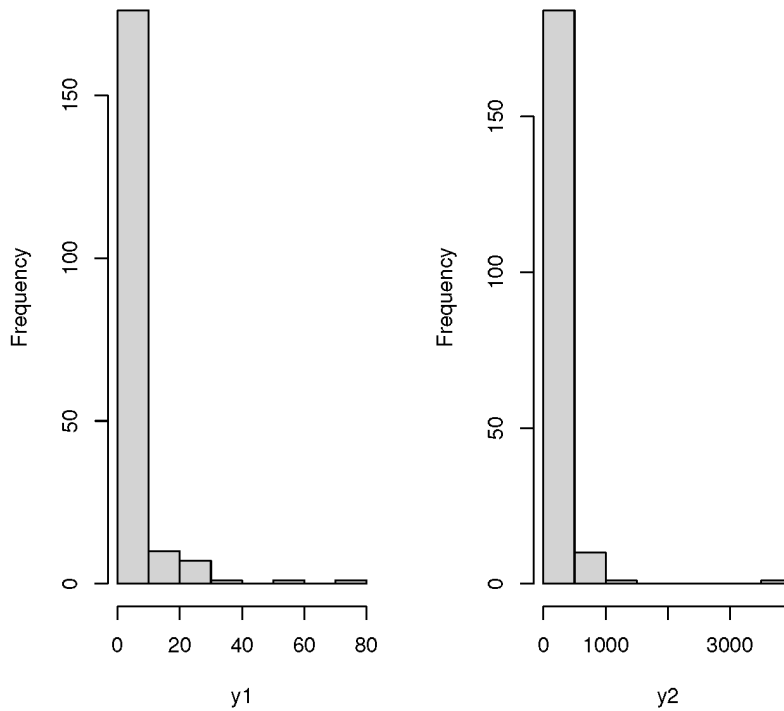


Figura 4.2: Histogramas de las variables respuestas del modelo CAR multivariado; y_1 para $d = 1$ (izquierda), y_2 para $d = 2$ (derecha).

La Figura 4.3 muestra los mapas de los datos simulados de las variables respuesta con el modelo bivariado expuesto anteriormente para $d = 1$ (izquierda) y $d = 2$ (derecha).

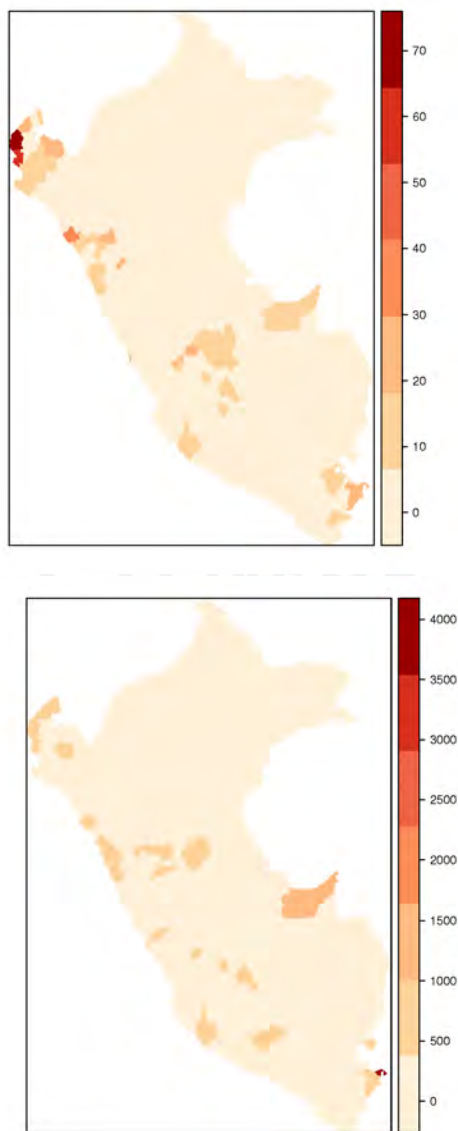


Figura 4.3: Datos simulados de las $K=2$ variables respuestas del modelo MCAR para $d=1$ (arriba) y $d=2$ (abajo).

Luego se ajustó el modelo multivariado CAR propio a través del INLA como se describe en la sección 3.3. Observamos que a pesar de la complejidad del modelo multivariado, el tiempo de empleo para la estimación es por debajo de los 10.6 segundos. En el Cuadro 4.1 se compara los valores de los coeficientes de regresión verdaderos con las estimaciones a posteriori. Se puede observar que las estimaciones de las medias a posteriori son cercanas a los valores originales. Las distribuciones a posteriori de los coeficientes de regresión se muestran en la Figura 4.4.

Cuadro 4.1: Valores de los coeficientes de regresión con las estimaciones a posteriori.

	original	media	d.s.
β_{01}	1	0.648	0.198
β_{02}	5	4.868	0.127

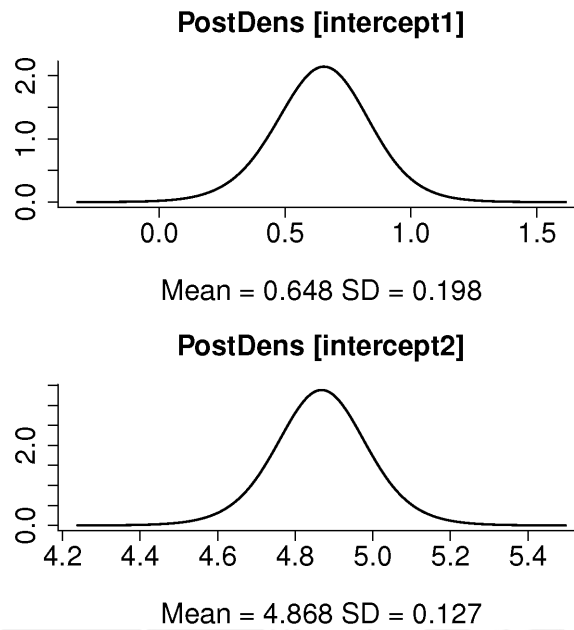


Figura 4.4: Estimación de la distribución marginal a posteriori de los coeficientes de regresión.

Además los intervalos de credibilidad incluyen el valor original de los coeficientes de regresión verdaderos, al igual que de los hiperparámetros como se muestra en el Cuadro 4.2

Cuadro 4.2: Estimaciones a posteriori de β_{01} , β_{02} y α , τ_1 , τ_2 y ρ .

Parámetros	original	Cuantil		
		2.5	50	97.5
β_{01}	1	0.248	0.648	1.037
β_{02}	5	4.613	4.868	5.123
α	0.8	0.727	0.879	0.954
τ_1	0.3	0.177	0.238	0.324
τ_2	0.5	0.422	0.521	0.641
ρ	0.8	0.662	0.759	0.833

Como podemos observar de los resultados, se ha podido recuperar los parámetros asignados inicialmente. La Figura 4.5 muestran los mapas de las estimaciones de las medias a posteriori a través el modelo MCAR ajustado.

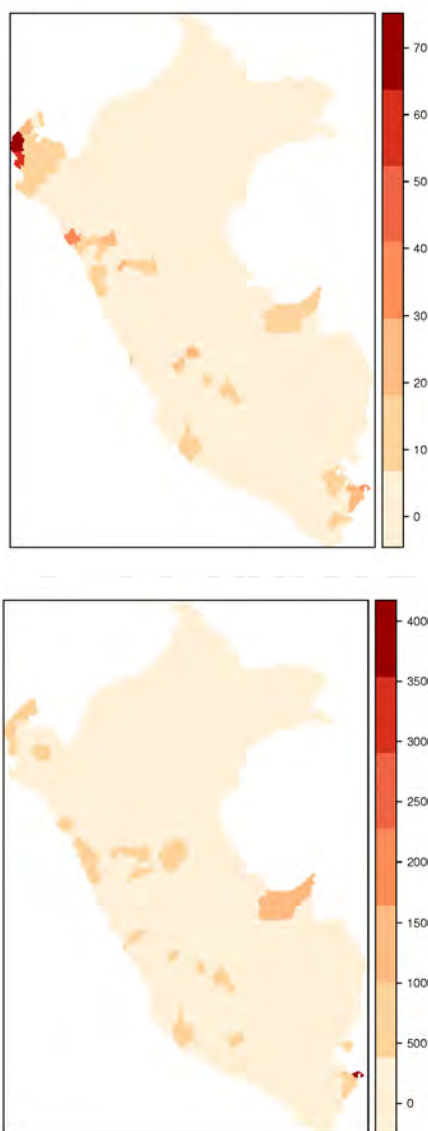


Figura 4.5: Mapas de las medias a posteriori estimadas de las variables respuestas a través del modelo CAR multivariado, para $d = 1$ (arriba) y $d = 2$ (abajo).

4.2. Simulación con interceptos y covariables

De forma similar, para simular datos del modelo de la sección 3.1 se asignaron los siguientes valores a los parámetros $\beta_{01} = 1$, $\beta_{02} = 5$, $\beta_{11} = -2$, $\beta_{12} = 3$, $\alpha = 0.8$, $\tau = (\tau_1, \tau_2) = (0.3; 0.5)$ y $\rho = 0.8$. Para simular datos espaciales de un modelo multivariado CAR se calculó la matriz de precisión $\Gamma \otimes (D - \alpha \cdot W)$, donde $\Gamma^{-1} = \begin{bmatrix} 1/\tau_1 & \rho/\sqrt{\tau_1\tau_2} \\ \rho/\sqrt{\tau_1\tau_2} & 1/\tau_2 \end{bmatrix}$ y $D = \text{diag}(w_{i+})$. Luego se simularon los efectos espaciales θ_{id} , $d = 1, 2$ para cada variable respuesta a partir de $N(0, \Gamma^{-1} \otimes (D - \alpha \cdot W)^{-1})$. A continuación se calculó $\mu_{id} = \exp(\beta_0 + \beta_1 x + \theta_{id})$, y finalmente generamos la variable respuesta según una distribución de Poisson, $Y_{id} \sim \text{Poisson}(\mu_{id})$. En la

Figura 4.6 se muestra los histogramas de ambas variables respuestas del modelo multivariado.

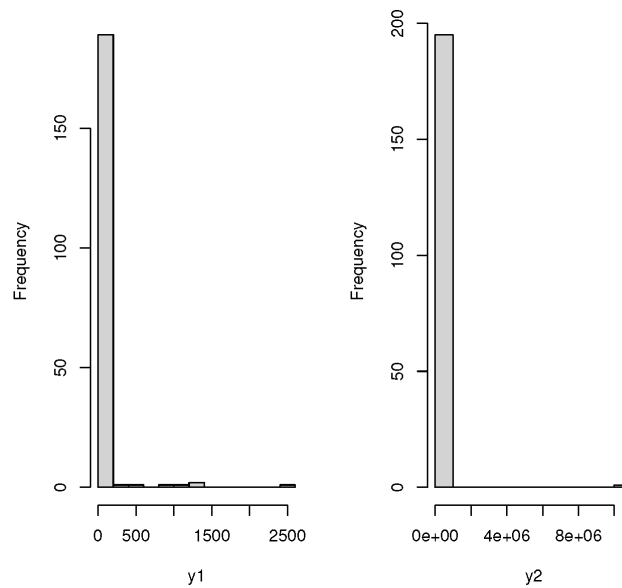


Figura 4.6: Histogramas de las variables respuestas del modelo CAR multivariado; y_1 para $d = 1$ (izquierda), y_2 para $d = 2$ (derecha).

La Figura 4.7 muestra los mapas de los datos simulados de las variables respuesta con el modelo bivariado expuesto anteriormente para $d = 1$ (izquierda) y $d = 2$ (derecha).

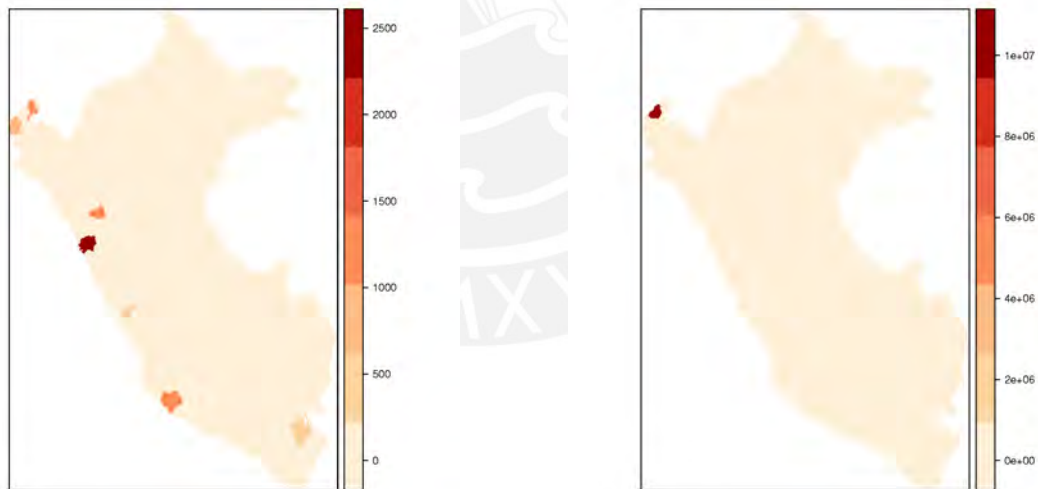


Figura 4.7: Datos simulados de las $K=2$ variables respuestas del modelo CAR multivariado, para $d = 1$ (izquierda) y $d = 2$ (derecha).

Luego se ajusto el modelo multivariado CAR propio implementado en el INLA. Observamos que el tiempo de empleo para la estimación está por debajo de los 10 segundos. En el Cuadro 4.3 compara los valores de los coeficientes de regresión verdaderos con las estimaciones a posteriori. Se puede observar que las estimaciones de las medias a posteriori son cercanas a los valores originales. Las distribuciones a posteriori de los coeficientes de regresión

se muestran en la Figura 4.8.

Cuadro 4.3: Valores de los coeficientes de regresión con las estimaciones a posteriori.

	original	media	d.s.
β_{01}	1	0.592	0.187
β_{02}	5	4.838	0.118
β_{11}	-2	-2.208	0.096
β_{12}	3	2.974	0.052

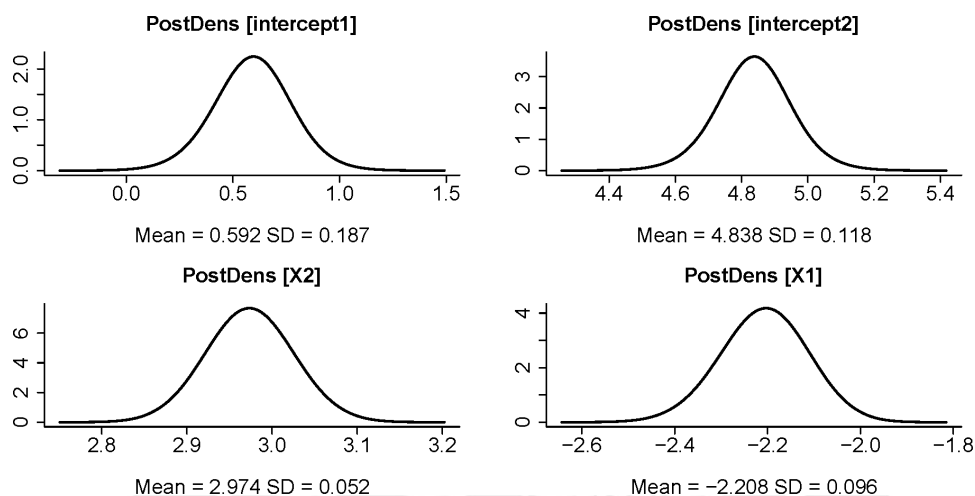


Figura 4.8: Estimación de la distribución marginal a posteriori de los coeficientes de regresión.

Además los intervalos de credibilidad incluyen el valor original de los coeficientes de regresión y de los hiperparámetros como se muestra en el Cuadro 4.4. Como podemos observar de los resultados se ha podido recuperar los parámetros asignados inicialmente.

Cuadro 4.4: Valores de los coeficientes de regresión con las estimaciones a posteriori

Parámetros	original	Cuantil		
		2.5	50	97.5
β_{01}	1	0.218	0.592	0.958
β_{02}	5	4.602	4.838	5.072
β_{11}	-2	-2.400	-2.208	-2.023
β_{12}	3	2.872	2.974	3.077
α	0.8	0.685	0.852	0.944
τ_1	0.3	0.186	0.253	0.341
τ_2	0.5	0.417	0.522	0.652
ρ	0.8	0.708	0.811	0.882

La Figura 4.9 muestra los mapas de las estimaciones de las medias a posteriori a través el modelo MCAR ajustado.

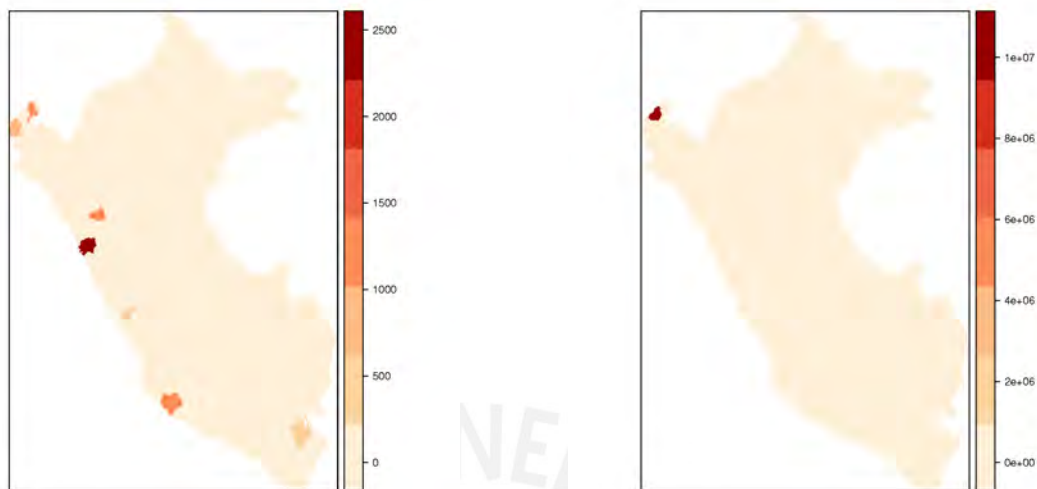


Figura 4.9: Mapas de las medias a posteriori estimadas de las variables respuestas a través del modelo CAR multivariado, para $d = 1$ (izquierda) y $d = 2$ (derecha).

4.3. Escenarios de simulaciones bajo diferentes ρ

Ahora como el parámetro ρ mide la correlación entre variables, la estimación del mismo es crucial, pues a menor valor menos correlacionadas estarán las variables, y a mayor valor más correlacionadas estarán las variables. Por ello realizamos simulaciones para diferentes valores del parámetro de autocorrelación ρ igual a 0.6, 0.5, 0.4, 0.3 y 0.2, y los mismos valores para el resto de parámetros. En particular, se tomó $\alpha = 0,8$ que también es de interés en este estudio, porque implica que existe una autocorrelación espacial significativa.

En el Cuadro 4.5 muestra los valores originales de los parámetros en las simulaciones y luego para cada valor diferente de ρ se muestran los cuantiles a posteriori estimados para los parámetros en cada escenario. Observamos que el verdadero valor de los parámetros se encuentra dentro de los intervalos de credibilidad, por ello concluimos que la estimación a posteriori es adecuada. Y la mediana a posteriori es muy cercana al verdadero valor del parámetro, en todos los casos. Asimismo no se evidenció un patrón en las estimaciones a posteriori para diferentes valores de ρ . Reflejando que el modelo es capaz de estimar adecuadamente los parámetros cuando la correlación entre las variables es baja a moderada.

Cuadro 4.5: Estimaciones a posteriori de β_{01} , β_{02} , β_{11} , β_{12} , α , τ_1 , τ_2 para diferentes valores de ρ

Rho	Parámetros	original	Cuantil		
			2.5	50	97.5
$\rho = 0.2$	β_{01}	1	0.219	0.601	0.973
	β_{02}	5	4.750	4.998	5.244
	β_{11}	-2	-2.400	-2.196	-2.001
	β_{12}	3	2.874	2.98	3.089
	α	0.8	0.699	0.861	0.949
	τ_1	0.3	0.188	0.261	0.363
	τ_2	0.5	0.404	0.507	0.635
	ρ	0.2	0.033	0.233	0.413
$\rho = 0.3$	β_{01}	1	0.204	0.598	0.982
	β_{02}	5	4.721	4.977	5.230
	β_{11}	-2	-2.401	-2.198	-2.004
	β_{12}	3	2.879	2.985	3.093
	α	0.8	0.722	0.874	0.956
	τ_1	0.3	0.197	0.264	0.355
	τ_2	0.5	0.410	0.515	0.646
	ρ	0.3	0.143	0.332	0.500
$\rho = 0.4$	β_{01}	1	0.207	0.591	0.964
	β_{02}	5	4.698	4.943	5.186
	β_{11}	-2	-2.413	-2.208	-2.014
	β_{12}	3	2.873	2.979	3.087
	α	0.8	0.689	0.857	0.947
	τ_1	0.3	0.188	0.257	0.351
	τ_2	0.5	0.406	0.509	0.635
	ρ	0.4	0.255	0.431	0.582
$\rho = 0.5$	β_{01}	1	0.199	0.587	0.964
	β_{02}	5	4.668	4.914	5.157
	β_{11}	-2	-2.417	-2.214	-2.02
	β_{12}	3	2.879	2.984	3.091
	α	0.8	0.697	0.861	0.949
	τ_1	0.3	0.189	0.255	0.343
	τ_2	0.5	0.415	0.519	0.647
	ρ	0.5	0.370	0.533	0.667
$\rho = 0.6$	β_{01}	1	0.191	0.587	0.975
	β_{02}	5	4.629	4.884	5.137
	β_{11}	-2	-2.413	-2.213	-2.023
	β_{12}	3	2.890	2.995	3.102
	α	0.8	0.710	0.870	0.954
	τ_1	0.3	0.190	0.258	0.351
	τ_2	0.5	0.414	0.519	0.650
	ρ	0.6	0.487	0.629	0.740

Capítulo 5

Aplicación

El presente capítulo tiene como objetivo aplicar el modelo multivariado CAR propio, a un conjunto de datos reales sobre las Infecciones Respiratorias Agudas (IRA) que habitualmente han sido una de las principales causas de mortalidad de los niños que residen en el área rural o en regiones de la sierra y selva del Perú.

Los datos corresponden a recuentos de personas que tienen problemas de IRA sin considerar neumonías (Y_{i1}) y neumonías (Y_{i2}) en $n = 196$ provincias del Perú en el año 2021. En este análisis estadístico se estudia la posible relación entre ambas enfermedades. A través de las provincias del Perú y su ubicación, es adecuado construir un grafo donde una arista representa si dos provincias son vecinas, es decir, comparten líneas fronterizas comunes, como se muestra en la Figura 5.2. La matriz de vecindad \mathbf{W} para este grafo se muestra en la Figura 5.1, donde los cuadrados amarillos representan valores cero de \mathbf{W} y los cuadrados rojos representan los valores igual a uno de \mathbf{W} .

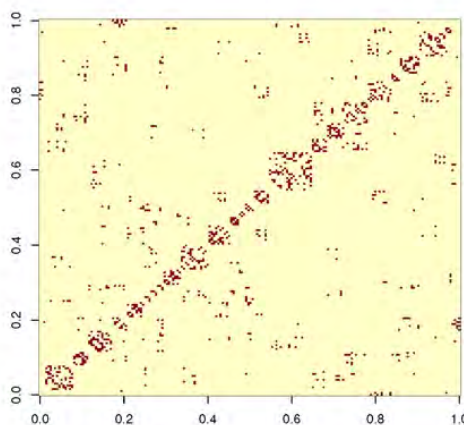


Figura 5.1: Matriz de Vecindad \mathbf{W} asociada a las provincias del Perú.



Figura 5.2: Representación del grafo de las provincias de Perú.

En general, se asume que Y_{id} es el número de casos de IRA sin considerar neumonía ($d = 1$) en la i -ésima provincia o el número de casos de neumonía ($d = 2$) en la i -ésima provincia. Consideramos $Y_{id} \sim \text{Poisson}(\mu_{id} = E_{id}R_{id})$, donde μ_{id} representa la media de casos de la enfermedad d en la provincia i , E_{id} se define como la población expuesta a la enfermedad d en la i -ésima provincia y R_{id} se define como la incidencia de la enfermedad d en la i -ésima provincia. Luego se aplicó el modelo multivariado MCAR detallado en la sección 3.2.

Ahora realizaremos la modelación de las tasas de las enfermedades en las provincias del Perú, por intermedio del modelo multivariado CAR, para así explicar el número de casos de las enfermedades con respecto a la población en cada provincia del Perú. Se ajustaron seis modelos con diferentes covariables (precip= media de precipitaciones, pobreza = porcentaje de pobreza, seguro = porcentaje con seguro de salud y agua.potable=porcentaje de agua potable) y efectos aleatorios espaciales (efecto), por lo que se calcularon criterios de selección para determinar el mejor modelo, los resultados se muestran en el Cuadro 5.1.

Cuadro 5.1: Criterios de selección de modelos.

Modelo	DIC	WAIC	-LPML	RMSE
M1: interc.+efecto	3345.48	3262.47	2922.59	8.8E-05
M2: interc.+precip.+efecto	3344.59	3261.85	2493.02	8.6E-05
M3: interc.+pobreza+efecto	3345.99	3263.75	2979.43	0.0001
M4: interc.+precip.+pobreza+efecto	3342.91	3259.88	2976.57	0.0001
M5: interc.+precip.+seguro+efecto	3344.76	3262.73	2501.45	9.0E-05
M6: interc.+precip.+agua.potable+efecto	3344.66	3262.57	2868.87	8.6E-05

De acuerdo a los criterios de selección el modelo M2, donde interviene la covariable precipitaciones (precip.) resultó ser el mejor pues presentó el menor LPML y RMSE, y segundo mejor WAIC. Seguido del modelo M4 quien tuvo el mejor WAIC pero quinto según el LPML. De acuerdo a estos resultados a continuación presentamos los resúmenes a posteriori para el modelo M2.

En el Cuadro 5.2 se muestran las estimaciones de la mediana posteriori y de los intervalos de credibilidad al 95 % (IC) de los parámetros obtenidos a través del modelo Poisson-CAR con offset descrito en la sección 3.2. Los IC al 95 % de los coeficientes de regresión no contienen el cero, por lo tanto son significativamente diferentes de cero.

Cuadro 5.2: Mediana a posteriori e intervalo de credibilidad al 95 % para los hiperparámetros.

	Cuantil		
	2.5	50	97.5
β_{01}	-3.636	-3.505	-3.373
β_{11}	-9.852	-9.561	-9.279
β_{02}	0.046	0.105	0.164
β_{12}	0.156	0.282	0.408
α	0.165	0.426	0.685
τ_1	0.490	0.602	0.737
τ_2	0.117	0.152	0.194
ρ	0.356	0.489	0.605

A partir del predictor lineal definido en la ecuación 3.5, se tiene que

$$\log\left(\frac{R_{id}}{E_{id}}\right) = \log(\mu_{id}) = X_{id}^{\top}\beta_d + \theta_{id}.$$

Entonces para interpretar el coeficiente de regresión asociado a una covariable se estudia esta medida cuando se aumenta en una unidad la covariable, así el cambio en la media μ_{id} se calcula a partir de la exponencial del coeficiente de regresión. Estos resultados se muestran

en el cuadro Cuadro 5.3. Podemos observar que por cada unidad en que se incrementan la media de las precipitaciones, la media de casos de IRA sin neumonía se reduce en 99.99 %, mientras al aumentar la media de las precipitaciones en una unidad, la media de casos de neumonías aumenta en 32.58 %.

Cuadro 5.3: Medida para interpretar los coeficientes de regresión del modelo de Poisson-CAR con offset

	e^{β}
β_{01} intercepto (IRA sin neumonía)	0.03005
β_{11} precipitaciones (IRA sin neumonía)	0.00007
β_{02} intercepto (neumonía)	1.11071
β_{12} precipitaciones (neumonía)	1.32578

Las distribuciones marginales a posteriori de los hiperparámetros se muestran en la Figura 5.3. La mediana a posteriori del parámetro de autocorrelación espacial ($\alpha = 0.426$) indica que existe evidencia de autocorrelación espacial moderada en las tasas de cada enfermedad entre las provincias vecinas. Así si una provincia tiene una tasa alta de la enfermedad respiratoria d , una región vecina también presenta una tasa alta de la enfermedad respiratoria d , para $d = 1, 2$. Los parámetros de precisión distintos indican que la variabilidad espacial de ambas enfermedades respiratorias IRA sin neumonía y neumonía son diferentes, siendo mayor para la tasa de neumonía. El valor de ρ que está alrededor de 0.50, indica que existe una correlación positiva entre los casos de enfermos con IRA sin neumonía y neumonías.

La Figura 5.4 muestra mapas de las estimaciones de las medias a posteriori de los efectos aleatorios espaciales $\Theta_{\bullet,d}$ a través del modelo CAR multivariado ajustado. Se puede observar claramente un patrón espacial similar entre ambas enfermedades. Finalmente, la Figura 5.5 muestra mapas de las estimaciones de las medias a posteriori de la incidencia de IRA sin neumonía y la incidencia de neumonías a través del modelo CAR multivariado ajustado. En primer lugar en efecto el mapa muestra que existe una correlación entre ambas incidencias, pues en general en una provincia con alta incidencia de enfermedades respiratorias sin neumonía también hay una alta incidencia de neumonías. Y en particular, se observa que las incidencias de ambas enfermedades son mayores en la región sur del país en arequipa, cuzco, en mayor medida en la costa en Lima, Lambayeque , así como en la región de la selva al norte.

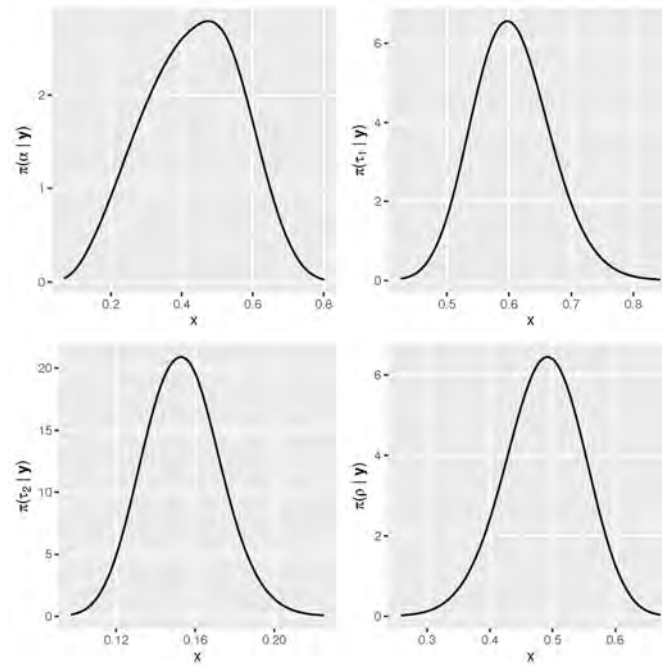


Figura 5.3: Gráficas de las distribuciones marginales a posteriori de los hiperparámetros.

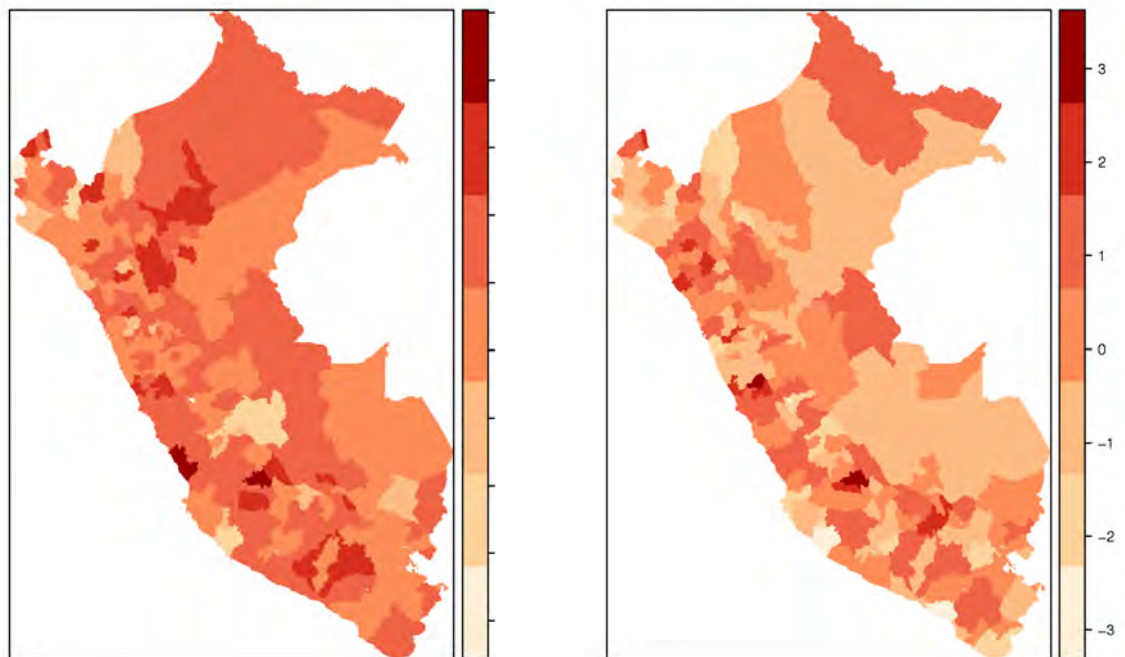


Figura 5.4: Mapas de las medias a posteriori estimadas de los efectos aleatorios espaciales a través del modelo CAR multivariado, para $d = 1$ (izquierda) y $d = 2$ (derecha).

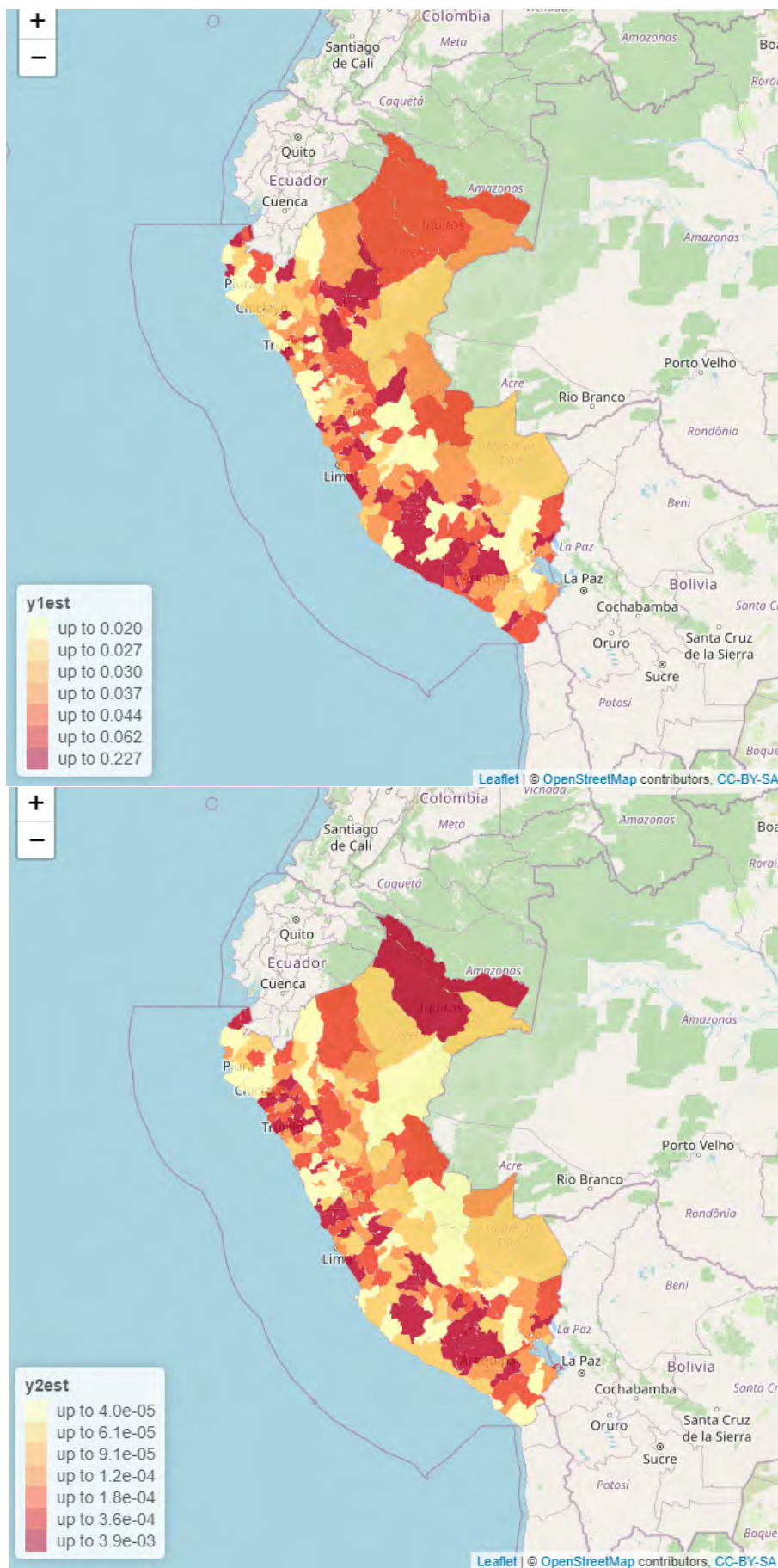


Figura 5.5: Mapas de las medias a posteriori estimadas de las variables respuestas a través del modelo CAR multivariado: $y1_{est}$ para $d = 1$ (arriba); $y2_{est}$ para $d = 2$ (abajo).

Capítulo 6

Conclusiones

El objetivo principal es estimar para así detectar la asociación entre enfermedades respiratorias a nivel provincial en el Perú, tomando en cuenta la distribución espacial de estas enfermedades en las provincias del Perú. Para ello se propuso aplicar un modelo espacial multivariado con efectos aleatorios condicionales autoregresivos al conjuntos de datos de áreas usando inferencia bayesiana de manera eficiente a través del método de integración aproximada anidada de Laplace (INLA). En ese estudio se observó la bondad de ajuste del modelo y la eficacia del método en cada simulación realizada ya que nos proporcionaba tiempos por debajo de los 15 segundos que en métodos clásicos tomaría mucho más tiempo.

Con respecto a los resultados obtenidos, podemos concluir que existe correlación positiva directa entre los casos de personas con IRA sin neumonía y neumonías. Además hay evidencia estadística de que el aumento del promedio de precipitaciones en una provincia reduce significativamente la incidencia de IRA sin neumonía en dicha provincia, mientras el aumento del promedio de precipitaciones en una provincia aumenta la incidencia de neumonías en dicha provincia. El valor de la autocorrelación espacial indica que existe evidencia de autocorrelación espacial moderada en las tasas de cada enfermedad entre las provincias vecinas.

Por último, el trabajo mostrado podría analizarse con más covariables y factores que puedan explicar adecuadamente la variable respuesta de la investigación, además el modelo planteado puede ampliarse para aplicar un modelo espacio-temporal empleando INLA, pues dada su eficiencia computacional se podría trabajar con mucho más datos. Los datos de IRA sin neumonía y neumonías también pueden ser usados para estudiar la evolución temporal del patrón espacial de las enfermedades respiratorias por ejemplo de acuerdo a meses o años.

Bibliografía

- Banerjee, S., Carlin, B. P. y Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*, Chapman and Hall/CRC.
- Bayes, T. (1763). Lii. an essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s, *Philosophical transactions of the Royal Society of London* (53): 370–418.
- Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2): 192–225.
- Blangiardo, M., Cameletti, M., Baio, G. y Rue, H. (2013). Spatial and spatio-temporal models with r-inla, *Spatial and spatio-temporal epidemiology* **4**: 33–49.
- Box, G. E. y Tiao, G. C. (2011). *Bayesian inference in statistical analysis*, Vol. 40, John Wiley & Sons.
- Brook, D. (1964). On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour systems, *Biometrika* **51**(3/4): 481–483.
URL: <http://www.jstor.org/stable/2334154>
- Dobson, A. J. y Barnett, A. G. (2018). *An introduction to generalized linear models*, Chapman and Hall/CRC.
- Gelman, A., Hwang, J. y Vehtari, A. (2014). Understanding predictive information criteria for bayesian models, *Statistics and computing* **24**(6): 997–1016.
- Gilks, W. R., Richardson, S. y Spiegelhalter, D. (1995). *Markov chain Monte Carlo in practice*, CRC press.
- Gómez-Rubio, V., Palmi-Perales, F., López-Abente, G., Ramis-Prieto, R. y Fernández-Navarro, P. (2019). Bayesian joint spatio-temporal analysis of multiple diseases, *SORT-Statistics and Operations Research Transactions* pp. 51–74.

- Held, L., Schrödle, B. y Rue, H. (2010). Posterior and cross-validators predictive checks: a comparison of mcmc and inla, *Statistical modelling and regression structures: Festschrift in honour of ludwig fahrmeir* pp. 91–110.
- Hu, H. (2008). Poisson distribution and application, *A Course in Department of Physics and Astronomy; University of Tennessee at Knoxville: Knoxville, TN, USA* .
- Lawson, A. B. (2018). *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*, Chapman and Hall/CRC.
- Lindsey, J. K. (2000). *Applying generalized linear models*, Springer Science & Business Media.
- Martínez-Beneito, M. A. y Botella-Rocamora, P. (2019). *Disease mapping: from foundations to multidimensional modeling*, CRC Press.
- Nelder, J. A. y Wedderburn, R. W. (1972). Generalized linear models, *Journal of the Royal Statistical Society: Series A (General)* **135**(3): 370–384.
- Outzen Berild, M., Martino, S., Gómez-Rubio, V. y Rue, H. (2021). Importance sampling with the integrated nested Laplace approximation, *arXiv e-prints* pp. arXiv–2103.
- Palmí-Perales, F., Gómez-Rubio, V. y Martínez-Beneito, M. A. (2021). Bayesian multivariate spatial models for lattice data with INLA, *ournal of Statistical Software* **98**(2).
- Perales, F. P. (2020). *Bayesian multivariate spatial models for the joint analysis of several diseases*, PhD thesis, Universidad de Castilla-La Mancha.
- Pereira da Silva, H. D. (2016). Aplicación de modelos bayesianos para estimar la prevalencia de enfermedad y la sensibilidad y especificidad de tests de diagnóstico clínico sin gold standard.
- Poisson, S.-D. (1837). *Recherches sur la probabilité des jugements.*, Bachelier Paris.
- Rue, H. y Held, L. (2005). *Gaussian Markov random fields: theory and applications*, Chapman and Hall/CRC.
- Rue, H., Martino, S. y Chopin, N. (2009). Approximate bayesian inference for latent Gaussian models by using integrated nested Laplace approximations, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71**(2): 319–392.
- Song, J. J. (2004). *Bayesian multivariate spatial models and their applications*, Texas A&M University.

- Spiegelhalter, D. J., Best, N. G., Carlin, B. P. y Van Der Linde, A. (2002). Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* **64**(4): 583–639.
- Valero, N., Larreal, Y., Arocha, F., Gotera, J., Mavarez, A., Bermudez, J., Moran, M., Maldonado, M. y Espina, L. M. (2009). Etiología viral de las infecciones respiratorias agudas, *Investigación Clínica* **50**(3): 359–368.
- Watanabe, S. y Opper, M. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory., *Journal of machine learning research* **11**(12).

