PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Inferencia bayesiana aproximada para el modelo multivariado block-NNGP

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

Carlos Alberto Gonzales Pizango

Asesora:

Zaida Jesús Quiroz Cornejo

Lima, 2024

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 *Inferencia bayesiana aproximada para el modelo multivariado block-NNGP*, del autor Carlos Alberto Gonzales Pizango, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 10 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 07/08/2024.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 7 de agosto de 2024

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

Dedicatoria

A mis padres, Martha y Antonio, por su incansable esfuerzo y dedicación en brindarnos una educación; sin su apoyo, nada de esto habría sido posible. A mis hermanos, Oscar, Miguel e Ingri, por su constante ánimo y respaldo. A Emma, por su amor, apoyo y paciencia.

Agradecimentos

A los profesores de la Maestría en Estadística, quienes fortalecieron mis conocimientos en todos los cursos. De manera especial, a mi asesora, la profesora Zaida, por su invaluable apoyo, enseñanzas, paciencia y motivación a lo largo de este camino. A los amigos que encontré en la PUCP, por su apoyo, desafíos y respaldo en momentos difíciles.

Resumen

El estudio de las especies de aves es un excelente indicador de la biodiversidad o la productividad. Se sabe que el calentamiento global y los cambios en el uso de la tierra por parte de los humanos están afectando la abundancia de aves. En este estudio nos enfocamos en las especies Morning Dove y American Robin, las especies más abundantes en América del norte. Las abundancias de estas especies pueden estar correlacionadas entre sí y mostrar una distribución espacial similar. Por lo tanto, proponemos modelar estos datos simultáneamente a través de modelos multivariados espaciales que se basan en compartir términos comunes de efectos aleatorios espaciales gaussianos. Para mejorar la eficiencia computacional, los procesos espaciales gaussianos se aproximan a un proceso gaussiano de vecinos más cercanos por bloques (block-NNGP). El modelo geoestadístico multivariado pertenece a la clase de modelos gaussianos latentes, por ello se usó el método de aproximación de Laplace anidada integrada (INLA) que permite una inferencia bayesiana rápida. El rendimiento del modelo propuesto se demuestra a través de simulaciones y la aplicación a los datos de especies de aves.

Palabras-clave: Block-NNGP, geoestadistica, procesos gaussianos, GRMF, INLA, modelo multivariado.

Abstract

The study of birds species is an excellent indicator of biodiversity or productivity. Global warming and changes human land us are considered major threats to biodiversity, affecting the abundance of bird species. In this study we focus on the Mourning Dove and American Robin, the most abundant birds species in the United States. The abundances of these species can be correlated between them and they would also be similar in nearby locations. Thus we propose to model these data simultaneously through multivariate models that relies on sharing common spatial Gaussian random effect terms. In order to improve the computational efficiency, each spatial Gaussian process is approximated to the block nearest neighbor Gaussian process (block-NNGP). Since the multivariate geostatistical model belongs to the class of Latent Gaussian Models, fast Bayesian inference can be carried out through the Integrated Nested Laplace Approximation (INLA) method. The good performance of the proposed model is shown through simulations and our application to the bird species real data.

Keywords: Block-NNGP, geostatistics, Gaussian process, GRMF, INLA, multivariate models.

Índice general

Capítulo 1

Introducción

1.1. Consideraciones preliminares

Actualmente diversas áreas se enfrentan a desafíos al momento de analizar bases de datos debido a que los datos son multivariados, se cuenta con varias variables de respuesta de interés y referenciados geográficamente (Banerjee et al., 2015). Ante estos desafíos podemos utilizar un modelo multivariado el cual nos permite modelar dos o más variables como variables respuesta, cuando se tiene conocimiento de una posible asociación lineal entre ellas, y se desea estimarlas de forma conjunta.

La expansión en el uso de modelos multivariados es cada vez mayor, las aplicaciones van desde las geociencias y las ciencias atmosféricas hasta la vigilancia del medio ambiente, la economía y otras áreas (Gneiting et al., 2010). Por ejemplo, la presión atmosférica y la temperatura son variables cuya relación es conocida, pues a mayor (menor) temperatura la presión baja (aumenta). De la misma forma, existe relación entre las concentraciones de los diferentes metales pesados emitidos a la atmósfera y transportadas con el viento y depositadas en el suelo (Cortés et al., 2016). Otro ejemplo es el estudio realizado por (Fabijańczyk et al. (2016), Ramos et al. (1999)) donde detallan la correlación entre la concentración de plomo (Pb) y zinc (Zn) en muestras de agua, suspensiones y sedimentos en un lago urbano. En la práctica se puede proponer un modelo para cada una de estas variables como variable respuesta, sin embargo probablemente ambas variables son explicadas por las mismas covariables, y dado que existe una relación directa o inversa entre ambas, puede ser más interesante proponer un modelo multivariado para estimar ambas variables de forma conjunta.

Por otro lado, los datos multivariados indexados por coordenadas espaciales son de interés en una larga lista de aplicaciones. En los ejemplos citados previamente se puede estudiar de forma conjunta la distribución espacial entre las dos variables presión y temperatura, o entre

las dos variables de metales pesados (Pb y Zn), en determinada región de estudio, pues el patrón espacial entre las variables respuesta también puede ser similar.

Se han propuesto o estudiado algunos métodos y modelos multivariados para datos georeferenciados (Myers, 1983; Chiles y Delfiner, 2009; Gneiting et al., 2010; Palmí-Perales et al., 2023). Por ejemplo, Gneiting et al. (2010) propone un modelo con efectos aleatorios espaciales, los cuales se definen a través de un proceso gaussiano multivariado. La dependencia espacial entre los efectos aleatorios es incorporada a través de una función de covarianza cruzada Matérn, tomando en cuenta la dependencia espacial no solo entre ubicaciones georeferenciados, si no también entre las variables respuesta. Si estamos modelando K variables de respuesta aleatorias, el número de parámetros debido a los efectos aleatorios especiales en el primer enfoque es $K(K+1)$, mientras que para el último enfoque es $2K$. En ambos casos, para n ubicaciones, la matriz de covarianza del proceso gaussiano asociado a la realización de los efectos aleatorios espaciales tiene un orden de $\mathcal{O}(2Kn)$.

Existen varios enfoques para tratar grandes bases de datos georeferenciados. Un enfoque es dividir el dominio espacial en bloques independientes disjuntos. También hay otros enfoques que inducen cero en la matriz de covarianza, como covariance-tapering, reducción de rango de la matriz de covarianza (low-rank models), proceso gaussiano predictivo (Banerjee et al. (2008)), entre otros. Otro enfoque particular es representar el proceso gaussiano como un proceso aleatorio gaussiano de Markov, induciendo la dispersión en la matriz de presición, es decir en la inversa de la matriz de covarianza (Rue y Held, 2005). Por ejemplo, esta dispersión puede lograrse mediante ecuaciones diferenciales parciales estocásticas (SPDE, Lindgren et al. (2011)), procesos gaussianos del vecino más cercano (NNGP, Datta et al. (2016)) o el proceso gausiano de bloques de vecinos más cercanos (block-NNGP,Quiroz et al. (2023)), un método que fusiona los enfoques NNGP y bloques independientes.

En el modelo jerárquico multivariado propuesto por Palmí-Perales et al. (2023), los efectos aleatorios espaciales compartidos siguen una aproximación del proceso gaussiano a través de ecuaciones diferenciales parciales estocásticas (SPDE). En esta tesis presentamos un modelo multivariado con efectos espaciales compartidos, similar al propuesto Palmí-Perales et al. (2023), pero se propone aproximar los procesos gaussianos usando el blockNNGP en vez del SPDE. Este modelo funciona para cualquier función de covarianza válida, no solo la Matérn como el SPDE. Este modelo es útil para modelar grandes conjuntos de datos. La inferencia se realiza a través del método de Aproximaciones Laplace Anidades Integradas (INLA) (Rue et al., 2009), permitiendo ajustar modelos multivariados espaciales cuyas variables respuesta son gaussianos o no gaussianas. El modelo es capaz de lograr cálculos eficientes y paralelos

evitando los procesos de muestreos de cadenas de Markov secuenciales, que son consumidores de tiempo y potencialmente problemáticos. El enfoque ofrece una solución general a las dificultades computacionales comúnmente encontradas al tratar con grandes conjuntos de datos multivariados espaciales, al mismo tiempo proporciona una herramienta efectiva y flexible para modelar datos multivariados.

1.2. Objetivos

El objetivo general de la tesis es proponer un modelo multivariado para datos georeferenciados, usando el enfoque de bloques de vecinos más cercanos (block-NNGP) desde el punto de vista de la inferencia bayesiana. De manera específica:

- Proponer, estudiar, propiedades, e implementar la estimación de modelos geostadísticos multivariados a través del método de bloque de vecinos más cercano (block-NNGP).
- Implementar métodos de inferencia bayesiana considerando la métodología de aproximación integrada anidada de Laplace (INLA)
- Realizar estudios de simulación acerca del modelo multivariado usando block-NNGP considerando diferentes escenarios.
- Aplicar el modelo propuesto a conjunto de datos reales.

1.3. Organización del trabajo

En el capítulo 2 se presentan una serie de conceptos preliminares asociados a los modelos geoestadísticos, como: autocorrelación espacial, procesos espaciales, isotropía, variogramas. También se revisa brevemente sobre inferencia bayesiana y el método INLA, estas definiciones permitirán comprender y aplicar adecuadamente el modelo propuesto. En el capítulo 3 se presenta el modelo propuesto, es decir, introduce el modelo multivariado espacial blockNNGP y la inferencia bayesiana para el modelo utilizando la metodología de aproximación integrada anidada de Laplace (INLA). En el capítulo 4, se presenta un estudio de simulación con los diferentes escenarios para la generación de datos, con el fin de evaluar la precisión y eficiencia de los métodos propuestos.

En el capítulo 5, se presentará la aplicación de los modelos propuestos usando en datos reales. En el capítulo 6, se presentan las conclusiones y trabajos futuros.

Capítulo 2

Conceptos y modelos

Se debe tener en cuenta que los datos geoestadísticos son mediciones sobre un fenómeno espacialmente continuo que se han recopilado en sitios específicos. Este tipo de datos pueden representar, abundancia de especies en determinados locales, riesgos de enfermedades en una zona, nivel de contaminación o contaminantes en diferentes estaciones de monitoreo, entre otros.

2.1. Estadística espacial

Según lo descrito por Grekousis (2020), los lugares no están aislados unos de otros. Se producen interacciones sociales, económicas y demográficas entre lugares cercanos como distantes. Estas interacciones y relaciones tienen una dimensión espacial, ya que ocurren en el espacio, por lo que la ubicación y la distancia son relevantes. Debemos definir qué tan *cerca* o *lejos* debe estar un objeto de otro para considerarse como tal y así determinar cómo representar un objeto para calcular estas métricas de distancia, ya sea como un punto, línea o polígono. Al aplicar métodos para analizar datos espaciales, debemos definir matemáticamente lo cerca que está un objeto cercano y cómo se define la contigüidad. Esto implica establecer una serie de parámetros geográficos para definir las relaciones espaciales entre objetos. Este proceso se conoce como conceptualización de las relaciones espaciales, y representa es una diferencia importante entre los métodos aplicados a los datos espaciales y no espaciales.

2.1.1. Autocorrelación espacial

La aplicación de métodos estadísticos en un contexto espacial plantea varios desafíos. Tobler (1970) resumió un aspecto fundamental que afecta cualquier análisis de datos referenciados espacialmente en la primera ley de la geografía: "Todo está relacionado con todo, pero

las cosas cercanas están más relacionadas que las lejanas". Esta ley resume de manera concisa el concepto estadístico de autocorrelación espacial, que establece que las observaciones cercanas se asemejan más entre sí que las situadas a mayor distancia.

Grekousis (2020) indica que la autocorrelación espacial es el grado de dependencia, asociación o correlación espacial entre el valor de una observación de una entidad espacial y los valores de las observaciones vecinas de las misma variable. Los términos asociación espacial o dependencia espacial son utilizados frecuentemente como sinónimos de autocorrelación espacial.

El concepto de autocorrelación es análogo al de correlación estadística utilizado para variables no espaciales. No obstante, hay una diferencia significativa, mientras que la correlación estadística se refiere a dos variables distintas sin considerar la ubicación o localización, la autocorrelación espacial se refiere al valor de una única variable en una ubicación específica en relación con los valores de la misma variable en ubicaciones vecinas.

La autocorrelación se presenta como una variable atributo de un conjunto de datos espaciales y muestra correlación consigo misma a distancias específicas. Esto implica que la ubicación o localización tiene un impacto en los valores de la variable de tal manera que promueve la agrupación de valores en áreas específicas. Un ejemplo común de autocorrelación espacial es la distribución de ingresos dentro de una ciudad, donde los hogares con mayores ingresos tienden a agruparse en ciertas regiones, mientras que hogares con ingresos más bajos tienden a agruparse en otras regiones.

2.1.2. Datos geoestadísticos

El concepto fundamental subyacente geoestadística supone que los datos son una realización de un proceso aleatorio, denominado proceso estocástico o campo aleatorio:

$$\{Y(s) : s \in D\},$$

donde D es un subconjunto de $\Re^2$ y el índice espacial s varía continuamente en toda la región D. Según esto debemos tener en cuenta que para una localización fija s_i, $Y(s_i)$ es una variable aleatoria a la que aplicamos las leyes de probabilidad; para una conjunto de eventos fijos de este proceso, observamos una función del espacio, es decir, los datos de las localizaciones $s_1, s_2, \cdots, s_N$. Los datos son sólo una realización de un proceso espacial, ya que no podemos observar el proceso en todos los puntos de D.

2.1.3. Procesos espaciales

Un *proceso espacial puntual* se refiere a un proceso estocástico $\{Y(s) : s \in D\}$ en el cual cada variable aleatoria $Y(s)$ representa la ubicación de un evento en el espacio. Una *realización* del proceso consiste en una colección de ubicaciones generadas de acuerdo con el modelo de proceso espacial puntual. En otras palabras, una realización representa un conjunto de datos que resulta de un modelo específico (ya sea observado o simulado).

En particular, el proceso $Y(s)$ es gaussiano, si para $n \geq 1$ y un conjunto de sitios $(s_1, \cdots, s_n)$, cualquier realización del proceso, por ejemplo, $(Y(s_1), \cdots, Y(s_n))^T$ tiene una distribución normal multivariante.

Hay dos conceptos subyacentes que nos brindan un punto de partida para modelar procesos espaciales puntuales, el de *estacionariedad* e *isotropía*. Podemos decir que un *proceso estacionario* es invariante a la traslación dentro de un espacio d-dimensional. Mientras que, un *proceso isótropico* es invariante a la rotación alrededor del origen. Es decir, las relaciones entre dos eventos en un proceso estacionario dependen sólo de sus posiciones relativas, no de las propias ubicaciones de los sucesos. Añadir un supuesto de isotropía (tomando como base un proceso estacionario) implica una restricción adicional donde las relaciones dependen sólo de la distancia y no de la dirección.

Estacionariedad

Se dice que un proceso espacial es *fuertemente estacionario* (a veces estacionariedad *estricta*) si para cualquier vector de separación $\mathbf{h}$, la distribución $(Y(s_1), \cdots, Y(s_n)$ es la misma que la de $(Y(s_1 + h), \cdots, Y(s_n + h)$, entonces:

$$F_{Y(s_1),Y(s_2),\cdots,Y(s_n)}(y_1, y_2, \cdots, y_n) = F_{Y(s_1+\mathbf{h}),Y(s_2+\mathbf{h}),\cdots,Y(s_n+\mathbf{h})}(y_1 + \mathbf{h}, y_2 + \mathbf{h}, \cdots, y_n + \mathbf{h}).$$

Otra condición que podríamos tener es la considerada *estacionariedad débil* (también llamada estacionariedad de segundo orden), que dice que un proceso espacial es débilmente estacionario si para cualquier vector de separación $\mathbf{h}$, tiene un media constante y función de covarianza que solo depende del vector de separación $\mathbf{h}$, es decir:

$$E(Y(s_i)) = \mu, \tag{2.1}$$

$$Cov(Y(s_i), Y(s_i + \mathbf{h})) = C(\mathbf{h}), \forall s_i, s_i + \mathbf{h} \in D.$$

Según esta condición la relación de covarianza entre los valores del proceso en dos lugares

cualesquiera puede resumirse mediante una función de covarianza $C(\mathbf{h})$ (que es una medida de la autocorrelación espacial), y esta función depende sólo del vector de separación $\mathbf{h}$. Un proceso débilmente estacionario no es necesariamente fuertemente estacionario, sin embargo, si el proceso gaussiano si se cumple esta premisa.

Existe otro tipo de estacionariedad llamada *estacionariedad intrínsica*. Aquí asumimos que un proceso espacial $Y(s)$ es intrínsicamente estacionario si para cualquier vector de separación $\mathbf{h}$, tenemos que $E(Y(s_i + \mathbf{h}) - Y(s_i)) = 0$ y definimos:

$$E[Y(s_i + \mathbf{h} - Y(s_i)]^2 = V(Y(s_i + \mathbf{h}) - Y(s_i)) = 2\gamma(\mathbf{h}), \forall s_i, s_i + \mathbf{h} \in D. \tag{2.2}$$

La ecuación (2.2) solo tiene sentido si $E[Y(s_i + \mathbf{h} - Y(s_i)]^2 = 2\gamma(\mathbf{h})$ depende únicamente de $\mathbf{h}$ y no de la elección de s_i (si esto ocurre el proceso es intrínsicamente estacionario). La función $2\gamma(\mathbf{h})$ se denomina *variograma* y $\gamma(\mathbf{h})$, *semivariograma*.

Se debe tener en cuenta que la estacionariedad intrínseca no nos proporciona información sobre la distribución conjunto de una colección de variables $Y(s_i), \cdots, Y(s_n)$, y por lo tanto no proporciona ninguna probabilidad. Sólo nos proporciona una descripción del comportamiento de las diferencias en lugar del comportamiento de los datos que observamos.

Isotropía

Otro concepto importante relacionado es el de la *isotropía*. La estacionariedad y la isotropía son propiedades de invarianza, la estacionariedad es invariante bajo la traslación y la isotropia es invariante bajo rotaciones y orientaciones. Podemos decir que un proceso espacial $Y(s)$ es isotrópico si para cualquier distancia $\|h\|$:

$$E(Y(s_i)) = \mu; \quad \text{y}$$

$$Cov(Y(s_i), Y(s_i + \mathbf{h})) = C(\|h\|), \forall s_i, s_i + \mathbf{h} \in D.$$

Si la función de semivariograma $2\gamma(\mathbf{h}) = C(\|h\|)$ depende de la distancia $\|h\|$, entonces se dice que el variograma es isotrópico. En caso contrario, se dice que es anisotrópico. Según Banerjee et al. (2015) este tipo de variogramas isotrópicos son populares debido a su simplicidad, capacidad de interpretación y porque existen formas paramétricas relativamente simples disponibles como candidatos para el semivariograma.

2.1.4. Variograma

Existe una relación fácilmente observable entre el variograma y la función de covarianza:

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var(Y(s_i+\mathbf{h}) - Y(s_i)) \\ &= Var(Y(s_i+\mathbf{h})) + Var(Y(s_i)) - 2Cov(Y(s_i+\mathbf{h}), Y(s_i)) \\ &= C(\mathbf{0}) + C(\mathbf{0}) - 2C(\mathbf{h}) \\ &= 2[C(\mathbf{0}) - C(\mathbf{h})]. \end{aligned}$$

Entonces,

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}). \tag{2.3}$$

Si el proceso espacial es ergódico (las propiedades estadísticas de un proceso son representativas del comportamiento estadístico promedio del proceso a lo largo del tiempo), entonces asumimos que $C(\mathbf{h}) \to 0$ mientras $\|\mathbf{h}\| \to \infty$, donde $\|\mathbf{h}\|$ denota de la longitud del vector $\mathbf{h}$, esto nos indica que la relación entre los valores de dos puntos disminuye a medida que los puntos se alejan más en el espacio. Si consideramos el límite de la ecuación (2.3) cuando $\mathbf{h}$ tiende a infinito, obtenemos que $lim\|\mathbf{h}\| \to \infty$ y la covarianza se acerca al valor constante $C(\mathbf{0})$.

Según lo descrito por Banerjee et al. (2015) y Waller y Gotway (2004) un proceso débilmente estacionario implica que un proceso sea intrínsicamente estacionario, pero lo contrario no es verdadero. Asimismo, es importante destacar que la definición de estacionariedad intrínseca es muy similar a la estacionaridad de segundo orden, donde la primera se define en términos del variograma y la segunda en términos de la función de covarianza. De hecho, el variograma es una generalización del la función de covarianza y, bajo el supuesto de estacionariedad de segundo orden, estas dos funciones están relacionadas.

Según los descrito por Waller y Gotway (2004) el semivariograma representado por la función $\gamma(\mathbf{h})$, es una parte fundamental en la geoestadística. Aunque los términos de *variograma* y *semivariograma* a menudo se utilizan indistintamente, hay una diferencia clara entre ellos: el variograma es el doble del semivariograma. Esta distinción puede no ser relevante en algunos cálculos, pero en otros casos puede ser crucial. Se distinguirá entre variograma y semivariograma, utilizando únicamente este último término.

El semivariograma es una función del proceso espacial y cumple ciertas propiedades:

I. $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$ es decir, la autocorrelación entre $Y(s_i)$ y $Y(u_i)$ es la misma que entre $Y(u_i)$ y $Y(s_i)$.

II. $\gamma(\mathbf{0}) = 0$, ya que por definición, $Var(Y(s_i) - Y(s_i)) = 0$.

III. $\gamma(\mathbf{h})/\|\mathbf{h}\|^2 \to 0$ donde $\|\mathbf{h}\|$ denota la longitud del vector $\mathbf{h}$.

IV. $\gamma(\cdot)$ debe ser condicionalmente negativa, es decir,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \gamma(s_i - s_j) \leq 0,$$

para cualquier número finito de ubicaciones $\{s_i : i = 1, \cdots, n\}$ y números reales $\{a_i, \cdots, a_n\}$ que satisfagan $\sum_{i=1}^{n} a_i = 0$. (Análogo a la condición positiva-definida de las funciones de covarianza y matrices de varianza-covarianza).

V. Si el proceso es isótropico, $\gamma(\mathbf{h}) \equiv \gamma(h)$, donde $h = \|\mathbf{h}\|$ (es decir, el semivariograma es únicamente función de la distancia).

Si realizamos un gráfico de $\gamma(\mathbf{h})$ en función de la distancia de separación $\|\mathbf{h}\|$, nos proporcionará información sobre la continuidad y variabilidad espacial del proceso. Adicionalmente, podríamos decir que a valores de pequeños de $\|\mathbf{h}\|$ (distancias cortas), esperamos que $Y(s_i+\mathbf{h})$ y $Y(\mathbf{h})$ fueran similares; y a medida que aumenta $\|\mathbf{h}\|$ se espera una menor simulitud entre $Y(s_i + \mathbf{h})$ y $Y(\mathbf{h})$, es decir, esperamos un $[Y(s_i + \mathbf{h}) - Y(s_i)]^2$ sea mayor. Este gráfico comienza en cero y, si las observaciones cercanas son más similares que que las alejadas, el semivariograma aumenta a medida que aumenta la distancia de separación. Este aumento en la variación de diferencias entre pares refleja la disminución de la autocorrelación espacial, ya que las observaciones $Y(s_i)$ y $Y(u_i)$ pueden variar más entre sí a medida que las ubicaciones s y u se alejan. Por lo general, el semivariograma se nivela hasta casi un valor constante, denominado umbral o meseta (*sill*), a una distancia de separación grande conocida como rango o alcance (*range*, r) que es el valor donde $\gamma(\mathbf{h})$ alcanza por primera vez último nivel (*sill*). Esta es la distancia a partir de la cual la autocorrelación especial entre observaciones se estabiliza, el rango puede interpretarse como la distancia a partir de la cual se puede asumir que la autocorrelación espacial no es tan fuerte. Más allá de la distancia, las observaciones no están correlacionadas espacialmente, lo que se refleja en una varianza (casi) constante de las diferencias entre pares. Como se ilustra en la figura 2.1.

Es importante tener en cuenta que si el proceso no es isotrópico, el semivariograma y la información que proporciona puede variar según la dirección, generando diferentes variogramas, uno por cada dirección. Si no hay autocorrelación entre $Y(s_i)$ y $Y(u_i)$, el semivarigrama será una línea horizontal.

La forma del semivariograma cerca del origen es relevante, ya que indica la suavidad o continuidad espacial de la variable estudiada. Una forma parabólica cerca del origen indica

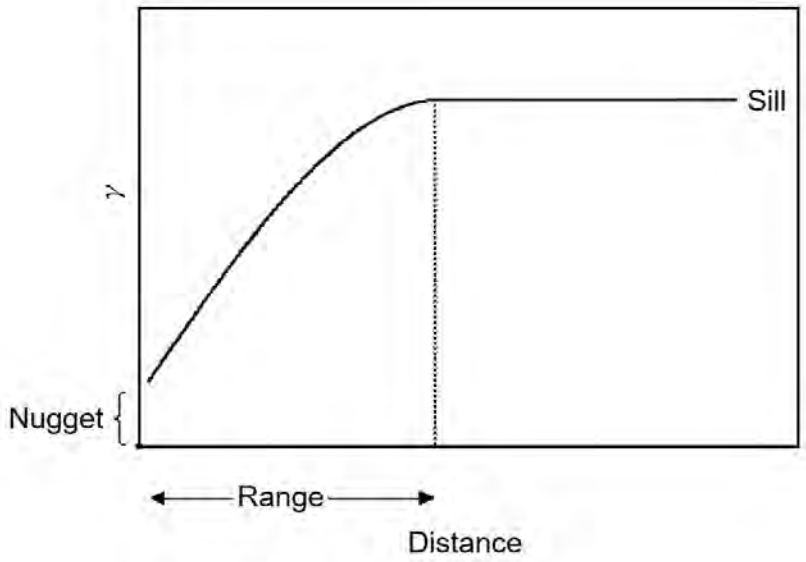

Figura 2.1: Semivariograma típico

una variable espacial suave, continua y diferenciable. Una forma lineal cerca del origen indica una variable continua pero no diferenciable, lo que implica que es menos regular. Una discontinuidad o salto vertical en el origen (cuando $lim_{h\to 0}\gamma(\mathbf{h}) = c_0 > 0$) indica que la variable espacial no es continua y presenta una variabilidad espacial muy irregular. Esta discontinuidad en geoestadística se conoce como efecto pepita (*nugget*, τ^2). Un gran efecto pepita implica que dos observaciones cercanas pueden tener valores muy diferentes, lo cual puede ser debido a errores de medición o a una discontinuidad espacial en el proceso (variabilidad no estructurada de los datos). Cuando existe efecto pepita, el umbral o meseta parcial (*partial sill*, σ^2) se define como la diferencia entre la varianza del proceso (*sill*) y el efecto pepita, es decir, $C(\mathbf{0}) - c_0$ (varianza marginal del proceso espacial). Este efecto se considera una discontinuidad y, por definición en la ecuación (2.2) el semivariograma en el origen es siempre cero (es decir, una distancia de separación cero).

Semivariogramas teóricos

Según Banerjee et al. (2015) y Waller y Gotway (2004), los semivariogramas isotrópicos gozan de popularidad debido a su simplicidad, facilidad de interpretación y, sobre todo, porque existen varias formas paramétricas relativamente simples que se pueden utilizar como opciones para el semivariograma. Denotando $\|h\|$ como h para simplicar la notación y sea ϕ un parámetro de decaimiento asociado al rango r, estas serían algunas de las funciones más conocidas se definen a continuación:

1. Lineal:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 h & \text{si } h > 0, \tau^2 > 0, \sigma^2 > 0, \\ 0 & \text{caso contrario.} \end{cases}$$

Nótese que $\gamma(h) \to \infty$ a medida que $h \to \infty$, por lo que este semivariograma no corresponde a un proceso débilmente estacionario (aunque es intrínsecamente estacionario). En este caso el *nugget* es τ^2 pero los valores de *sill* y *range* son ambos infinitos. Para otros variogramas (como los siguientes), el umbral es finito, pero sólo se alcanza asintóticamente.

2. Esférica:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 & \text{si } h \geq 1/\phi, \\ \tau^2 + \sigma^2\{\frac{3\phi h}{2} - \frac{1}{2}(\phi h)^3\} & \text{si } 0 < d \leq 1/\phi, \\ 0 & \text{caso contrario.} \end{cases}$$

La razón principal de la popularidad de este semivariograma es su capacidad de presentar de manera clara y visual los valores del *nugget*, *sill* y el *range*. Es valido en $r = 1, 2$ o 3 dimensiones, pero para $r > 4$ no cumple la condición de ser una matriz de varianza espacial definida positiva, que es necesaria para especificar una distribución de probabilidad conjunta válida.

3. Exponencial:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - exp(-\phi h)) & \text{si } h > 0 \\ 0 & \text{caso contrario.} \end{cases}$$

El semivariograma exponencial sube más lentamente desde el origen que el semivariograma esférico. Sin embargo, este modelo se aproxima asintóticamente al umbral. Es común utilizar el concepto de alcance efectivo (*effective range*), que nos indica la distancia a la cual no existe una correlación espacial persistente. Para definir este concepto, es necesario realizar una conversión de escala desde γ a C (la cual es posible en este caso, ya que que existe el $lim_{h\to\infty}\gamma(h)$, la función exponencial no solo es intrínsecamente estacionaria, sino también débilmente estacionaria). El semivariograma exponencial se relaciona con la función de covarianza exponencial:

$$C(t) = \begin{cases} \tau^2 + \sigma^2 & \text{si } h = 0, \\ \sigma^2 exp(-\phi h) & \text{si } h > 0. \end{cases} \tag{2.4}$$

Es común definir el alcance efectivo (h_0), como la distancia a la cual la correlación se vuelve insignificante, generalmente considerada cuando cae por debajo del 0.05. Si igualamos $exp(-\phi h_0)$ a este valor, obtenemos $t_0 \approx 3/\phi$, ya que $log(0.05) \approx -3$. La forma de la ecuación (2.4) nos proporciona una idea de por qué el componente pepita (τ^2) a menudo se interpreta como una "varianza de efecto no espacial", mientras que el umbral parcial (σ^2) se interpreta como una "varianza de efecto espacial".

4. Gaussiano:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - exp(-\phi^2 h^2)) & \text{si } h > 0, \\ 0 \quad \text{caso contrario.} \end{cases}$$

Este modelo es parábolico cerca del origen, lo cual nos indica que es un proceso espacial muy suave. En la práctica los procesos que siguen este tipo de modelo son poco frecuentes, aunque el modelo gaussiano es considerado generalmente el mejor según los criterios automáticos de ajuste de modelos. A pesar que es un modelo semivariográfico válido, su uso a menudo puede conducir a casos especiales en las ecuaciones de predicción espacial. Según Waller y Gotway (2004) este modelo corresponde a un proceso determinista y, por lo tanto, contradice el supuesto de aleatoriedad subyacente en geoestadística y recomienda usar este modelo con precaución y solo cuando se disponga de una gran cantidad de datos cercanos entre sí para evaluar su comportamiento cerca del origen.

5. Powered exponencial:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - exp(-|\phi^2|^p)) & \text{si } h > 0, \\ 0 \quad \text{caso contrario.} \end{cases}$$

Para $0 < p \leq 2$ da lugar a una familia de variogramas válidos. Las formas gaussiana y exponencial son los miembros frecuentes de esta clase.

6. Racional cuadrático:

$$\gamma(h) = \begin{cases} \tau^2 + \frac{\sigma^2 h^2}{(\phi + h^2)} & \text{si } h > 0, \\ 0 \quad \text{caso contrario.} \end{cases}$$

7. Wave:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2(1 - \frac{sin(\phi h)}{\phi h}) & \text{si } h > 0, \\ 0 \quad \text{caso contrario.} \end{cases}$$

Este modelo es un ejemplo de semivariograma que no es monotónicamente creciente.

8. Matérn:

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2[\ 1 - \frac{(2\sqrt{\nu}h\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu}h\phi)] & \text{si } h > 0, \\ 0 \quad \text{si } d = 0. \end{cases} \tag{2.5}$$

Esta familia de modelos también es denominado modelo *K-Bessel* en geoestadística, debido a su dependencia de la función de Bessel K_ν. El modelo original fue propuesto por Matérn (1960), posteriormente Handcock y Stein (1993) y Handcock y Wallis (1994) presentaron interpretaciones atractivas para ν como para ϕ. Específicamente, $\nu > 0$ es un parámetro que controla la suavidad del campo aleatorio realizado, mientas que ϕ es un parámetro de decaimiento espacial. La función $\Gamma(\cdot)$ se refiere a la función gamma común, y K_ν es la función de Bessel modificada del orden ν. Casos especiales de este modelo son la función exponencial $\nu = 1/2$ y la función gaussiana ($\nu \to \infty$). La función de covarianza de Matérn a menudo se reparametriza utilizando $\alpha = 2\sqrt{2}\phi$, junto con $\eta = \sigma^2\phi^{2\nu}$ y ν. Una ventaja de esta familia de modelos es que el comportamiento del semivariograma cerca del origen puede estimarse a partir de los datos en lugar de suponer una forma específica. Sin embargo, el cálculo de K_ν necesario para esta estimación puede ser engorrosa y, al igual que con el modelo gaussiano, requiere tener algunos datos muy espaciados.

Gráficos de estas funciones teóricas del semivariograma se muestran en la Figura 2.2.

2.1.5. Semivariograma empírico

Un modelo de variograma se selecciona trazando el *semivariograma empírico* (Matheron (1963), que es una estimación no paramétrica sencilla del semivariograma, y luego comparándolo con las diferentes formas teóricas disponibles según las opciones mencionadas en la sección 2.1.4. El semivariograma puede ser fácilmente estimado bajo el supuesto de estacionariedad intrínseca de forma que se cumplan las ecuaciones (2.1) y (2.2). Se puede definir el variograma como

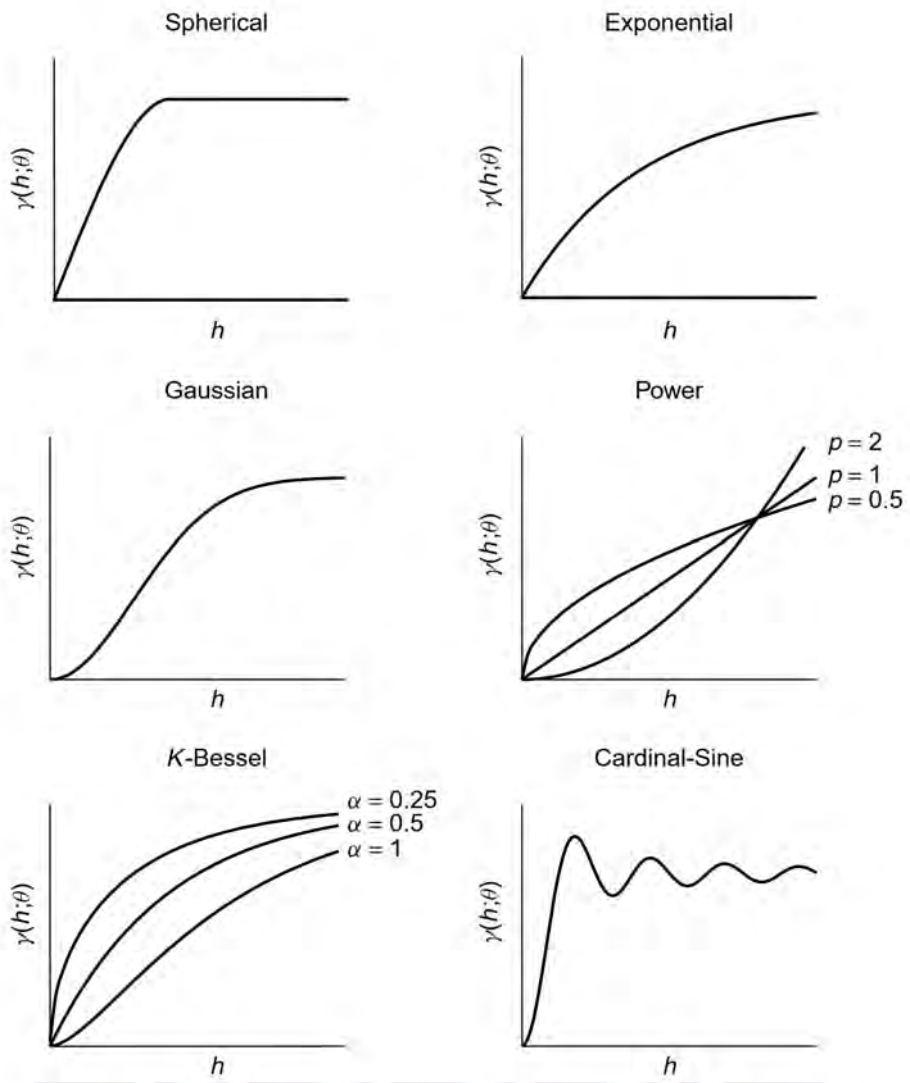

Figura 2.2: Semivariogramas teóricos: a) Esférica, b) Exponencial, c) Gaussiano, d) Power law, d) K-Bessel (Matérn), d) Cardinal-Sine (Wave). Fuente: Waller y Gotway (2004).

$$\begin{aligned} 2\gamma(\mathbf{h}) &= Var(Y(s_i+\mathbf{h}) - Y(s_i)) \\ &= E[(Y(s_i+\mathbf{h}) - Y(s_i))^2] - [E(Y(s_i+\mathbf{h}) - Y(s_i))]^2. \end{aligned}$$

De la ecuación (2.1), $E[Y(s_i)] = \mu$ para todo i, por lo que el segundo término es cero. Así para estimar el semivariograma sólo necesitamos estimar $E[(Y(s_i+\mathbf{h}) - Y(s_i))^2]$. Dado que las expectativas son sólo promedios estadísticos, una forma de estimar este término es promediar todas las diferencias al cuadrado $[Y(s_i) - Y(s_j)]^2$ para pares de observaciones tomadas a la misma distancia en las misma dirección (es decir, para todos los s_i, s_j, tales que $s_i - s_j = \mathbf{h}$). Esta estimación esta esencialmente basada en el método de los momentos (MOM), su análogo en la estimación convencional de la varianza muestral (s_2) y esta dado

por:

$$\hat{\gamma}(h) = \frac{1}{2N(h)} \sum_{(s_i,s_j)\in N(h)} [Y(s_i) - Y(s_j)]^2 \tag{2.6}$$

donde $N(h)$ es el conjunto de pares de puntos tales que $\|s_i - s_j\| = h$ y $|N(h)|$ es el número de pares de este conjunto. Hay que tener en cuenta que, a menos que las observaciones estén distribuidas en una cuadrícula regular, las distancias entre los pares de puntos serán todas diferentes, lo que hace que la estimación directa del semivariograma no sea útil tal cual. En cambio, se puede dividir el espacio en intervalos $I_1 = (0, h_1)$, $I_2 = (h_1, h_2)$ así sucesivamente hasta $I_K = (h_{K-1}, h_K)$ utilizando una cuadrícula generalmente regular, donde $0 < h_1 < \cdots < h_K$. Si representamos los valores h de cada intervalo utilizando el punto medio del intervalo, podemos modificar la definición de $N(h)$ a:

$$N(h_k) = \{(s_i, s_j) : \|s_i - s_j\| \in I_k\}, \quad k = 1, \cdots, K.$$

Cada intervalo debe ser lo suficientemente pequeño para que conservemos una resolución espacial suficiente para definir la estructura del semivariograma, pero también lo suficientemente amplios para asegurar que al menos 30 pares de puntos en cada intervalo (Banerjee et al., 2015).

En resumen, el proceso de estimación del semivariograma implica trazar una estimación empírica, visualizarla y ajustarla a un modelo teórico que se ajuste a los datos. Sin embargo, debido al ruido presente en la estimación empírica, en un escenario de datos reales, varios modelos diferentes (mecionados en 2.1.4) pueden parecer apropiados. Normalmente, el ajuste se ha realizado de forma subjetiva o mediante el método de prueba y error, seleccionando los valores de los parámetros de pepita, umbral y rango que nos den una buena correspondencia con el semivariograma empírico. La evaluación de "buena correspondencia"puede realizarse visualmente o utilizando algún criterio de ajuste, como mínimos cuadrados u otros. De manera más formal, este proceso se puede tratar como un problema de estimación estadística, utilizando rutinas de maximización no lineal para encontrar los valores de los parámetros que minimicen algún criterio de ajuste.

Si se tiene un modelo de distribución para los datos, se puede utilizar el método de máxima verosimilitud (o máxima verosimilitud restringida, REML) para obtener estimaciones confiables de los parámetros. Sin embargo, es más conveniente e intuitivo trabajar directamente con el modelo de la covarianza $C(h)$, al utilizar este enfoque, también podemos obtener una inferencia completa, que incluye la obtención de distribuciones posteriores para todas las

incógnitas relevantes. En conclusión, el semivariograma es una herramienta útil para explorar la autocorrelación espacial.

2.2. Medidas de distancia

Según lo descrito anteriormente la estadística espacial se basa en la idea de que los atributos medidos en elementos cercanos tienden a ser más similares que aquellos más lejanos. Por lo tanto, es esencial contar con descripciones matemáticas de la cercanía y la lejanía para cuantificar y utilizar este concepto en el análisis estadístico. Una forma simple de describir esto es a través de la distancia entre dos características. En el análisis de variogramas, se requieren cálculos precisos de distancias entre ubicaciones para evaluar la intensidad de la asociación espacial. Según Grekousis (2020), Banerjee et al. (2015), Waller y Gotway (2004) existen múltiples métodos para medir la distancia, se describe algunas medidas de distancia para cuantificar el grado de proximidad entre dos características especiales:

1. *Distancia geodésica*: para conjuntos de datos que abarcan áreas espaciales relativamente pequeñas, la distancia euclidiana ordinaria es una aproximación adecuada. Sin embargo, cuando se trata de dominios más grandes, como todo un territorio continental, la curvatura de la tierra introduce distorsiones debido a las diferencias en los incrementos de longitud y latitud. Supongamos que estamos utilizando el sistema de coordenadas longitud/latitud para representar lugares en la superficie de la Tierra, y tenemos dos lugares, $s_1 = (\lambda_1, \phi_1)$ y $s_2 = (\lambda_2, \phi_2)$, donde λ representa la longitud y ϕ la latitud. La distancia más corta entre estas dos ubicaciones a lo largo de la superficie de la Tierra esférica se calcula mediante la fórmula de la distancia esférica, dado por:

$$d(s_1, s_2) = (6378) \cdot \arccos\left[\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \cos\left(\lambda_1 - \lambda_2\right)\right], \tag{2.7}$$

   donde 6378 kilómetros es el radio de la Tierra (esférica).

2. *Distancia euclidiana*: es la distancia entre dos puntos A y B conectados por una línea recta. Supongamos que trabajamos con un sistema de coordenadas proyectadas y tenemos dos ubicaciones $s_1 = (u_1, v_1)$ y $s_2 = (u_2, v_2)$, en un plano bidimensional. Entonces la distancia más corta entre estas dos ubicaciones en un mapa plano viene dato por:

$$d(s_1, s_2) = \sqrt{(u_2 - u_1)^2 + (v_2 - v_1)^2}. \tag{2.8}$$

   Usando la notación de vectores, esta distancia puede ser denotada como $\|s_2 - s_1\|$. Esta

es conocida como la norma euclidiana, y la medida de distancia en la ecuación se llama distancia euclidiana. Si utilizamos las coordenadas de longitud y latitud, $s_1 = (\lambda_1, \phi_1)$ y $s_2 = (\lambda_2, \phi_2)$, la distancia resultante no tomará en cuenta la curvatura de la Tierra. En general, la medida de distancia euclidiana no debería ser utilizada para calcular distancias entre conjuntos de coordenadas de longitud y latitud, especialmente si las distancias abarcan una extensa área.

3. *Distancia Manhattan o City-Block*: es la diferencia vertical más horizontal (medida a lo largo de los ejes) entre los puntos A y B. Si para medir la distancia no podemos medirlo a través de una línea recta y en su lugar necesitamos seguir una serie de segmentos perpendiculares, surge la idea de la distancia *City-Block* entre dos lugares, $s_1 = (u_1, v_1)$ y $s_2 = (u_2, v_2)$:

$$d(s_1, s_2) = |(u_2 - u_1)| + |(v_2 - v_1)|. \tag{2.9}$$

4. *Distancia de Minkowski*: es una generalización de las distancias euclidiana y de Manhattan:

$$d(s_1, s_2) = ((u_2 - u_1)^p + (v_2 - v_1)^p)^{1/p}. \tag{2.10}$$

Para $p = 1$ obtenemos la distancia de Manhattan y para $p = 2$, la distancia euclidiana.

Banerjee et al. (2015) considera algunos enfoques diferentes para calcular distancias en la Tierra utilizando métricas euclidianas, clasificándolas como las que surgen de las coordenadas esféricas clásicas y las que surgen de proyecciones planas (como por ejemplo *chord*, *naive Euclidean*).

Waller y Gotway (2004) indica que en muchas aplicaciones, sobre todo en las ciencias ambientales, la medida de distancia euclidiana puede no ser realista (por ejemplo, las montañas, los dominios de forma irregular, etc., presentan barreras en la medición de distancias y pueden causar ruido). Considerando que dos puntos situados a ambos lados de una barrera pueden estar físicamente cerca, puede ser poco realista suponer que están relacionados. Plantean que los análisis geoestadísticos pueden realizarse tambien utilizando distancia no euclidianas (por ejemplo, distancia *City-Block* siempre que se cumplan dos condiciones: 1) la medida de distancia es una métrica válida en $\Re^2$ (es decir, debe no ser negativa, simétrica y satisfacer la desigualdad del triángulo); 2) el semivariograma utilizado con esta métrica debe satisfacer propiedad de un semivariograma.

Aunque el cálculo de distancia geodésica (ecuación (2.7)) es sencillo, las métricas euclidia-

nas son populares por su simplicidad y facilidad de interpretación. Además, la modelización estadística de las correlaciones espaciales se basa a menudo en funciones de correlación que sólo son válidas con métricas euclidianas.

2.3. Inferencia bayesiana

La inferencia clásica busca infererir sobre el parámetros o vector de parámetros θ de una población a través de una muestra representativa $\boldsymbol{y}$. En el caso de la inferencia bayesiana, se introduce un conocimiento previo de los parámetros $\boldsymbol{\theta}$, es necesario tener presente los siguientes conceptos:

- **Probabilidad Condicional**: mide el grado de factibilidad de la ocurrencia de un evento A si se conoce que el evento B ya ocurrió. Sea B un evento $P(B) > 0$. La probabilidad condicional de un evento A dato el evento B esta dado por:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.11}$$

- **Teorema de Bayes**: Sean A y B dos eventos cualesquiera

$$P(A|B) = \frac{P(A|B)P(A)}{P(B)}, \forall j = 1, 2, \cdots, n \tag{2.12}$$

- **Distribución a priori**: esta distribución representa el conocimiento previo (puede ser conocimiento externo, opiniones de expertos, etc.) que tenemos sobre un parámetro $\boldsymbol{\theta}$ antes de realizar un experimento. Se denota por $p(\boldsymbol{\theta})$.

- **Verosimilitud**: representa como creemos que se comportan los datos $\boldsymbol{y}$ si conocemos $\boldsymbol{\theta}$. Se define a través de la distribución muestral $p(\boldsymbol{y}|\boldsymbol{\theta})$.

- **Distribución a posteriori**: es la distribución de $\boldsymbol{\theta}$ que resume la información previa sobre $\boldsymbol{\theta}$ y la que se obtiene por los datos. Se denota por $p(\boldsymbol{\theta}|\mathbf{y})$.

Sea el conjunto de datos observados $\boldsymbol{y}$, nuestro conocimiento sobre $\boldsymbol{\theta}$ es actualizado, mediante el teorema de Bayes obteniendo:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})}, \tag{2.13}$$

donde $p(\boldsymbol{y}) = \int p(\boldsymbol{\theta})p(\boldsymbol{y}|\boldsymbol{\theta})d\theta$. Como $p(\boldsymbol{y})$ no depende de $\boldsymbol{\theta}$, podemos escribir como:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.14}$$

La principal ventaja de la inferencia bayesiana es su capacidad de obtener una distribución a posteriori, lo cual es útil en modelos complejos con varios parámetros como los jerárquicos o anidados (Gelman et al., 2014).

2.3.1. Modelos jerárquicos

Al utilizar el enfoque bayesiano en el análisis estadístico, se considera tantos los datos observados como las incógnitas como variables aleatorias. Esto nos permite combinar modelos complejos de datos con conocimientos externos u opiniones de expertos. Este enfoque, especifica el modelo de distribución $f(\boldsymbol{y}|\boldsymbol{\theta})$ para los datos observados $\boldsymbol{y} = (y_1, \cdots, y_n)$ dado un conjunto de parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, y se asume que $\boldsymbol{\theta}$ es una cantidad aleatoria que se obtiene mediante muestro de una distribución a priori $\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})$, donde $\boldsymbol{\lambda}$ es un conjunto de hiperparámetros. Si conocemos $\boldsymbol{\lambda}$, la inferencia sobre $\boldsymbol{\theta}$ se basa en su distribución a posteriori,

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{\lambda}) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{p(\boldsymbol{y}|\boldsymbol{\lambda})} = \frac{p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})}{\int p(\boldsymbol{y}, \boldsymbol{\theta}|\boldsymbol{\lambda})d\theta} = \frac{f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})d\theta}. \tag{2.15}$$

Se puede observar la contribución tanto de los datos (representados por la verosimilitud f) como del conocimiento u opinión externa (representado por la *a priori*) π para la obtención de la distribución a posteriori. En la práctica, dado que $\boldsymbol{\lambda}$ no será conocida, se requerirá frecuentemente una distribución de segundo nivel (o *hiperprior*) $h(\boldsymbol{\lambda})$, luego la ecuación (2.15) se sustituirá por:

$$p(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})d\lambda}{\int f(\boldsymbol{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})d\theta d\lambda}. \tag{2.16}$$

Waller y Gotway (2004) nos indica que la estructura jerárquica en la ecuación (2.16) implica tres niveles de especificación distribucional, siendo el nivel $\boldsymbol{\theta}$ el mayor interés:

1. Un vector de efectos fijos que relaciona las covariables con el resultado esperado Y.
2. Un vector de efectos aleatorios basado normalmente e un campo aleatorio gaussiano o una distribución gaussiana multivariante con correlación espacial.
3. Un vector de parámetros que define las correlaciones espaciales (matriz de varianza-covarianza) de los efectos aleatorios.

2.3.2. Aproximación de Laplace anidada Integrada (INLA)

En la mayoría de casos, los modelos geoestadísticos se pueden describir como modelos de regresión aditiva estructurada. Uno de los subconjunto más populares, son los modelos gaussianos latentes, donde el campo latente es gaussiano y controlado por unos pocos hiperparámetros, mientras que las variables de respuesta son gaussianas o no gaussianas. Debido a la naturaleza del modelo complejo las marginales a posteriori no son fácilmente estimables. Para abordar este tipo de modelos, se pueden aplicar métodos de Monte Carlo con cadenas de Markov (MCMC), aunque estos métodos no están exentos de problemas, sobre todo en términos de convergencia y tiempo requerido para la estimación de los parámetros. Ante ello Rue et al. (2009) y una revisión actualizada Rue et al. (2017) plantea la utilización de una aproximación de Laplace anidada integrada (INLA, *Integrated nested Laplace approximation*), método que permite realizar inferencia bayesiana aproximada para la clase de modelos gaussianos latentes (MGL) y nos permite calcular directamente aproximaciones precisas de las distribuciones marginales a posteriori.

Los tres compenentes principales del INLA son:

1. **Modelos gaussianos latentes (MGL)**, es un modelo jerárquico esta compuesto por los siguientes niveles:

   a) *Vector aleatorio* ($\boldsymbol{Y}$): Sea $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^\top$ donde las variables aleatorias Y_i condicionadas a $\boldsymbol{X}, \boldsymbol{\theta}$ son independientes, entonces la función de probabilidad o densidad de $\boldsymbol{Y}$ es:

   $$\pi(y|x,\theta) = \prod_{n=1}^{n} \pi(y_i|x_i,\theta).$$

   De manera general, si $Y_i \sim\ FD(\mu_i, \delta)$ donde μ_i es la media y δ es algún parámetro de escala o dispersión, entonces se puede usar una función de enlace $g(.)$ para asociar μ_i con el predictor lineal η_i, de la siguiente forma:

   $$g(E(Y)) = \eta_i = \alpha + \sum_{j=1}^{n_f} f^{(j)}(u_{ji}) + \sum_{k=1}^{n_\beta} \beta_k z_{ki}; \tag{2.17}$$

   donde:

   - α: intercepto, efecto "fijo"
   - z_{ki}: covariables
   - β_k: coeficientes de regresión, efectos "fijos".
   - $f^{(j)}(\cdot)$: funciones desconocidas de las covariables, efectos estructurados.

b) *Campo gaussiano latente (**X**)*: En este nivel se incluyen los efectos aleatorios a los cuales se les asigna distribución normal. Se asume que $\boldsymbol{X} = (X_1, \cdots, X_j)$ condicionado a $\boldsymbol{\theta}$ tiene una distribución normal $N(\mu, \Sigma(\theta))$. En particular, en la ecuación (2.17), se tiene que $X = \{\alpha, \beta_k, f^{(j)}\}$ el el vector de todas las variables latentes gaussianas.

c) *Hiperparámetros (**θ**)*: Sea $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p \in \Theta)$, el vector de hiperparámetros (está compuesto por los parámetros desconocidos restantes, que no son necesariamente gaussianos), tal que la densidad de θ es: $\pi(\theta)$.

2. **Campos Aleatorio markovianos gaussiano (CAMG)**: un vector aleatorio $\boldsymbol{X} = (X_1, \cdots, X_J)$ es un CAMG (GMRF, *Gaussian Markov random field*) si y solamente si, tiene distribución normal $N(\boldsymbol{\mu}, \boldsymbol{Q})$ con media μ y matriz de precisión Q (definida positiva), tal que la función de densidad de conjunta de $X|\theta$ está dada por:

$$\pi_{\boldsymbol{X}|\boldsymbol{\theta}}(x) = (2\pi)^{-n/2}|Q(\boldsymbol{\theta})|^{1/2} \exp\{\frac{-1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{Q}(\boldsymbol{\theta})(\boldsymbol{x} - \boldsymbol{\mu})\},$$

donde la matriz de precisión es definida como la inversa de la matriz de covarianza, $\boldsymbol{Q}(\boldsymbol{\theta}) = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta})$ y $\boldsymbol{Q}$ es una matriz dispersa (*sparse*), es decir, tiene muchos elementos iguales a cero.

El enfoque principal se centra en la marginal a posteriori del campo latente y de los hiperparámetros. INLA ofrece una aproximación de estas densidades marginales a posteriori, las cuales pueden ser utilizadas para calcular estadísticas de resumen aproximadas de interés, como medias, varianzas o cuantiles a posteriori.

Según Martino y Riebler (2019) el INLA se puede aplicar a los modelos gaussianos latentes (MGL) que cumplan los siguientes supuestos:

a) Cada variable respuesta depende sólo de uno de los elementos de campo gaussiano latente $\boldsymbol{X}$.

b) El tamaño del vector de hiperparámetros $\boldsymbol{\theta}$ es pequeño (<15).

c) El campo latente $\boldsymbol{X}$, puede ser grande, pero está dotado de algunas propiedades de independencia condicional (Markov), de modo que la matriz de precisión $\boldsymbol{Q}(\boldsymbol{\theta})$ sea dispersa.

d) El interés inferencial reside en las marginales a posteriori $\pi(x_i|y)$ y $\pi(\theta_j|y)$.

3. **Aproximaciones de Laplace**: de los componentes anteriores tenemos el vector aleatorio, campo aleatorio markoviano gaussiano y los hiperparámetros. Con ello podemos

calcular la función de probabilidad o densidad (fdp) conjunta a posteriori:

$$\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta}) \prod_{i=1}^{n} \pi(y_i|\boldsymbol{x}, \boldsymbol{\theta}),$$

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

Como la fdp conjunta generalmente no es fdp de una distribución conocida, se puede emplear métodos de simulación para obtener muestras de la distribución a posteriori (mediante el uso de métodos como el MCMC u otros alternativos).

El INLA propone calcular aproximaciones de la marginales a posteriori:

$$\pi(x_j|y) = \int \pi(x_j|\theta, y)\pi(\theta|y)d\theta,$$

$$\pi(\theta_p|y) = \int \pi(\theta|y)d\theta_p,$$

para $j = 1, \cdots, J$ y $p = 1, \cdots, P$.

A partir de estas fdp, es posible obtener estimaciones puntuales e intervalos de los parámetros de interés. Sin embargo, En muchas ocasiones las integrales mencionadas no se pueden calcular de forma analítica, por lo tanto INLA propone calcular estas integrales mediante aproximaciones numéricas.

Pasos para la aproximación INLA:

a) Aproximar la marginal a posteriori de $\boldsymbol{\theta}$: $\pi(\boldsymbol{\theta}|\boldsymbol{y}) \rightarrow \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$.

La función de densidad a posteriori de $\boldsymbol{\theta}|\boldsymbol{y}$ puede obtenerse a través de:

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

donde el denominador corresponde a la distribución condicional completa de $\boldsymbol{x}$.

Es importante destacar que las fdp del numerador son conocidas (son componentes del MGL) y tienen distribuciones conocidas y se pueden obtener muestras de ellas. Sin embargo, la fdp del denominador (condicional completa de $\boldsymbol{x}$) no siempre tiene una distribución conocida de la cual se puede obtener muestras. Por esta razón, INLA propone utilizar una aproximación gaussiana para aproximar la marginal

de $\boldsymbol{\theta}$ de la siguiente manera:

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

donde $x^* = argmax_x(\pi|\boldsymbol{\theta},\boldsymbol{y})$ es la moda de $\pi(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{y})$ que puede ser obtenida a través de algún método de optimización. La aproximación en la ecuación $(3a)$ es una aproximación de Laplace. A partir de esta aproximación tenemos valores de $\boldsymbol{\theta}$ con mayor masa de probabilidad.

b) Explorar $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ y usar integración numérica para aproximar la fdp marginal de θ_p:

$$\pi(\theta_p|\boldsymbol{y}) \approx \int_{\Theta_p} \tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_k \tilde{\pi}(\theta^{(pk)}|\boldsymbol{y})w_{pk}$$

En particular para calcular tales marginales necesitamos explorar bien la distribución $\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ y seleccionar valores más probables de $\boldsymbol{\theta}$ para la integración numérica.

c) Aproximar la función de densidad $\pi(x_j|\boldsymbol{\theta},\boldsymbol{y}) \to \tilde{\pi}(x_j|\boldsymbol{\theta},\boldsymbol{y})$.

La aproximación de la distribución marginal $\pi(x_j|\boldsymbol{\theta},\boldsymbol{y})$ puede ser obtenida mediantes: aproximación gaussiana, aproximación de Laplace o aproximación de Laplace simplificada. Siendo la aproximación de Laplace:

$$\tilde{\pi}(x_j|y_\theta) \approx \frac{\pi(\boldsymbol{x},\boldsymbol{\theta}|\boldsymbol{y})}{\pi_{GG}(x_j|x_j,\boldsymbol{y},\boldsymbol{\theta})}|_{x_j=x_j^*(x_j,\theta)},$$

donde $x_j^*(x_j,\theta)$ es la moda de la distribución de $x_j|x_j,\boldsymbol{y},\boldsymbol{\theta}$. La aproximación es muy buena, tanto como $x_j|x_j,y,\theta$ sea casi gaussiana.

d) Integración numérica para aproximar la marginal de x_j:

$$\pi(x_j|\boldsymbol{y}) \approx \int_\Theta \tilde{\pi}(x_j|\boldsymbol{\theta},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_k \tilde{\pi}(x_j|\boldsymbol{\theta}^{(k)},\boldsymbol{y})\tilde{\pi}(\boldsymbol{\theta}^{(k)}|\boldsymbol{y})w_k,$$

donde $\boldsymbol{\theta}^{(k)} = (\theta_1^{(k)},\theta_2^{(k)},\cdots,\theta_P^{(k)})$ y w_k representa el conjunto de sus pesos correspondientes.

Capítulo 3

Modelo multivariado blockNNGP

Los datos especiales suelen ser multivariantes en el sentido de que en cada unidad se miden múltiples resultados (es decir, más de uno). Este conjunto de datos contiene observaciones referenciadas geográficamente, en otras palabras, tanto el valor como el lugar donde se recoge (las coordenadas) aparecen en el conjunto de datos. Entre las diversas aplicaciones de los modelos geoestadísticos, es común encontrarnos con la necesidad de realizar interpolación espacial en lugares donde no se han tomado muestras, considerando tanto la dependencia entre las mediciones en un lugar concreto como la asociación entre las mediciones en distintos lugares. En este contexto, a menudo resulta crucial modelar simultáneamente múltiples variables espaciales.

3.1. Definición del modelo multivariado geoestadístico

Sea $\boldsymbol{Y_k} = (Y_{1,k}, \ldots, Y_{n,k})^\top$ un vector aleatorio de n variables aleatorias de interés, donde cada variable aleatoria $Y_{i,k}$ para $i = 1, \ldots, n$ ubicaciones, y $k = 1, \ldots, K$ sigue la misma distribución univariada denotada por FD, así $Y_{i,k} \sim FD\left(\mu_{i,k}, \tau_k^2\right)$, donde $\mu_{i,k}$ es la media de $Y_{i,k}$ y τ_k^2 representa un posible parámetro de escala o dispersión. En particular, siguiendo a Palmí-Perales et al. (2023), asumimos que

$$g(\mu_{i,k}) = \boldsymbol{X}_{ik}^\top \boldsymbol{\beta}^{(k)} + \sum_{j=1}^{k} w_{i,j}, \tag{3.1}$$

donde $g(.)$ es una función de enlace, $\boldsymbol{\beta}^{(k)}$ es un vector de coeficientes de regresión, $\boldsymbol{X}_{ik}^T$ es un vector de covariables y $w_{i,j}$ es un efecto aleatorio espacial estructurado del campo espacial latente $\{w_j(s)\}$. Por ejemplo, para $K = 2$, bajo esta definición se asume que la media para un local i y $k = 2$, es decir $\mu_{i,2}$, se explica por una media global asociada a las covariables

y por el efecto aleatorio espacial en el mismo local i y $k = 1$. Esto implica que en general asumimos que la distribución espacial de una variable respuesta puede ser usada para explicar la distribución espacial de otra variable, y esto se puede generalizar a más de dos variables respuestas.

En general, $\boldsymbol{w_{Sk}} = (w_{1,k}, w_{2,k}, \ldots, w_{n,k})^\top$ es una realización del campo espacial $\{w_k(s)\}$, para $k = 1, \ldots, K$ en $\boldsymbol{S} = \{s_1, \ldots, s_n\}$, un conjunto fijo de n ubicaciones en $\boldsymbol{D} \in \Re^2$. Se suele asumir que estos términos de efectos aleatorios provienen de un proceso gaussiano $\{w_k(s)\}$ definido utilizando una función de covarianza válida $\boldsymbol{C(.)}$. Por ejemplo, una función de covarianza exponencial con elementos $C(s_i, s_h) = \sigma^2 \exp(-\phi\|s_i - s_h\|)$, donde $\|s_i - s_h\|$ es la distancia euclídeana entre dos localizaciones; σ^2 es la varianza marginal, y el decaimiento espacial ϕ está relacionado con el llamado rango efectivo (r), la distancia a la que la correlación decae a 0.1, es decir, $\phi = 2/r$.

3.1.1. Modelo multivariado geoestadístico usando blockNNGP

En este trabajo se propone aproximar cada proceso gaussiano $\{w_k(s)\}$ usando el enfoque blockNNGP propuesto por Quiroz et al. (2023), el cual se basa en dividir el dominio espacial en varios bloques dependientes bajo ciertas restricciones, donde los bloques cruzados permiten capturar la dependencia espacial a gran escala, mientras que cada bloque individual captura la dependencia espacial en una escala más pequeña.

Sea $\boldsymbol{w_{Sk}}$ la realización del proceso gaussiano para la k-ésima variable respuesta, es decir tiene distribución normal con vector de medias cero y matriz de covarianza $\boldsymbol{C_{sk}}$. Consideremos una partición de $\boldsymbol{D}$ en M bloques $(b_1, \ldots, b_M)$, donde $U_{m=1}^{M} b_m = \boldsymbol{D}, b_m \cap b_l = \emptyset$, para todos los pares de bloques b_m y b_l. El vector $\boldsymbol{w_{kb_m}} = \{w_{s_i,k}; s_i \in b_m\}$ para la variable respuesta k y $\dim(\boldsymbol{w_{kb_m}}) = n_{km}$ tal que $\sum_{m=1}^{M} n_{km} = n$. En este contexto la densidad conjunta de $\boldsymbol{w_{Sk}} = (w_{s_1,k}, \ldots, w_{s_n,k})^\top$ está dada por:

$$\pi(\boldsymbol{w_{Sk}}) = \pi(\boldsymbol{w_{kb_1}}) \prod_{m=2}^{M} \pi(\boldsymbol{w_{kb_m}} | \boldsymbol{w_{kb_1}}, \ldots, \boldsymbol{w_{kb_{m-1}}}). \tag{3.2}$$

Se define un subconjunto $\boldsymbol{N(b_m)}$ que está compuesto por los nb bloques vecinos "pasados"del bloque b_m. Luego la función densidad conjunta de $\boldsymbol{w_S}$ (Ecuación (3.2)) se aproxima por:

$$\widetilde{\pi}(\boldsymbol{w_{Sk}}) = \pi(\boldsymbol{w_{kb1}}) \prod_{m=2}^{M} \pi(\boldsymbol{w_{kb_m}} | \boldsymbol{w_{N(kb_m)}}).$$

En particular, si $\boldsymbol{w_{Sk}} \sim N(0, \boldsymbol{C_{Sk}}(\theta_{ki}))$, se aproxima la función de densidad conjunta de

$\boldsymbol{w_{Sk}}$ por:

$$\widetilde{\pi}(w_{Sk}) = \prod_{m=1}^{M} f(w_{kb_m}|B_{kb_m}w_{N(kb_m)}, F_{kb_m}),$$

donde f es la densidad conjunta de una distribución normal, $\boldsymbol{B}_{kb_m} = \boldsymbol{C}_{kb_m,\boldsymbol{N(kb_m)}}\boldsymbol{C}^{-1}_{\boldsymbol{N(kb_m)}}$ y $\boldsymbol{F}_{kb_m} = \boldsymbol{C}_{kb_m} - \boldsymbol{C}_{kb_m,\boldsymbol{N(kb_m)}}\boldsymbol{C}^{-1}_{kb_m}\boldsymbol{C}_{\boldsymbol{N(kb_m)},kb_m}$, siendo $\boldsymbol{C}_{kb_m}$, $\boldsymbol{C}^{-1}_{\boldsymbol{N(kb_m)}}$ y $\boldsymbol{C}_{kb_m,\boldsymbol{N(kb_m)}}$ submatrices de $\boldsymbol{C}_{sk}$.

En general, si $\boldsymbol{G}$ es un grafo encadenado y $\boldsymbol{G}^b$ es un grafo de cadena acíclico, entonces $\widetilde{\pi}(\boldsymbol{w_S})$ es una función de densidad conjunta. En particular, dado que $\boldsymbol{w_{Sk}}$ es un vector aleatorio normalmente distribuido con media cero, matriz de covarianza $\boldsymbol{C}_{sk}$ y densidad conjunta que $\pi(\boldsymbol{w_S})$. Para un grafo encadenado $\boldsymbol{G}$ y un grafo de cadena acíclico $\boldsymbol{G^b}$, entonces $\widetilde{\pi}(\boldsymbol{w_S})$ representa la función de densidad conjunta de una distribución normal de n-variables con media cero y matriz de precisión (inversa de matriz de covarianza) definida positiva $\widetilde{\boldsymbol{Q}}_{\boldsymbol{Sk}} = (\boldsymbol{B}^{\boldsymbol{T}}_{\boldsymbol{Sk}}\boldsymbol{F}^{-\boldsymbol{1}}_{\boldsymbol{Sk}}\boldsymbol{B_{Sk}})$. Cuanto mayor sea el número de bloques, esta matriz de precisión es más dispersa (está llena de ceros). Para construir el proceso espacial, se define un conjunto de locales $\boldsymbol{U} \in D$ disjunto de $\boldsymbol{S}$. Y a partir de la distribución condicional de $\boldsymbol{U}$ dado $\boldsymbol{S}$ se define la fdp conjunta para cualquier conjunto de locales $\boldsymbol{V} = \{S \cup U\}$ en $\boldsymbol{D}$. Se prueba que se construye un proceso espacial con media cero y matriz de precisión dispersa. Esta característica permite definir el blockNNGP como un campo aleatorio markoviano gaussiano. Para más detalles sobre el blockNNGP, revisar Quiroz et al. (2023).

3.1.2. Esquema de bloques

El dominio espacial debe dividirse en varias regiones, cómo crear estas regiones y cuántas regiones necesitamos son preguntas importantes. Idealmente, se asume que los datos son homogéneos dentro de las regiones, es decir, las regiones tienen un número similar de ubicaciones, esta característica asegura cálculos paralelos más rápidos. Se presenta dos tipos de particiones: i) particionado regular: El particionado regular es muy simple, simplemente divide el dominio espacial en M regiones rectangulares disjuntas; y ii) particionado por teselación de Voronoi: El enfoque de teselación de Voronoi propuesto por Green y Sibson (1978) divide el dominio espacial en M regiones disjuntas b_1, b_2, ..., b_M, con centros $c_1, \ldots, c_M$, de tal manera que los puntos dentro de B_m están más cerca de c_m que cualquier otro centro c_j, para $j \neq m$.

Con respecto al número de bloques, se debe tomar en cuenta que si el número de bloques es muy pequeño la matriz de precisión sería menos dispersa (tendría menos ceros), por otro lado el número de bloques tampoco debe ser muy grande porque aunque la matriz de dispersión

sería más dispersa, se estaría perdiendo mucha información y si el rango es pequeño, la inferencia no sería adecuada. En general se recomienda, tomar en cuenta que el tamaño del bloque no debe ser menos al rango que se espera, y se recomienda de forma empírica aproximadamente $M = \sqrt{n}$.

3.2. Inferencia bayesiana usando INLA

El modelo presentado en la sección anterior pertenece a la clase modelos gaussianos latentes (MGL), por lo tanto, las estimaciones posteriores de los parámetros se pueden realizar utilizando el método Aproximación de Laplace iterada y anindada (INLA, de las siglas en inglés Integrated nested Laplace approximation) propuesto por Rue et al. (2009).

Primero presentamos el modelo de forma jerárquica como un MGL. Se define el vector aleatorio $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_K)^\top$, el campo gaussiano latente $\boldsymbol{x}$ donde se incluyen todas las distribuciones a priori gaussianas y el vector de hiperparámetros θ. Entonces podemos escribir el modelo como un MGL de la siguiente manera:

$$\begin{aligned} \boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}), && \text{hiperparámetros} \\ \boldsymbol{x} \mid \boldsymbol{\theta} &\sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{Q}^{-1}\right), && \text{Campo gaussiano latente} \\ \pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) &= \prod_{k=1}^{K} \prod_{i=1}^{n} \pi(y_{i,k}|\boldsymbol{x}, \boldsymbol{\theta}), && \text{Verosimilitud de las observaciones} \end{aligned}$$

donde específicamente asumimos que las variables aleatorias $Y_{i,k}$ son condicionalmente independientes dadas el campo gaussiano latente y los hiperparámetros. El campo latente se define como $\boldsymbol{x} = \left(\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \ldots, \boldsymbol{\beta}^{(k)}, \boldsymbol{w_{S1}}, \boldsymbol{w_{S2}}, \ldots, \boldsymbol{w_{SK}}\right) = (\boldsymbol{\beta}, \boldsymbol{w})$. Suponemos una distribución a priori normal para los coeficientes de regresión, $\boldsymbol{\beta} \sim N\left(0, \boldsymbol{Q}_\beta^{-1}\right)$ y una distribución a priori blockNNGP para cada proceso espacial k-ésimo, $\boldsymbol{w_k} \mid \boldsymbol{\theta_k} \sim \text{blockNNGP}\left(0, [\widetilde{\boldsymbol{Q}}\left(\boldsymbol{\theta_k}\right)]^{-1}\right)$. Entonces,

$$\boldsymbol{Q} = \begin{bmatrix} \boldsymbol{Q}_\beta & \mathbf{0} \\ \mathbf{0} & \widetilde{\boldsymbol{Q}} \end{bmatrix},$$

donde $\widetilde{\boldsymbol{Q}} = \text{diag}(\widetilde{\boldsymbol{Q}}(\theta_k))$. El posible vector de hiperparámetros se define como:

$$\boldsymbol{\theta} = \left(\theta_1, \ldots, \theta_k, \tau_1^2, \ldots, \tau_k^2\right) = \left(\phi_1, \sigma_1^2, \phi_2, \sigma_2^2, \ldots, \phi_K, \sigma_K^2, \tau_1^2, \ldots, \tau_K^2\right),$$

donde $\boldsymbol{\theta_k} = \{\phi_k, \sigma_k^2\}$. Dado que el rango efectivo $\rho^k = 2/\phi^k$, siguiendo la propuesta de Fuglstad et al. (2019), se asignó una distribución a priori conjunta compleja penalizada (PC

prior) para ρ_k y σ_k, es decir,

$$\pi\left(\sigma_k, \rho_k\right) = \tilde{\lambda}_{1k}\tilde{\lambda}_{2k}\rho_k^{-2}\exp\left(-\tilde{\lambda}_{1k}\rho_k^{-1} - \tilde{\lambda}_{2k}\sigma_k\right), \quad \sigma_k > 0, \rho_k > 0,$$

donde

$$\tilde{\lambda}_{1l} = -\log\left(\alpha_{1l}\right)\rho_{0l} \quad \text{y} \quad \tilde{\lambda}_{2l} = -\frac{\log\left(\alpha_{2l}\right)}{\sigma_{0l}},$$

y $\mathrm{P}\left(\rho_l < \rho_{0l}\right) = \alpha_{1l}$ y $\mathrm{P}\left(\sigma_l > \sigma_{0l}\right) = \alpha_{2l}$, donde $\rho_{0l}, \sigma_{0l}, \alpha_{1l}$, y α_{2l} son determinados según sus conocimientos previos. Para τ_k^2 se asumió una distribución gamma inversa informativa.

La distribución conjunta a posteriori del MGL puede calcularse utilizando la distribución de verosimilitud de $\mathbf{Y}$, la distribución del campo gausiano latente $\boldsymbol{x}$ y la distribución de hiperparámetros $\boldsymbol{\theta}$ como:

$$\begin{aligned}
&\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})\pi(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}) \\
&\pi(\boldsymbol{x}, \boldsymbol{\theta}|\boldsymbol{y}) \propto \pi(\boldsymbol{\theta})|\boldsymbol{Q}(\boldsymbol{\theta})|^{1/2}\exp\left\{\frac{-1}{2}\boldsymbol{x}^T\boldsymbol{Q}(\boldsymbol{\theta})x\right\}\prod_{k=1}^{K}\prod_{i=1}^{n}\pi(y_{i,k}|\boldsymbol{x}, \boldsymbol{\theta}).
\end{aligned}$$

INLA proporciona aproximaciones para las marginales a posteriori de las variables latentes y los hiperparámetros calculadas numéricamente a partir de:

$$\begin{aligned}
\widetilde{\pi}(x_l|\boldsymbol{y}) &= \int \widetilde{\pi}(x_l|\boldsymbol{\theta}, \boldsymbol{y})\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\boldsymbol{\theta} \quad \text{y} \\
\widetilde{\pi}(\theta_t|\boldsymbol{y}) &= \int \widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})d\theta_{-t},
\end{aligned} \tag{3.3}$$

donde $\widetilde{\pi}(.)$ representan aproximaciones de las funciones de densidad. En resumen, la idea principal de INLA se divide en las siguientes tareas: 1) a partir de una aproximación gaussiana de la distribución condicional completa de x, obtiene la aproximación de Laplace para la a posteriori conjunta de hipeparámetros dados los datos $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$; 2) a continuación, proporciona una aproximación de la distribución condicional del campo latente dados los datos y los hiperparámetros $\widetilde{\pi}(x_l|\boldsymbol{\theta}, \boldsymbol{y})$; y 3) por último, explora $\widetilde{\pi}(\boldsymbol{\theta}|\boldsymbol{y})$ en una cuadrícula y la utiliza para integrar $\boldsymbol{\theta}$ y θ_{-t} numéricamente en la ecuación (3.3).

3.2.1. Predicción

Para una nueva ubicación $u_0 \notin S$, podemos obtener muestras posteriores del efecto aleatorio espacial $w_{u_0,k}$ a partir de $w_{u_0,k} \mid \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{y} \sim N(m_{k0}, v_{k0})$, la media y varianza son respectivamente:

$$m_{k0} = C^{\top}_{u_0,N(u_0)}(\boldsymbol{\theta_k})C^{-1}_{N(u_0),N(u_0)}(\boldsymbol{\theta_k})\boldsymbol{w_k}(N(u_0)),$$

$$v_{k0} = \sigma_k^2 - C_{u_0,N(u_0)}^{\top}(\boldsymbol{\theta_k}) C_{N(u_0),N(u_0)}^{-1}(\boldsymbol{\theta_k}) C_{u_0,N(u_0)}(\boldsymbol{\theta_k}),$$

donde $N(u_0)$ es un conjunto de vecinos para u_0 en S. Luego, el muestreo se realiza a través de la distribución predictiva a posteriori:

$$y_{u_0,k} \mid \boldsymbol{w}_{\boldsymbol{u_0}}^{\star}, \boldsymbol{w}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{y}) \overset{ind}{\sim} FD\left(\mu_{u_0}^{\star}, \tau_k^2\right),$$

donde $\mu_{u_0}^{\star} = g^{-1}(\boldsymbol{X}(s_{ik})^{\top}\boldsymbol{\beta}^{(k)} + \sum_{j=1}^{k} w_{u_0,j})$ y $\boldsymbol{w}_{\boldsymbol{u_0}}^{\star} = (w_{u_0,1}, w_{u_0,1}, \ldots, w_{u_0,k})$.

Capítulo 4

Estudio de Simulación

En este capítulo, se desarrollan los estudios de simulación de recuperación de parámetros con el objetivo de evaluar el rendimiento de nuestra implementación del modelo geoestadístico multivariado usando block-NNGP e INLA. Se desarrolló un estudio de simulación del modelo planteado en el capítulo 3. Se ajustaron los modelos NNGP, Block-NNGP regular y Block-NNGP irregular.

Para evaluar el rendimiento de los modelos blockNNGP multivariados, presentamos el siguiente experimento de simulación. Generamos s_i localizaciones aleatorias en un dominio espacial $[0,1]\times[0,1]$ donde $i = 1, \ldots, 2500$. En estos locales se simularon los efectos espaciales aleatorios para $K = 2$, es decir a partir de dos procesos gaussianos (PG) espaciales, es decir para $k = 1, 2$, $\{w_k(s)\}$ es un proceso gaussiano con media cero y función de covarianza exponencial $\boldsymbol{C}_k$ con elementos $C_k(s_i, s_j) = \sigma_k^2 \exp(-\phi_k \|s_i - s_j\|)$. Establecemos $\sigma_1^2 = 1.5$, $\sigma_2^2 = 1$ y los rangos efectivos $r_1 = 0.2$ ($\phi_1 = 10$), $r_2 = 0.33$ ($\phi_2 = 6$). También generamos las covariables $x_i \sim N(0,1)$, y asumimos que $\boldsymbol{X}_{ik} = (1, x_i)^T$. Bajo esta configuración se simularon datos a partir de un modelo gaussiano multivariado y a partir de un modelo poisson multivariado como se presenta en la siguientes secciones. Cabe resaltar que en ambos estudios los datos se dividieron en un conjunto de entrenamiento y otro conjunto de prueba para la predicción, específicamente el 80 % de los datos se utilizó para el entrenamiento y el 20 % restante para la predicción.

4.1. Modelo gaussiano multivariado

Los valores de los parámetros de regresión vienen dados por $\boldsymbol{\beta}^{(1)} = (\beta_{01}, \beta_{11})^T = (1, 5)^T$, $\boldsymbol{\beta}^{(2)} = (\beta_{02}, \beta_{12})^T = (2, 3)^T$ y los valores de los efecto pepita son $\tau_1^2 = 0.1$ y $\tau_2^2 = 0.3$. Por último, simulamos 2500 muestras del modelo multivariante presentado en la sección 3, para

simular variables de respuesta gaussianas, es decir,

$$Y_{i,k} \mid \boldsymbol{x}, \boldsymbol{\theta} \overset{ind}{\sim} N\left(\boldsymbol{X}_{ik}^T\boldsymbol{\beta}^{(k)} + \sum_{j=1}^{k} w_{i,j}, \tau^2\right); \quad i = 1, \ldots, n; \quad k = 1, 2.$$

Para los datos de entrenamiento $n = 2000$, ajustamos los modelos blockNNGP eligiendo $M = 64, 100$ bloques regulares, $M = 64, 128$ bloques irregulares y $nb = 2, 4, 6$ bloques vecinos, y el modelo NNGP (caso particular del blockNNGP con $M = n$ bloques) con $nb = 10, 20, 30$ vecinos. Usamos la misma configuración para la dos variables respuesta k=1,2. La inferencia bayesiana completa se llevó a cabo implementando el blockNNGP en `R-INLA`. En estas simulaciones, utilizamos una a priori compleja penalizada para φ_k y σ_k, considerando $\rho_k = 2/\phi_k$, asumimos que $P(\rho_k < 0.075) = 0.05$ y $P(\sigma_k > 5) = 0.05$.

Para comparar los modelos ajustados se calcularon los siguiente criterios: el criterio de información ampliamente aplicable (WAIC - widely applicable information criterion), el logaritmo de la verosimilitud pseudo-marginal (LPML - logarithm of the pseudo marginal likelihood), el error cuadrático medio de estimación para la variable respuesta Y_k, (MSE Y_k) y el error cuadrático medio de predicción para la variable respuesta Y_k (MSP Y_k).

Los resultados de la evaluación de criterios de comparación entre los modelos ajustados se muestran en el cuadro 4.1. Podemos observar que el modelo NNGP es mejor cuando aumenta el número de vecinos según el LPML, sin embargo este resultado no se replica según el WAIC. En general, los modelos block-NNGP con bloques regulares e irregulares muestran una mejor bondad de ajuste que los modelos NNGP, pues tienen menor WAIC y mayor LPML que los modelos NNGP. Según los resultados para los modelos blockNNGP regulares como se esperaba, el modelo es mejor para un menor número de bloques y un mayor número de bloques vecinos. En el caso de los modelos block-NNGP con bloques irregulares este patrón es similar según el LPML, pero no es tan claro con el WAIC.

Por otro lado, en todos los modelos ajustados, el MSE es menor para la variable respuesta 1, esto tiene sentido pues puede reflejar que la variable 2 depende del efecto aleatorio espacial de la variable 1, por lo tanto contiene el error de estimación de esta variable. En particular, el MSE de las variable respuestas son mejores para los modelos NNGP con menos vecinos $nb = 10$, sin embargo como ya vimos este modelo presenta el peor ajuste según los criterios WAIC y LPML. Por otro lado, para los modelos blockNNGP con bloques regulares, el MSE es mejor cuando se tienen más bloques vecinos. En el caso de los modelos blockNNGP con bloques irregulares, resultó ser mejor el modelo con 28 bloques y 6 bloques vecinos. Con respecto a la predicción del modelo, según los valores de MSP el mejor modelo es el modelo

NNGP con 30 vecinos seguido del modelo blockNNGP con 64 bloques regulares y cuatro bloques vecinos, y el modelo blockNNGP con 100 bloques regulares y 6 bloques vecinos. Por último, en general también se observa que el tiempo disminuye a medida que aumentamos el número de vecinos, o tenemos menos bloques. La ventaja del modelo blockNNGP con bloques irregulares en términos de tiempo se observa para el modelos con 64 bloques y dos bloques vecinos, pues incluso demora menos que el modelo NNGP con 10 vecinos. Y aunque sus criterios de selección no sean los mejores, son muy similares a los otros modelos blockNNGP ajustados.

Cuadro 4.1: Modelo Gaussiano Multivariado - Resumen de métricas de desempeño de los modelos NNGP, blockNNGP - Regular y blockNNGP - irregular

Modelos	NNGP			BlockNNGP Regular				BlockNNGP Irregular			
	nb=10	**nb=20**	**nb=30**	**M=64 nb=2**	**M=64 nb=4**	**M=100 nb=4**	**M=100 nb=6**	**M=64 nb=2**	**M=64 nb=4**	**M=128 nb=4**	**M=128 nb=6**
WAIC	6591.053	6733.25	6786.071	6318.788	**6298.883**	6303.394	6299.547	6317.426	6322.037	**6285.01**	**6287.83**
LPML	-4575.861	-4333.693	-4233.602	-3630.977	**-3628.620**	-3629.02	**-3628.835**	-3710.297	-3707.684	-3741.907	-3732.966
MSE Y_1	0.024	0.025	0.030	0.043	0.043	0.043	0.042	**0.040**	0.041	**0.039**	**0.039**
MSE Y_2	0.098	0.161	0.166	0.177	0.173	0.175	0.174	**0.161**	**0.161**	**0.149**	**0.153**
MSP Y_1	0.305	0.303	0.305	0.330	0.328	**0.323**	0.324	**0.322**	**0.321**	0.563	0.562
MSP Y_2	0.739	0.746	0.733	0.752	**0.753**	**0.769**	**0.766**	0.777	0.775	1.303	1.304
Time (sec)	1188.46	2111.911	2840.857	1233.533	4411.686	1227.529	2107.42	802.056	1982.629	1439.135	2026.592

El desempeño de la estimación de los parámetros de cada modelo se muestra en el cuadro 4.2. En general, para los modelos blockNNGP, los parámetros reales se encuentran dentro de los intervalos de credibilidad (95 % IC) o las estimaciones de la media a posteriori están cerca del valor real de los parámetros, siendo mejor la estimación usando bloques regulares. Sin embargo para los modelos NNGP, sobre todo para los parámetros espaciales no se recuperan bien los valores originales de los parámetros.

Este resultado se observa mejor a través de las distribuciones marginales a posteriori de los parámetros de decaimiento y la varianza marginal para ambas variables, las cuales se muestran en la Figura 4.1. Este resultado muestra que la los modelos NNGP no estiman correctamente los parámetros espaciales, bajo ningún número de vecinos. Mientras que entre los modelos blockNNGP ajustados, los modelos con bloques regulares consiguieron estimar mejor los parámetros.

Las distribuciones marginales a posteriori del efecto pepita para ambas variables también se muestran en la Figura 4.1. Se observa que los modelos blockNNGP estimaron correctamente este parámetro. En el caso del modelo NNGP, se estima correctamente el efecto pepita de la variable respuesta 1, pero el efecto pepita de la variable respuesta 2 no es estimado correctamente por el modelo con 10 vecinos y es ligeramente subestimado cuando se usan 20 o 30 vecinos.

Cuadro 4.2: Modelo Gaussiano Multivariado - Resumen de las estimaciones de las medias a posteriori e intervalos de credibilidad (Lím. Inf., Lím. Sup.) al 95 % de los parámetros de los modelos NNGP, blockNNGP - Regular y blockNNGP - irregular

Modelos		NNGP			BlockNNGP Regular				BlockNNGP Irregular			
	Valor Original	nb=10	nb=20	nb=30	M=64 nb=2	M=64 nb=4	M=100 nb=4	M=100 nb=6	M=64 nb=2	M=64 nb=4	M=128 nb=4	M=128 nb=6
Variable 1												
β_{01}	1	1.615	1.69	1.709	1.611	1.573	1.577	1.5810	1.601	1.599	1.587	1.571
Lím. Inf.		1.507	1.548	1.546	1.236	1.161	1.179	1.18	1.413	1.391	1.405	1.378
Lím. Sup.		1.722	1.834	1.877	1.992	1.982	1.974	1.983	1.788	1.804	1.765	1.758
β_{11}	5	4.98	4.985	4.989	4.98	4.98	4.98	4.98	4.983	4.982	4.979	4.979
Lím. Inf.		4.946	4.952	4.958	4.955	4.955	4.955	4.955	4.958	4.958	4.954	4.953
Lím. Sup.		5.015	5.018	5.021	5.005	5.004	5.005	5.005	5.008	5.007	5.004	5.004
σ_1^2	1.5	1.228	1.209	1.188	1.338	1.348	1.331	1.338	1.194	1.226	1.32	1.314
Lím. Inf.		1.116	1.082	1.053	1.041	1.032	1.033	1.039	1.008	1.026	1.12	1.114
Lím. Sup.		1.35	1.347	1.34	1.723	1.752	1.719	1.732	1.41	1.46	1.554	1.548
ϕ_1	10	12.086	12.935	12.862	12.032	12.058	12.214	12.138	13.471	13.042	12.465	12.57
Lím. Inf.		10.567	11.206	10.987	8.895	8.828	9.009	8.93	10.968	10.542	10.135	10.241
Lím. Sup.		13.732	14.861	14.905	15.562	15.846	15.836	15.726	16.238	15.858	15.026	15.148
τ_1^2	0.1	0.1	0.096	0.103	0.104	0.104	0.104	0.104	0.103	0.104	0.103	0.102
Lím. Inf.		0.076	0.074	0.082	0.089	0.088	0.088	0.088	0.087	0.088	0.087	0.086
Lím. Sup.		0.128	0.121	0.126	0.122	0.121	0.12	0.121	0.121	0.121	0.12	0.12
Variable 2												
β_{02}	2	2.2	1.898	2.059	2.279	2.238	2.246	2.255	2.197	2.209	2.15	2.124
Lím. Inf.		2.026	0.91	0.734	1.487	1.437	1.435	1.437	1.909	1.877	1.893	1.844
Lím. Sup.		2.372	2.535	3.127	3.091	3.064	3.089	3.115	2.487	2.543	2.408	2.405
β_{12}	3	2.953	2.958	2.961	2.956	2.957	2.957	2.957	2.962	2.963	2.956	2.956
Lím. Inf.		2.906	2.913	2.918	2.922	2.923	2.923	2.923	2.927	2.928	2.921	2.921
Lím. Sup.		3	3.002	3.003	2.991	2.991	2.991	2.991	2.997	2.998	2.991	2.991
σ_2^2	1	1.012	1.977	2.493	1.234	1.162	1.193	1.204	0.911	0.937	0.884	0.88
Lím. Inf.		0.869	1.149	1.249	0.76	0.725	0.723	0.729	0.718	0.724	0.719	0.71
Lím. Sup.		1.173	3.615	5.036	2.003	1.883	1.97	1.994	1.148	1.207	1.078	1.085
ϕ_2	6	8.058	2.783	2.181	5.078	5.574	5.402	5.384	7.481	7.368	8.279	7.952
Lím. Inf.		6.367	1.123	0.817	2.659	2.898	2.739	2.713	5.308	5.067	6.201	5.83
Lím. Sup.		10.056	4.87	4.123	8.216	8.949	8.903	8.86	10.01	10.114	10.784	10.405
τ_2^2	0.3	0.273	0.334	0.329	0.295	0.293	0.293	0.293	0.288	0.287	0.278	0.281
Lím. Inf.		0.214	0.286	0.284	0.262	0.259	0.26	0.26	0.253	0.252	0.243	0.246
Lím. Sup.		0.342	0.389	0.38	0.332	0.33	0.33	0.33	0.327	0.325	0.316	0.32

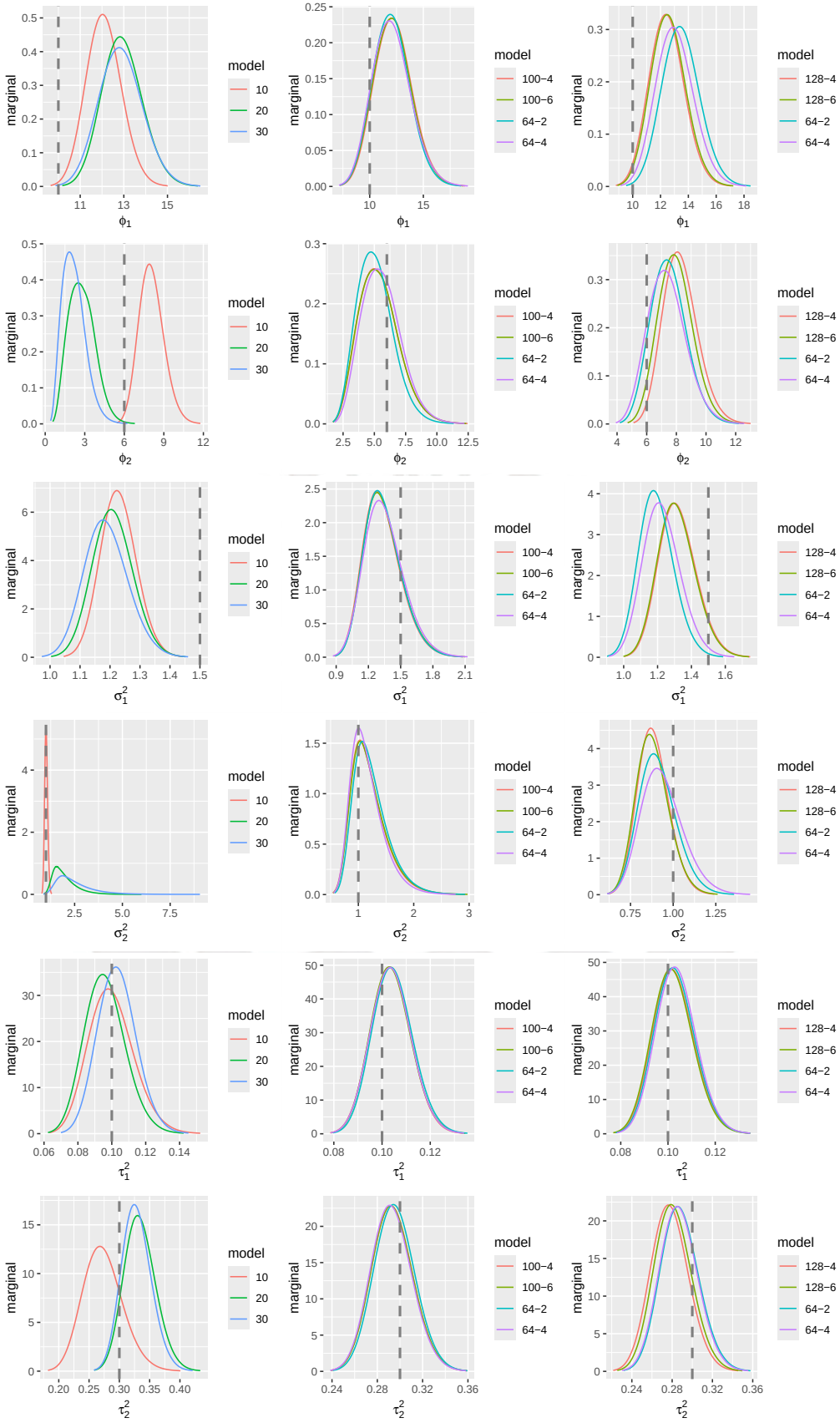

Figura 4.1: Modelo gaussiano multivariado - Densidades marginales a posteriori del parámetro de decaimiento espacial ϕ_k , varianza marginal σ_k^2 y efecto pepita τ_k^2, para los modelos NNGP (izquierda), blockNNGP regular (centro), blockNNGP irregular (derecha). La línea vertical punteada gris representa el verdadero valor del parámetro.

Dado que el rendimiento de todos los modelos blockNNGP con bloques regulares es bastante similar, continuamos el análisis con el blockNNGP regular (64-4). La Figura 4.2 muestra los efectos aleatorios espaciales reales (panel superior) y las estimaciones de las medias a posteriori (panel inferior), para la variable 1 (izquierda) y la variable 2 (derecha). Es evidente que los efectos aleatorios espaciales son recuperados, pero destacamos que el efecto aleatorio espacial w_2 de la variable respuesta 2 está un poco más suavizado que el efecto aleatorio original.

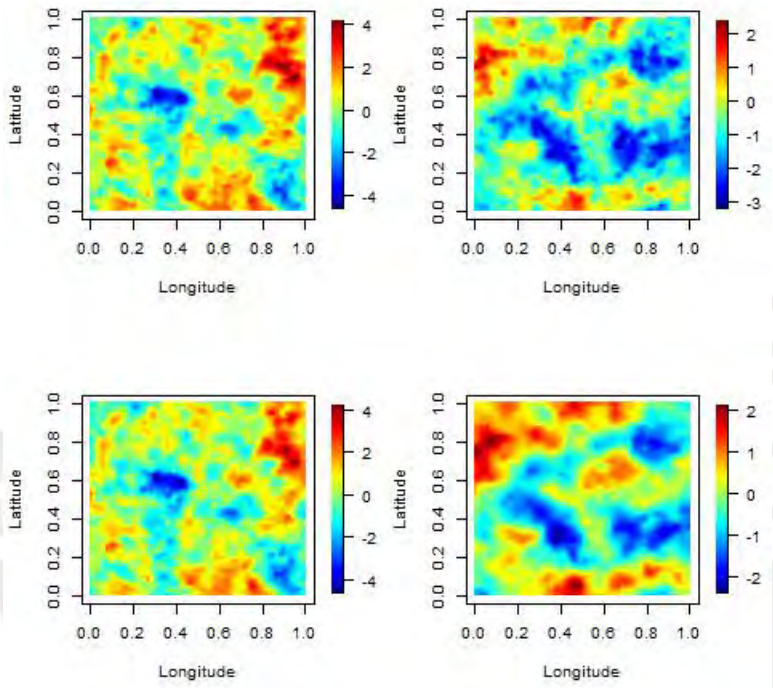

Figura 4.2: Modelo gaussiano multivariado - Simulación BlockNNGP regular (M=64, nb=4). Arriba: Efectos aleatorios originales w_1 (izquierda) y w_2 (derecha). Abajo: Estimaciones de media a posteriori de efectos aleatorios.

En la Figura 4.3 también se comparan los verdaderos efectos aleatorios espaciales con sus estimaciones de medias a posteriori (panel superior). En general, podemos observar que los efectos aleatorios espaciales son recuperados. Sin embargo, la media a posteriori del efecto aleatorio espacial w_2 tiene mayor variabilidad. Esto podría explicar el patrón de suavidad mostrado para este efecto aleatorio en la Figura 4.2. Este resultado tiene sentido ya que todo el efecto aleatorio espacial de la variable de respuesta 2 contiene un efecto aleatorio espacial compartido de la variable de respuesta 1. La Figura 4.3 también compara la verdadera variable de respuesta con sus estimaciones de la media a posteriori (panel inferior). Muestra que la verdadera variable de respuesta se recupera bien, con más variabilidad para la variable de respuesta 2, debido a la influencia de los efectos aleatorios espaciales.

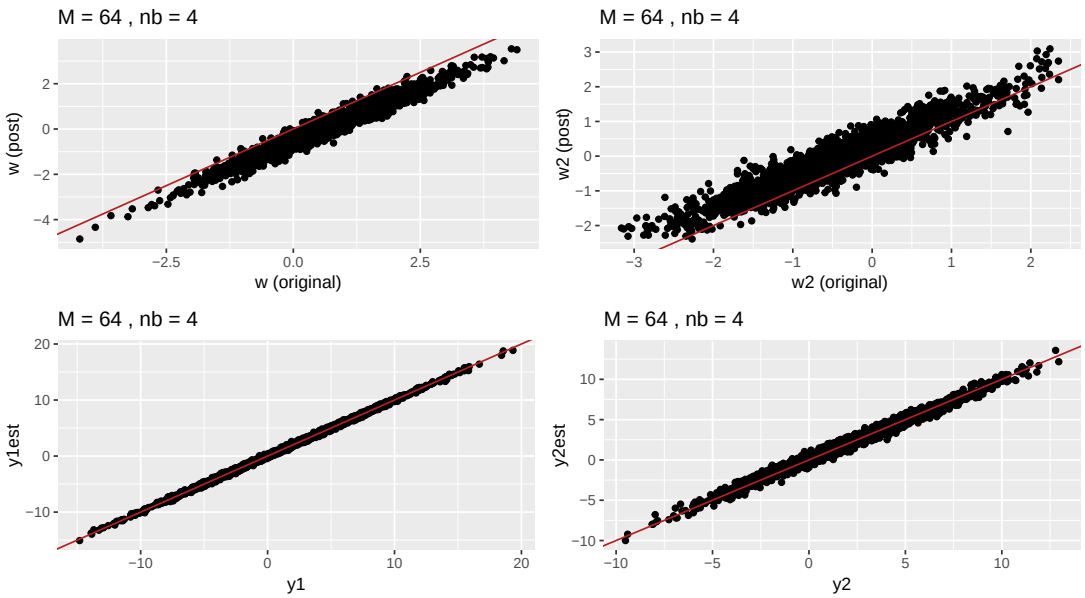

Figura 4.3: Modelo gaussiano multivariado - Simulación BlockNNGP regular (M=64, nb=4). Arriba: Comparación de los datos originales w_1 (izquierda), w_2 (derecha) y la media a posteriori del efecto aleatorio w_i, respectivamente. Abajo: Comparación de los datos originales y_1 (izquierda), y_2 (derecha) y la media a posteriori $y.est$, respectivamente.

En cuanto a la predicción, la Fig. 4.4 compara los valores originales de las variables respuesta (usando los datos de prueba no usados para la estimación) con respecto a las predicciones obtenidas ajustando el modelo blockNNGP regular (64-4). En general, podemos observar que las variables de respuesta son recuperadas de forma eficiente. Vemos que la predicción de y_2 presenta mayor variabilidad, esto explica que el MSP sea mayor para la variable 2, en efecto, este resultado refleja que la variable 2 depende del efecto aleatorio espacial de la variable 1.

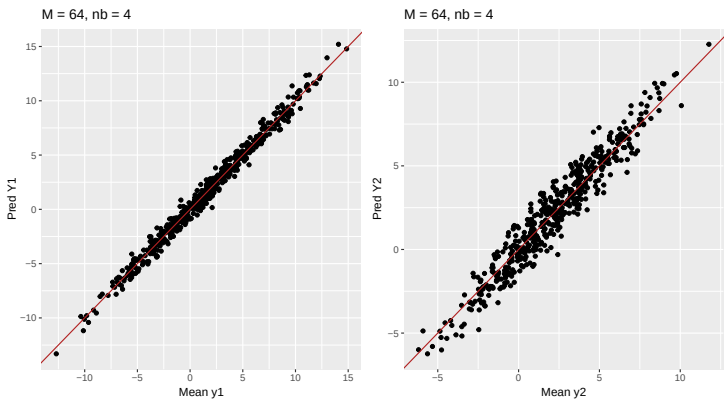

Figura 4.4: Modelo gaussiano multivariado - Predicción BlockNNGP regular (M=64, nb=4). Comparación de los datos originales y_1 (izquierda), y_2 (derecha) y la media a posteriori de las predicciones, respectivamente.

4.2. Modelo Poisson multivariado

Los valores de los parámetros de regresión vienen dados por $\boldsymbol{\beta}^{(1)} = (\beta_{01}, \beta_{11})^T = (1, 5)^T$, $\boldsymbol{\beta}^{(2)} = (\beta_{02}, \beta_{12})^T = (2, 3)^T$. Simulamos $n = 2500$ muestras del modelo multivariante presentado en la sección 3, para variables de respuesta de Poisson, es decir, de

$$Y_{i,k} \mid \boldsymbol{\beta}, \boldsymbol{\theta} \overset{ind}{\sim} \text{Poisson}\left(\mu_{i,k} = \exp(\boldsymbol{X}_{ik}^T \boldsymbol{\beta}^{(k)} + \sum_{j=1}^{k} w_{i,j})\right); i = 1, \ldots, n; k = 1, 2.$$

Asignamos la misma configuración previa presentada en la sección anterior. Los resultados de la evaluación de criterios se muestran en la Tabla 4.3. Podemos observar que según el WAIC y LPML, el modelo NNGP es mejor cuando aumenta el número de vecinos. Los resultados para los modelos blockNNGP tienen menor WAIC y mayor LPML que los modelos NNGP, por lo tanto según estos criterios se ajustan mejor. En particular los modelos con bloques regulares presentan mejor desempeño que los modelos con bloques irregulares y en general, como se esperaba, el modelo es mejor para un menor número de bloques y un mayor número de bloques vecinos.

Cuadro 4.3: Modelo Poisson multivariado - Resumen de métricas de desempeño del modelos NNGP, blockNNGP - regular y blockNNGP - irregular

Modelo	NNGP			
	nb=10	**nb=20**	**nb=30**	
WAIC	12756.637	12591.47	12501.619	
LPML	-44203.896	-36984.9	-33091.595	
MSE Y_1	1.252	1.374	1.435	
MSE Y_2	0.334	0.353	0.349	
MSP Y_1	34.181	35.303	34.771	
MSP Y_2	1.388	1.152	1.177	
Time (sec)	485.785	998.656	2067.947	
Modelo	BlockNNGP Regular			
	M=64 nb=2	**M=64 nb=4**	**M=100 nb=4**	**M=100 nb=6**
WAIC	**11910.712**	**11911.454**	11911.873	**11911.756**
LPML	**-13269.175**	**-13346.719**	**-13380.822**	-13417.016
MSE Y_1	2.042	2.033	2.032	2.024
MSE Y_2	0.396	**0.393**	**0.393**	0.395
MSP Y_1	38.805	38.581	40.159	**37.046**
MSP Y_2	1.479	1.455	**1.189**	**1.232**
Time (sec)	563.772	983.281	864.055	2181.74
Modelo	BlockNNGP Irregular			
	M=64 nb=2	**M=64 nb=4**	**M=128 nb=4**	**M=128 nb=6**
WAIC	11995.574	12005.034	12051.552	12057.255
LPML	-17301.097	-16422.125	-19631.115	-18763.531
MSE Y_1	**1.942**	**1.942**	**1.914**	**1.875**
MSE Y_2	**0.370**	**0.393**	**0.368**	0.399
MSP Y_1	**37.531**	43.644	37.595	**37.064**
MSP Y_2	1.360	1.604	**1.303**	1.698
Time (sec)	700.002	1986.93	805.414	1781.033

Con respecto al MSE y MPE, en general, el MSE y MPE es mayor para la variable

1. Además según estos criterios los modelos NNGP estiman y predicen mejor la variable respuesta, pese a que los criterios anteriores no los favorecen. Según el MSE y MSP el mejor modelo blockNNGP es el modelo con 128 bloques irregulares y 4 bloques. vecinos. Por último, el tiempo aumenta a medida que aumentamos el número de vecinos. Vemos que el modelo blockNNGP con 64 bloques y dos bloques vecinos es próximo al modelo NNGP con 10 vecinos, a pesar de que usa un mayor número de vecinos.

El desempeño de la estimación de los parámetros de cada modelo se muestra en el cuadro 4.4. En general, todos los parámetros reales de los modelos blockNNGP se encuentran dentro de los intervalos de credibilidad al 95 % o las estimaciones de las medias a posteriori están cerca del valor real de los parámetros. Por otro lado, los modelos NNGP si bien estiman correctamente la mayoría de los parámetros, no consiguen estimar el parámetro de decaimiento espacial, asociado al rango de la variable respuesta 2.

Cuadro 4.4: Modelo poisson multivariado - Resumen de las estimaciones de las medias a posteriori e intervalos de credibilidad (Lím. Inf., Lím. Sup.) al 95 % de los parámetros de los enfoques NNGP, blockNNGP - Regular y blockNNGP - irregular

Modelos		NNGP			BlockNNGP Regular				BlockNNGP Irregular			
	Valor verdadero	nb=10	nb=20	nb=30	M=64 nb=2	M=64 nb=4	M=100 nb=4	M=100 nb=6	M=64 nb=2	M=64 nb=4	M=128 nb=4	M=128 nb=6
Variable 1												
β_{01}	1	1.094	1.082	1.107	0.958	0.915	0.911	0.916	1.042	1.102	0.998	1.027
Lím. Inf.		0.943	0.857	0.836	0.391	0.282	0.287	0.321	0.773	0.801	0.78	0.791
Lím. Sup.		1.239	1.294	1.364	1.489	1.506	1.494	1.469	1.303	1.392	1.209	1.254
β_{11}	-0.2	-0.193	-0.199	-0.195	-0.19	-0.191	-0.19	-0.19	-0.192	-0.193	-0.195	-0.197
Lím. Inf.		-0.238	-0.240	-0.234	-0.222	-0.222	-0.222	-0.222	-0.224	-0.225	-0.228	-0.229
Lím. Sup.		-0.149	-0.158	-0.155	-0.159	-0.159	-0.159	-0.159	-0.159	-0.16	-0.163	-0.164
σ_1^2	1.5	1.680	1.656	1.672	1.875	1.839	1.819	1.8	1.761	1.732	1.603	1.65
Lím. Inf.		1.502	1.442	1.400	1.355	1.295	1.275	1.349	1.47	1.357	1.35	1.379
Lím. Sup.		1.877	1.903	1.997	2.661	2.451	2.417	2.405	2.108	2.198	1.922	1.971
ϕ_1	10	10.152	9.989	9.414	9.091	9.335	9.45	9.54	9.562	10.124	10.5	10.732
Lím. Inf.		8.739	8.466	7.470	6.027	6.744	6.851	6.944	7.768	7.419	8.457	8.735
Lím. Sup.		11.714	11.628	11.522	12.43	13.065	13.29	12.486	11.486	13.313	12.566	12.961
Variable 2												
β_{02}	-1	-1.201	-0.935	-1.091	-1.475	-1.491	-1.508	-1.495	-1.603	-1.572	-1.545	-1.32
Lím. Inf.		-1.580	-1.688	-1.762	-2.239	-2.304	-2.307	-2.275	-1.954	-1.999	-1.85	-1.732
Lím. Sup.		-0.721	0.139	-0.360	-0.704	-0.663	-0.699	-0.708	-1.255	-1.125	-1.237	-0.805
β_{12}	0.5	0.503	0.504	0.505	0.492	0.491	0.491	0.491	0.486	0.475	0.482	0.478
Lím. Inf.		0.420	0.426	0.427	0.421	0.420	0.420	0.420	0.413	0.403	0.409	0.405
Lím. Sup.		0.586	0.582	0.582	0.564	0.563	0.563	0.563	0.560	0.548	0.555	0.55
σ_2^2	1	1.064	1.344	1.211	0.831	0.798	0.799	0.786	0.786	0.770	0.808	0.828
Lím. Inf.		0.610	0.878	0.911	0.575	0.558	0.560	0.519	0.582	0.495	0.611	0.583
Lím. Sup.		1.755	2.054	1.547	1.202	1.184	1.178	1.179	1.041	1.204	1.057	1.136
ϕ_2	6	3.338	2.607	3.272	5.716	6.153	6.223	6.363	7.273	6.036	7.139	5.126
Lím. Inf.		2.005	1.256	2.650	3.615	3.639	3.727	3.910	5.260	2.897	4.877	2.796
Lím. Sup.		5.116	4.332	4.152	7.905	8.455	8.532	9.184	9.476	10.195	9.505	7.798

Los distribuciones marginales a posteriori de los parámetros de decaimiento y la varianza marginal se muestran en la Figura 4.5. En general, vemos que las marginales de los modelos blockNNGP contienen el parámetro original, además las marginales son más similares en los

modelos blockNNGP con bloques regulares aún con diferentes número de bloques y bloques vecinos. Las distribuciones marginales son muy variables en los modelos NNGP, y en muchos casos no se recuperan los parámetros originales. Las estimaciones a posteriori del parámetro ϕ_2 de los modelos NNGP difieren del valor original del parámetro.

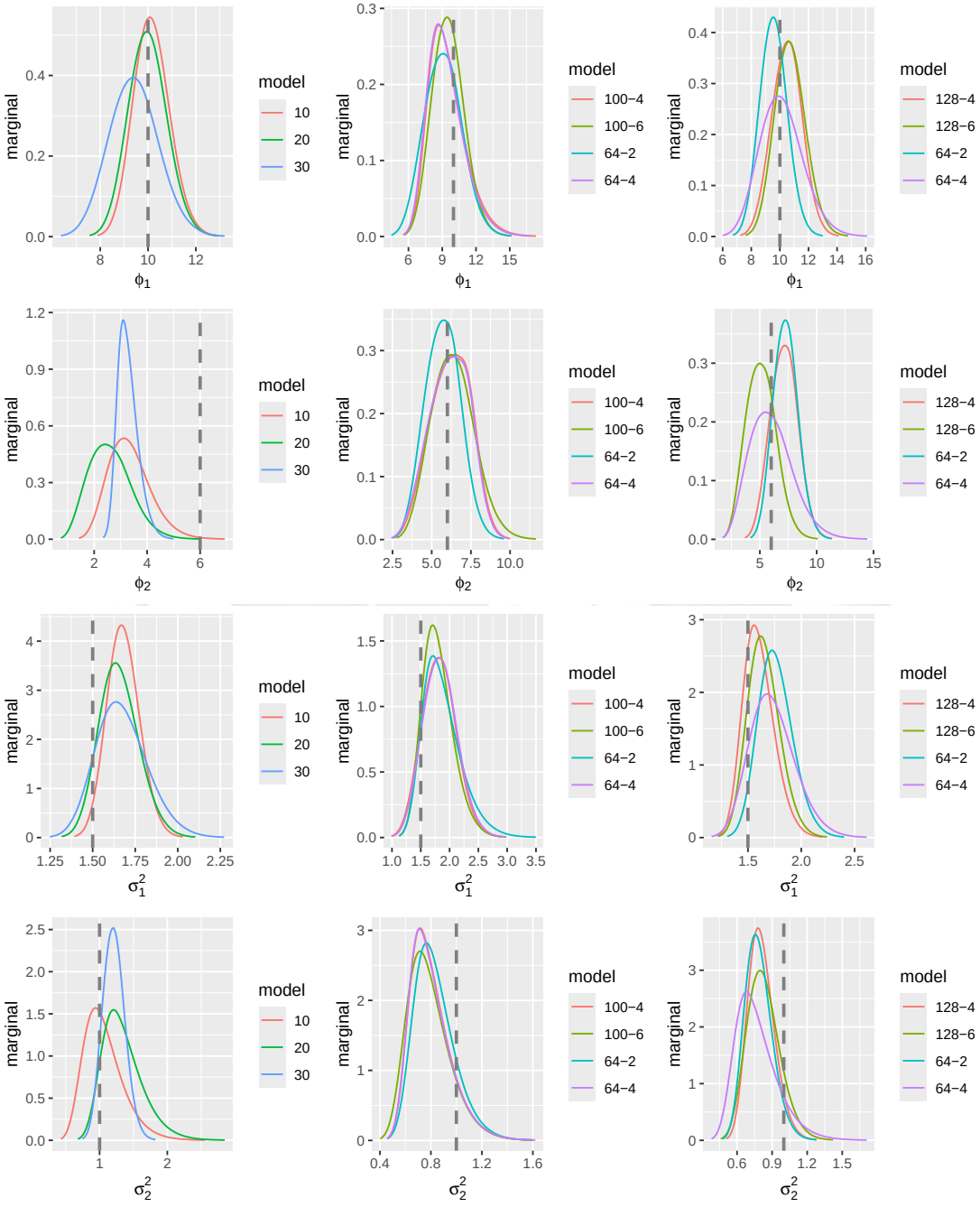

Figura 4.5: Modelo poisson multivariado - Densidades marginales a posteriori del parámetro de decaimiento espacial ϕ_k y varianza marginal σ_k^2 para los enfoques NNGP (izquierda), blockNNGP regular (centro), blockNNGP irregular (derecha). La línea vertical punteada gris representa el verdadero valor del parámetro.

Dado que el rendimiento de todos los modelos blockNNGP regular es bastante similar,

continuamos el análisis con el blockNNGP regular (64-2).

La Figura 4.6 muestra los efectos aleatorios espaciales reales (panel superior) y las estimaciones medias a posteriori (panel inferior) para la variable 1 (izquierda) y la variable 2 (derecha). Los efectos aleatorios espaciales son recuperados, pero destacamos que el efecto aleatorio espacial w_2 de la variable 2 es más suavizado que el efecto aleatorio original.

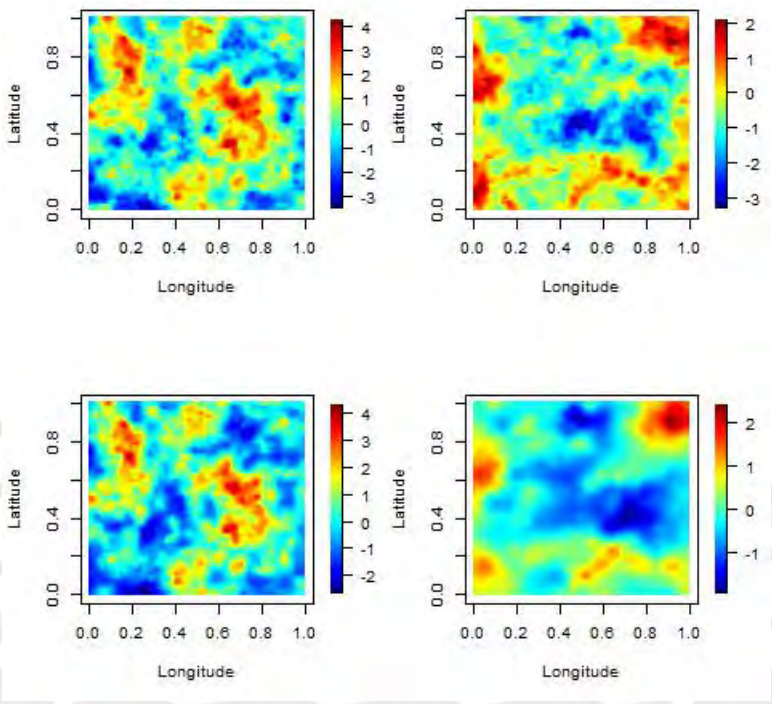

Figura 4.6: Modelo Poisson multivariado - Simulación BlockNNGP regular (M=64, nb=2). Arriba: Efectos aleatorios originales w_1 (izquierda) y w_2 (derecha). Abajo: Estimaciones de media a posteriori de efectos aleatorios.

En la figura 4.7 también se comparan los verdaderos efectos aleatorios espaciales con sus estimaciones de medias a posteriori (panel superior). En general, podemos observar que los efectos aleatorios espaciales. Sin embargo, media a posteriori del efecto aleatorio espacial w_2 de la variable 2 tiene mayor variabilidad. Esto podría explicar el patrón de suavidad mostrado para este efecto aleatorio en la Figura 4.6. Este resultado tiene sentido ya que todo el efecto aleatorio espacial de la variable de respuesta 2 tiene un efecto aleatorio espacial compartido de la variable de respuesta 1 y este efecto aleatorio espacial w_2 se usa para explicar la variable de respuesta 2, de manera similar que para los modelos gaussianos. La Figura 4.7 también compara la verdadera variable de respuesta con sus estimaciones de media a posteriori (panel inferior). Muestra que la verdadera variable de respuesta se recupera bien, con más variabilidad para la variable de respuesta 2, debido a la influencia de los efectos aleatorios espaciales.

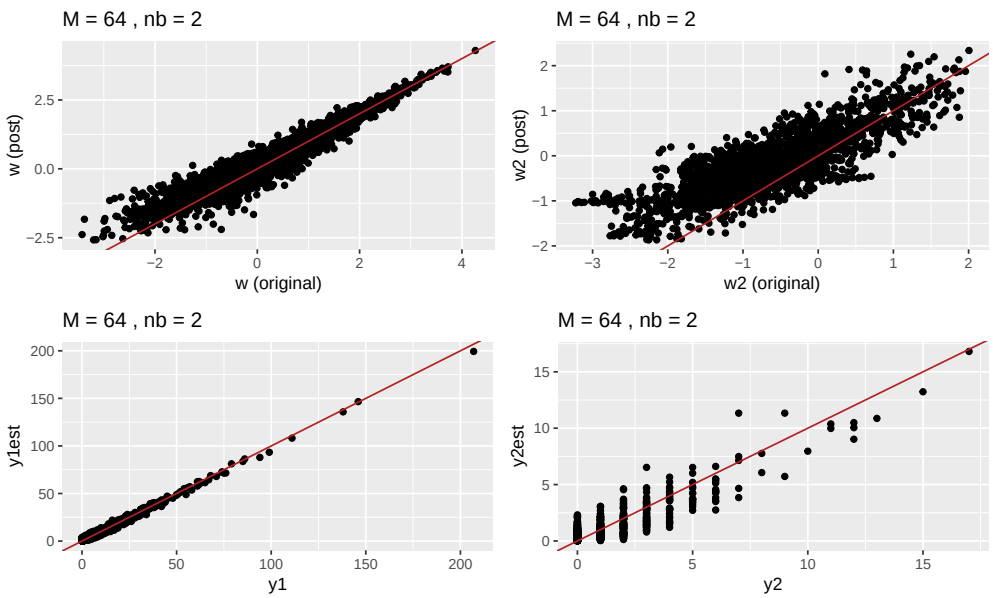

Figura 4.7: Modelo poisson multivariado - Simulación BlockNNGP regular (M=64, nb=2). Arriba: Comparación de los datos originales w_1 (izquierda), w_2 (derecha) y la media a posteriori del efecto aleatorio w_i, respectivamente. Abajo: Comparación de los datos originales y_1 (izquierda), y_2 (derecha) y la media a posteriori $y.est$, respectivamente.

En cuanto a la predicción, la Fig. 4.8 compara los valores de la variable respuesta (datos de prueba) con la predicción estimada a través del modelo blockNNGP regular (64-2). En general, podemos observar que las predicciones son razonables dada la complejidad del modelo multivariado.

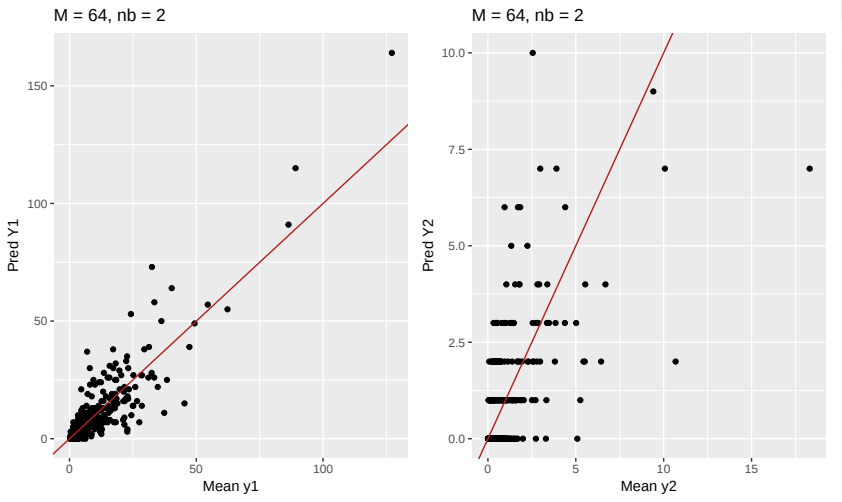

Figura 4.8: Modelo Poisson multivariado - Predicción BlockNNGP regular (M=64, nb=2). Comparación de los datos originales y_1 (izquierda), y_2 (derecha) y la media de la predicción $y.pred$, respectivamente.

Capítulo 5

Aplicación

En este capítulo se presentan dos aplicaciones para mostrar el alcance del modelo propuesto. En la primera aplicación se ajustan modelos blockNNGP gaussianos multivariados en datos de concentraciones de plomo y zinc. Estos datos fueron modelados por Palmí-Perales et al. (2023) usando ecuaciones diferenciales parciales estocásticas (SPDE). Resaltamos que esta aplicación solo ilustra el uso del modelo, pues la base de datos es pequeña. En segundo lugar, se aplican modelos de Poisson multivariados para estudiar la distribución espacial de la abundancia de dos tipos de especies de aves en América del norte.

5.1. Aplicación 1

Los datos corresponden al conjunto de datos *meuse* que proporciona las ubicaciones y mediciones de $n = 155$ de metales pesados de la capa superficial del suelo recogidas en una llanura aluvial junto al río Meuse, cerca del pueblo de Stein (Países Bajos). En particular, nos centramos en las concentraciones de plomo y zinc. Los valores de las concentraciones se consideran en la escala logarítmica. Y se asume que las concentraciones de plomo y zinc transformadas logarítmicamente (y_1 y y_2) se distribuyen normalmente. Se utiliza como covariable la distancia al río Meuse, normalizada a [0,1]. Ajustamos los modelos blockNNGP presentados en la sección 4.1, eligiendo $M = 64$ bloques irregulares y $nb = 2$ bloques vecinos (Fig. 5.1), el modelo NNGP con $nb = 10$ vecinos y el modelo SPDE.

La inferencia bayesiana completa se llevó a cabo utilizando `R-INLA`. Usando las distancias entre los locales, se asume que el rango varía entre 100 y 4440.764, luego fijamos una distribución a priori $\phi \sim U(0.0002, 0.01)$. El cuadro 5.1 muestra el WAIC, tiempo y las estimaciones de los parámetros a través de los modelos ajustados. De acuerdo al WAIC, el modelo blockNNGP tiene mejor bondad de ajuste que los modelos NNGP y SPDE. En términos de

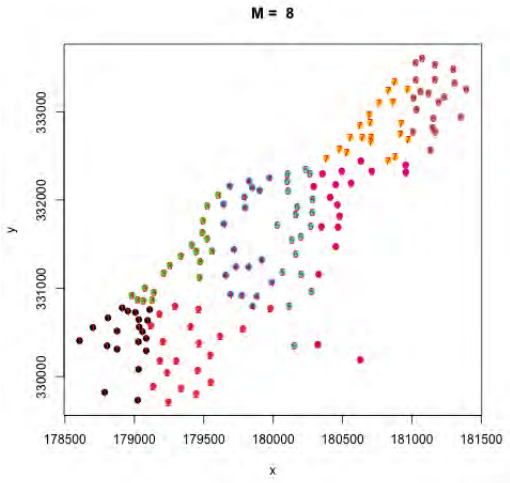

Figura 5.1: Bloques irregulares ($M = 8$) para *meuse data*.

tiempo, el SPDE corre en menos tiempo.

A partir de los resultados del modelo blockNNGP, Los resultados para los coeficientes de regresión son similares para todos los modelos. La media a posteriori del rango efectivo para la variable plomo es de aproximadamente 333 km y para la variable zinc es de aproximadamente 400 km. Mientras que partir de los resultados SPDE, la media a posteriori del rango efectivo para la variable plomo es de 242 km y para la variable zinc es de 1010 km. Las varianzas marginales para la variable concentración de plomo es similar en los tres modelos, sin embargo la varianza marginal de la variable concentración de zinc es menor según el SPDE. Luego el SPDE captura menor variabilidad espacial que los otros modelos. Finalmente el efecto pepita del modelo blockNNGP es menor para ambas variables, es decir la variabilidad no espacial restante según este modelo es menor.

Cuadro 5.1: Aplicación. Resumen de las estimaciones de las medias a posteriori de los parámetros.

	NNGP nb=10	**blockNNGP M=64,nb=2**	**SPDE**
Variable 1			
β_{01}	5.360	5.353	5.353
β_{11}	-2.299	-2.306	-2.340
σ_1^2	0.277	0.262	0.218
ϕ_1	0.006	0.006	0.0083
τ_1^2	0.00014	6.147286e-05	0.0096
Variable 2			
β_{02}	6.603	6.591	6.604
β_{12}	-2.831	-2.825	-2.919
σ_2^2	0.069	0.074	0.0404
ϕ_2	0.0052	0.0047	0.00198
τ_2^2	2.724819e-05	5.432874e-05	4.651163e-05
Tiempo (segundos)	18.1	22.3	13.4
WAIC	-1794.24	**-1813.62**	-1167.37

5.2. Aplicación 2

Analizamos los datos de abundancia durante la temporada de cría de aves Mourning Dove y American Robin. Los datos fueron recolectados por la Encuesta de Aves Reproductoras de América del Norte (BBS, The North American Breeding Bird Survey) en 2019. La encuesta se realiza en carreteras y abarca una amplia área geográfica que incluye los Estados Unidos, Alaska, Canadá y partes de México. Las rutas de la encuesta tienen una longitud de 39.4 km con 50 paradas. En cada parada, los participantes realizan un conteo de puntos de 3 minutos y registran las aves vistas o escuchadas. Dado que los recuentos no son censos completos, existe diferentes factores potenciales en el proceso de conteo afecten la probabilidad de detección.

Se analiza la abundancia de cada especie de aves en las mismas $n = 2076$ ubicaciones (Fig.5.2). Se observa que la abundancia de ambas especies es mayor en la costa este de Estados Unidos. A partir de estos mapas, la especie American Robin es más abundante en la región central y en la costa este, mientras que la especie Mourning Dove es más abundante en la costa central y norte de la costa este.

Se remarca que, a pesar de que los conteos de mourning dove y american robin no muestran una correlación lineal fuerte, como muestra el coeficiente de correlación de Pearson con un valor de 0.09, los mapas muestran que la distribución especial de ambas variables muestran cierta similitud. Además, también se estudió la evidencia de autocorrelación espacial para cada variable a través de variogramas y la distribución espacial similar a través del variograma cruzado (Figura 5.3). En primer lugar, los variogramas revelan indicios de autocorrelación espacial para la abundancia de cada especie de ave. El rango para la abundancia de la ave Mourning Dove es de aproximadamente 7 grados, mientras que el American Robin es de aproximadamente 13 grados. El cross-variograma muestra una autocorrelación espacial positiva entre la abundancia de ambas especies, con un rango asociado de 7 grados.

Específicamente, se asume que $Y_k = (Y_{1,k}, \ldots, Y_{n,k})^\top$ donde $Y_{i,1}$ representa la cantidad de aves de la especie mourning dove observadas en el local s_i e $Y_{i,2}$ representa la cantidad de aves de la especie American Robin observadas en el local s_i. Según el modeo definido en la sección 4.2 se asume que las variables aleatorias $Y_{i,k}$, condicionadas a $\mathbf{x}, \boldsymbol{\theta}$, son independientes, entonces:

$$Y_{i,1}|\mathbf{x}, \boldsymbol{\theta} \overset{ind}{\sim} \text{poisson}\,(\mu_{i,1}) \quad i = 1, \ldots, n;$$

$$Y_{i,2}|\mathbf{x}, \boldsymbol{\theta} \overset{ind}{\sim} \text{poisson}\,(\mu_{i,2}) \quad i = 1, \ldots, n.$$

Evaluamos qué efectos podrían explicar la estimación de la abundancia de estas dos especies de aves. Para ello utilizamos la data de controles rutinarios recolectados por el observador

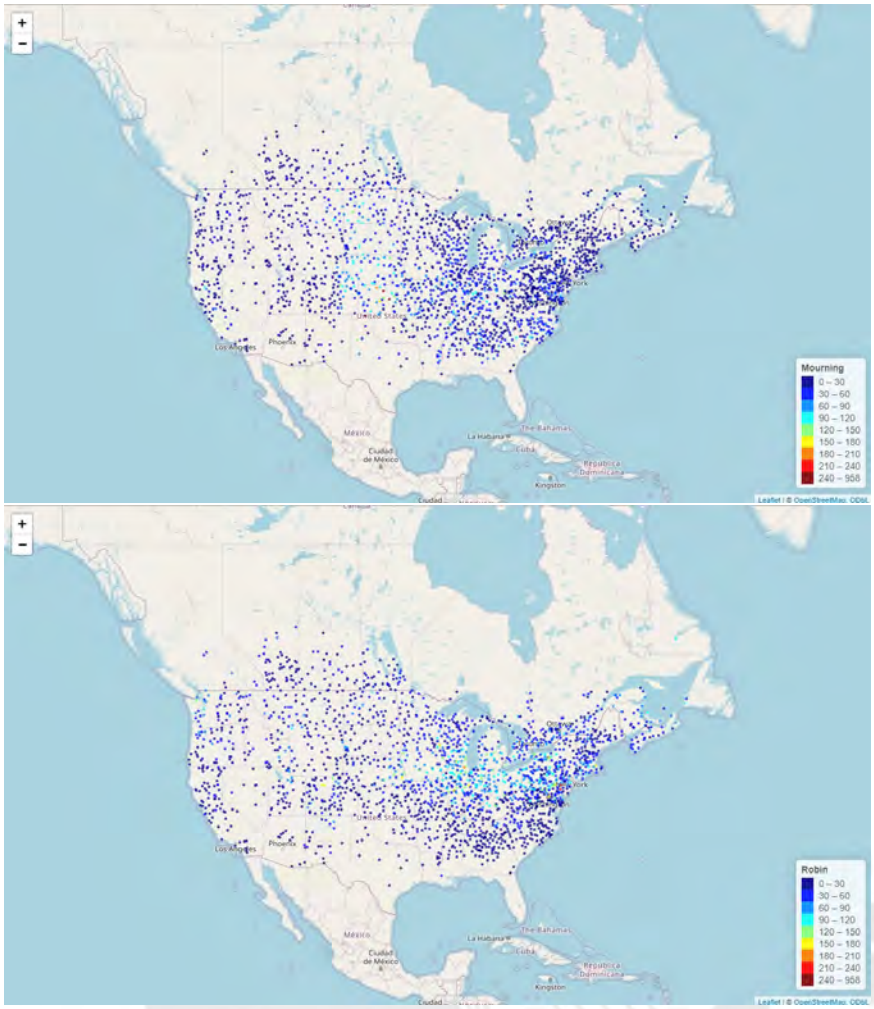

Figura 5.2: Arriba: Abundancia especie Mourning Dove (*Zenaida macroura*), Abajo: Abundancia American Robin (*Turdus migratorius*).

temperatura (Dinges et al. (2022)), el número de automóviles que pasan durante los conteos realizados en cada parada (Veh), covariable incluida en el modelo propuesto por Griffith et al. (2010). Se considera dos formas de incluir como covariable los recuentos de vehículos (un efecto simple y una transformación), en el caso de la transformación añadimos 1 a cada punto de datos y tomamos el logaritmo para incluirlo como predictor en el modelo, denotado como $log(Veh + 1)$, esta transformación estabiliza la varianza y elimina la asimetría de los datos de recuento de vehículos según Griffith et al. (2010). Asimismo evaluamos los efectos de la longitud ($Long$), latitud (Lat), precipitación ($Prec$), elevación ($Elev$) y temperatura máxima en grados Fahrenheit ($TMax$), como covariables para estimar la abundancia de cada especie.

De esta forma, las medias de las abundancias de ambas especies se modelan a través de

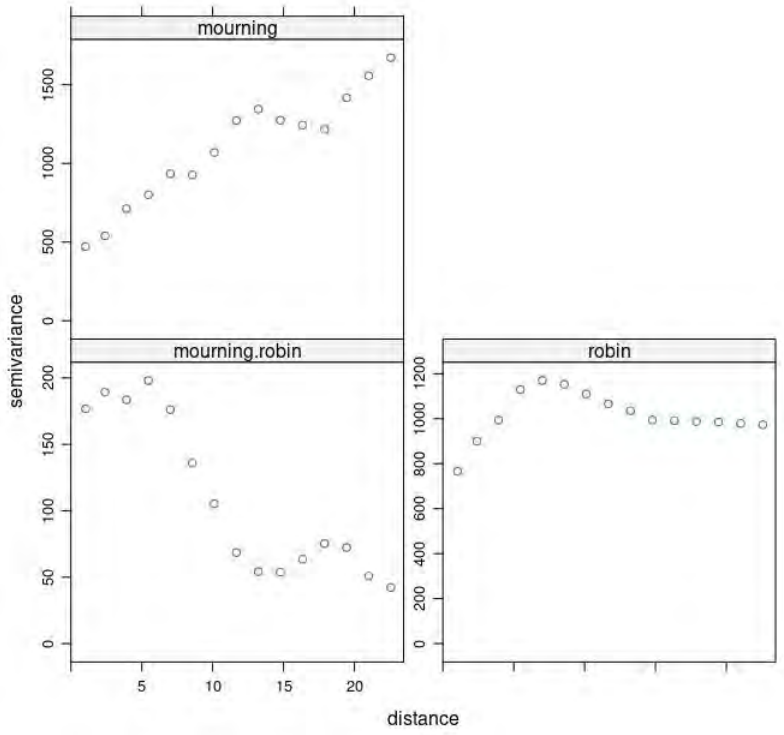

Figura 5.3: Variograma de los datos de abundancia de American Robin (panel superior izquierdo), variograma cruzado de los datos de abundancia de Mourning Dove y American Robin (panel inferior izquierdo) y variograma de los datos de abundancia de American Robin (panel inferior derecho).

los siguientes predictores lineales:

$$\log(\mu_{i,1}) = \eta_{i,1} = \boldsymbol{Z}_{i1}^{\top}\boldsymbol{\beta}^{(1)} + w_{i,1};$$

$$\log(\mu_{i,2}) = \eta_{i,2} = \boldsymbol{Z}_{i2}^{\top}\boldsymbol{\beta}^{(2)} + w_{i,1} + w_{i,2}, \tag{5.1}$$

donde $\boldsymbol{Z}_{ik}$ representa una covariable o vector de covariables en el local s_i definido para la variable respuesta abundancia de mourning dove (k=1) y abundancia de American Robin (k=2), respectivamente.

Para ajustar el modelo se asumió una a priori blockNNGP para cada especie, tal que:

$$\boldsymbol{w_k}|\boldsymbol{\theta_k} \sim \text{blockNNGP}(0, \widetilde{\boldsymbol{Q}}(\boldsymbol{\theta_k})^{-1}); k = 1, 2.$$

Se observó que la distancia máxima entre las localizaciones es aproximadamente 60 grados (un grado equivale aprox. 111.31 km), por ello utilizamos una a priori compleja penalizada para ϕ_k y σ_k, considerando $\rho_k = 2/\phi_k$, asumimos que $P(\rho_k < 60) = 0.95$ y $P(\sigma_k > 100) = 0.05$.

Ajustamos los modelos blockNNGP eligiendo $M = 22$ bloques regulares y $M = 25$ bloques irregulares con $nb = 2$ bloques vecinos 5.4). La inferencia bayesiana se llevó a cabo utilizando nuestra implementación en `R-INLA`.

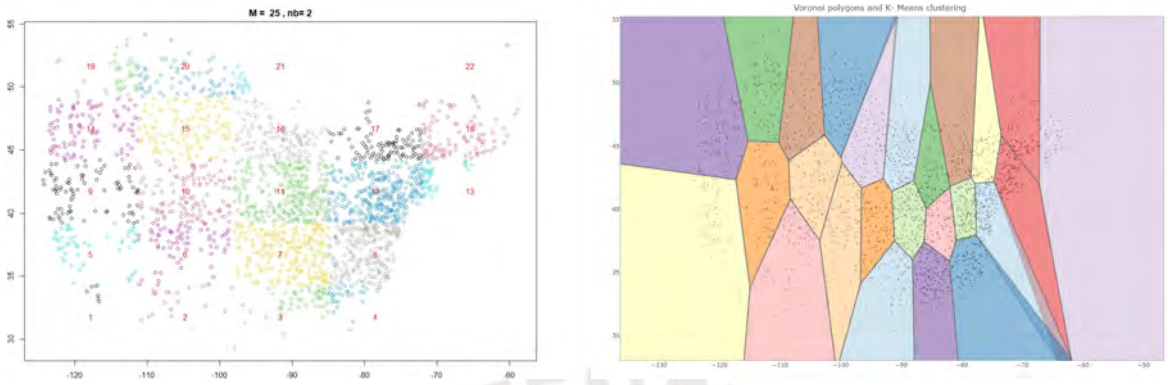

Figura 5.4: Ejemplo de configuración de bloques usados en los modelos blockNNGP. Izquierda: Bloques regulares (M = 22). Derecha: bloques irregulares (M = 25).

A través de un análisis preliminar, las covariables número de automóviles (Veh), precipitación ($Prec$) y elevación ($Elev$) resultaron no significativas. Al ajustar los modelos blockNNGP sólo la longitud, latitud, $\log(Veh + 1)$ y temperatura máxima fueron variables significativas para estimar la densidad poblacional de cada especie. Para los mejores modelos, se consideraron las siguientes covariables:

Modelo I : $Lat_1, Lat_2, \log(Veh + 1)_1, \log(Veh + 1)_2$

Modelo II : $Long_1, Lat_1, Lat_2, \log(Veh + 1)_1, \log(Veh + 1)_2$

Modelo III : $Long_1, Lat_1, Lat_2, TMax_1, \log(Veh + 1)_1, \log(Veh + 1)_2,$

donde el subíndice 1 o 2, indican que la covariable ertenece al predictor $\eta_{i,1}$ o $\eta_{i,2}$, respectivamente, definidos en la ecuación (5.1). Las métricas evaluadas incluyen WAIC, LPML, MSE (para Y_1 e Y_2) y el tiempo de ejecución. Los resultados se muestran en el cuadro 5.2. El modelo de bloques irregulares ($M = 25$ con $nb = 2$) tiene menor tiempo de ejecución y con desempeño similar en ambos enfoques en la mayoría de los casos. Considerando ello, el modelo III con bloques irregulares tiene uno de los valores más bajos para WAIC y el mayor LPML, lo que indica una mejor bondad de ajuste y predictibilidad en comparación con los otros modelos. El modelo III tiene un tiempo de ejecución bajo, y los MSE para Y_1 e Y_2 son similares en todos los modelos siendo ligeramente mejores para el modelo II con bloques regulares. En general, el modelo III podría considerarse bueno en términos de ajuste del modelo además que es eficiente en términos de recursos computacionales. En función de las observaciones anteriores, el **modelo III** podría considerarse una buena elección.

Cuadro 5.2: Aplicación 2. Resumen de métricas de desempeño por modelo con blockNNGP regulares e irregulares.

	Block-NNGP Regular (M = 22, nb = 2)					Block-NNGP Irregular (M = 25, nb = 2)				
	WAIC	LPML	MSE Y_1	MSE Y_2	Tiempo (Seg.)	WAIC	LPML	MSE Y_1	MSE Y_2	Tiempo (Seg.)
Modelo I	28364.434	-39035.176	9.036	3.424	5895	28480.877	-38256.550	**10.042**	**3.705**	2311
Modelo II	28356.284	-39075.420	**8.945**	**3.422**	3634	28481.979	-38220.802	10.049	3.726	2443
Modelo III	**28353.904**	**-38871.534**	9.031	3.450	4546	**28472.484**	**-38096.521**	10.060	3.753	2360

A continuación se usaron las covariables del **modelo III** bajo diferentes enfoques ajustando el modelo SPDE, blockNNGP regular ($M = 25, 49, 100$) e irregulares ($M = 25, 50, 100$) con $nb = 2, 4, 6$ bloques vecinos, y el modelo NNGP con $nb = 10, 30, 60$ número de vecinos. Las métricas del desempeño de los modelos ajustados se muestran en el cuadro 5.3.

Cuadro 5.3: Aplicación 2. Resultados de los modelos **SPDE**, **NNGP**, **blockNNGP - regular** y **blockNNGP - irregular**. Resumen de las métricas de desempeño y tiempo (en segundos).

Modelos	Bloques	Número Vecinos	WAIC	LMPL	MSE Y_1	MSE Y_2	Tiempo (seg.)	Observaciones (por bloque)
SPDE	—	—	55994.26	-29663.68	373.588	242.298	39	—
NNGP	—	nb = 10	28489.40	-37653.90	10.195	3.824	726	—
	—	nb = 30	28478.87	-37619.45	10.211	3.806	1600	—
	—	nb = 60	28515.24	-37345.61	10.570	3.869	4219	—
BlockNNGP Regular	M = 25	nb = 2	**28353.904**	-38871.534	**9.031**	**3.450**	4546	94
	M = 25	nb = 4	28513.16	**-37199.59**	10.574	3.857	12186	94
	M = 49	nb = 2	28480.76	-37657.86	10.213	3.83	1487	51
	M = 49	nb = 4	28483.6	-37572.1	10.284	3.81	6037	51
	M = 100	nb = 2	**28474.21**	-37825.68	**10.019**	**3.794**	1316	27
	M = 100	nb = 4	28482.01	-37584.37	10.251	3.821	6260	27
	M = 100	nb = 6	28494.5	**-37475.45**	10.389	3.845	4691	27
BlockNNGP Irregular	M = 25	nb = 2	28472.48	**-38096.52**	10.060	3.753	2360	83
	M = 25	nb = 4	**28467.76**	**-38191.65**	10.053	3.748	6680	83
	M = 50	nb = 2	28470.42	-38637.46	9.793	3.676	3146	42
	M = 50	nb = 4	28474.73	-38341.89	9.953	3.728	3989	42
	M = 100	nb = 2	28483.56	-39276.29	**9.625**	**3.566**	2094	21
	M = 100	nb = 4	28475.68	-39062.07	9.659	3.582	1976	21
	M = 100	nb = 6	**28444.41**	-38942.78	**9.529**	**3.572**	2278	21

Según el criterio de WAIC el NNGP y blockNNGP tienen mejor bondad de ajuste que el SPDE, siendo ligeramente mejor el modelo blockNNGP. Según el LPML, el SPDE es el mejor modelo, sin embargo tiene los peores MSE. En términos del LPML el NNGP es similar o mejor que el blockNNGP. Según el MSE, los blockNNGP estiman mejor que los modelos NNGP. El modelo más rápido es el SPDE. Los modelos blockNNGP demoran más que el NNGP con pocas observaciones vecinas, pero en muchos casos es similar o más rápido que los modelos NNGP con 60 observaciones. Los modelos blockNNGP con bloques irregulares son más rápidos que usando bloques regulares. En general los blockNNGP tienen buen ajuste y tiempos de ejecución, siendo mejores en muchos casos que los otros modelos NNGP y SPDE. EL modelo blockNNGP es ligeramente mejor cuando se usa $M = 25, 100$ con $nb = 2$. Los resultados de las medias a posteriori e IC al 95 % se presentan en el cuadro 5.4.

Cuadro 5.4: Aplicación 2. Resultados de los enfoques **NNGP**, **BlockNNGP regular** y **BlockNNGP Irregular** para el Modelo III. Resumen de las estimaciones de las medias a posteriori e intervalos de credibilidad al 95 % de los parámetros.

Modelos	SPDE	NNGP			Block-NNGP Regular							Block-NNGP Irregular						
		nb=10	nb=30	nb=60	M=25 nb=2	M=25 nb=4	M=49 nb=2	M=49 nb=4	M=100 nb=2	M=100 nb=4	M=100 nb=6	M=25 nb=2	M=25 nb=4	M=50 nb=2	M=50 nb=4	M=100 nb=2	M=100 nb=4	M=100 nb=6
American Robin																		
β_{01}	0.650	0.300	-0.115	-2.388	0.905	-0.297	0.251	-0.315	0.580	-0.156	-0.565	0.253	0.067	1.156	0.951	1.250	1.555	1.400
Lím. Inf.	-0.676	-1.927	-2.753	-25.313	-0.311	-3.705	-2.039	-3.305	-1.465	-2.816	-4.575	-1.622	-1.824	-0.379	-0.667	-0.057	0.189	0.001
Lím. Sup.	1.955	2.490	2.405	20.506	2.096	2.717	2.470	2.505	2.584	2.368	3.286	2.048	1.888	2.675	2.537	2.544	2.914	2.781
β_{11} $Long_1$	0.012	0.014	0.012	0.005	0.015	0.011	0.013	0.011	0.014	0.012	0.010	0.015	0.012	0.019	0.019	0.018	0.019	0.019
Lím. Inf.	0.005	0.000	-0.005	-0.166	0.007	-0.011	-0.001	-0.007	0.001	-0.005	-0.016	0.003	0.001	0.010	0.009	0.010	0.011	0.010
Lím. Sup.	0.020	0.027	0.028	0.176	0.022	0.031	0.027	0.029	0.026	0.028	0.036	0.026	0.023	0.027	0.028	0.025	0.027	0.027
β_{21} Lat_1	0.088	0.098	0.102	0.109	0.087	0.104	0.098	0.106	0.092	0.103	0.108	0.103	0.099	0.090	0.094	0.087	0.084	0.086
Lím. Inf.	0.063	0.058	0.057	-0.116	0.065	0.052	0.058	0.056	0.056	0.059	0.042	0.069	0.065	0.061	0.063	0.062	0.057	0.059
Lím. Sup.	0.113	0.137	0.148	0.334	0.110	0.161	0.138	0.158	0.129	0.150	0.176	0.138	0.134	0.120	0.125	0.113	0.111	0.114
β_{31} $TMax_1$	-0.007	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.008	-0.007	-0.008	-0.008	-0.008
Lím. Inf.	-0.008	-0.010	-0.010	-0.010	-0.011	-0.010	-0.011	-0.011	-0.011	-0.010	-0.010	-0.011	-0.010	-0.010	-0.010	-0.010	-0.011	-0.011
Lím. Sup.	-0.006	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.005	-0.006	-0.005
β_{41} $\log(Veh+1)_1$	0.069	0.089	0.088	0.087	0.093	0.087	0.089	0.088	0.095	0.088	0.087	0.091	0.090	0.089	0.091	0.088	0.090	0.092
Lím. Inf.	0.062	0.070	0.069	0.068	0.074	0.069	0.071	0.069	0.076	0.069	0.069	0.072	0.071	0.070	0.072	0.069	0.071	0.073
Lím. Sup.	0.076	0.108	0.106	0.105	0.113	0.105	0.108	0.106	0.114	0.106	0.106	0.110	0.108	0.108	0.110	0.107	0.109	0.111
σ_1^2	1.152	1.348	1.461	52.942	0.860	2.151	1.410	1.513	1.341	1.489	3.087	1.292	1.277	1.207	1.268	1.168	1.245	1.221
Lím. Inf.	0.986	1.099	1.106	32.184	0.766	0.581	1.125	0.935	1.093	1.102	0.057	1.063	1.006	1.031	1.099	0.995	1.076	0.977
Lím. Sup.	1.338	1.680	1.919	84.364	0.960	6.492	1.815	2.226	1.688	1.980	16.345	1.581	1.553	1.422	1.539	1.347	1.451	1.461
ϕ_1	1.044	0.607	0.564	0.015	1.083	0.523	0.574	0.581	0.600	0.550	1.583	0.630	0.641	0.684	0.632	0.709	0.645	0.702
Lím. Inf.	0.857	0.471	0.412	0.009	0.947	0.117	0.428	0.372	0.459	0.396	0.028	0.500	0.511	0.566	0.496	0.602	0.536	0.569
Lím. Sup.	1.264	0.747	0.741	0.024	1.252	1.315	0.719	0.902	0.737	0.740	9.967	0.768	0.821	0.806	0.736	0.843	0.754	0.884
Mourning Dove																		
β_{02}	6.556	7.517	7.933	6.520	6.589	8.634	7.406	7.965	7.122	8.064	8.370	7.720	6.691	6.680	6.967	6.238	6.448	6.307
Lím. Inf.	4.639	3.312	1.882	-18.624	4.421	-1.595	1.664	2.328	3.136	-0.864	-0.008	4.482	3.348	4.220	4.355	4.209	4.247	4.061
Lím. Sup.	8.500	11.783	14.244	31.600	8.809	18.864	13.198	13.763	11.386	17.306	16.665	11.113	10.104	9.184	9.636	8.288	8.665	8.572
β_{12} Lat_2	-0.097	-0.127	-0.139	-0.159	-0.099	-0.164	-0.125	-0.139	-0.115	-0.145	-0.156	-0.130	-0.107	-0.101	-0.107	-0.087	-0.094	-0.089
Lím. Inf.	-0.143	-0.222	-0.278	-0.482	-0.151	-0.385	-0.252	-0.272	-0.211	-0.335	-0.341	-0.210	-0.188	-0.160	-0.170	-0.135	-0.146	-0.143
Lím. Sup.	-0.052	-0.033	-0.009	0.164	-0.049	0.054	-0.001	-0.011	-0.025	0.026	0.029	-0.054	-0.029	-0.043	-0.045	-0.039	-0.042	-0.036
β_{22} $\log(Veh+1)_2$	0.011	0.032	0.032	0.031	0.033	0.031	0.033	0.032	0.037	0.032	0.031	0.034	0.031	0.031	0.033	0.028	0.029	0.031
Lím. Inf.	0.003	0.006	0.005	0.005	0.005	0.005	0.007	0.005	0.011	0.006	0.005	0.008	0.005	0.004	0.007	0.001	0.002	0.004
Lím. Sup.	0.019	0.058	0.058	0.057	0.060	0.057	0.059	0.058	0.064	0.058	0.057	0.061	0.057	0.058	0.060	0.056	0.056	0.058
σ_2^2	2.714	3.359	5.646	105.774	1.659	24.422	5.311	5.007	3.511	14.213	11.988	2.837	3.041	2.274	2.428	2.196	2.261	2.375
Lím. Inf.	2.291	2.697	3.038	64.054	1.345	5.926	3.311	3.278	2.301	3.122	1.965	2.006	2.087	1.693	2.039	1.811	1.880	1.973
Lím. Sup.	3.1977	4.276	11.352	169.935	1.937	91.405	8.615	7.985	5.534	57.237	44.382	3.846	4.322	2.851	2.945	2.630	2.710	2.823
ϕ_2	1.039	0.256	0.169	0.008	0.603	0.057	0.167	0.186	0.252	0.106	0.131	0.311	0.292	0.391	0.353	0.404	0.388	0.363
Lím. Inf.	0.842	0.199	0.072	0.005	0.500	0.009	0.095	0.108	0.148	0.014	0.019	0.220	0.195	0.300	0.283	0.329	0.316	0.298
Lím. Sup.	1.274	0.314	0.283	0.013	0.770	0.140	0.255	0.272	0.371	0.268	0.420	0.431	0.417	0.530	0.421	0.493	0.466	0.438

El análisis de las estimaciones a posteriori del modelo blockNNGP regular (100-2) por tener uno de los mejores desempeño y ser uno de los modelos más rápidos. Para la especie American Robin, para el intercepto β_{01} el promedio de la abundancia de aves de esta especie cuando todas las covariables y efectos aleatorios son cero es de $\exp(0.58) = 1.79$. El efecto de las covariables longitud (β_{11}), latitud (β_{21}), TMax (β_{31}) y Log(Veh+1) (β_{41}) según sus coeficientes asociados indican cómo cada una de estas afectan al promedio de la especie. Por ejemplo, por cada unidad de aumento de un grado en la longitud la abundancia de aves de la especie American Robin aumenta en un $(\exp(0.014) - 1)100\,\% = 1.41\,\%$. El efecto de la latitud indica que la media de abundancia de aves American Robin aumenta en un $(\exp(0.092) - 1)100\,\% = 9.64\,\%$ por cada grado en que se aumenta en la latitud. Para la especie Mourning Dove, los parámetros se interpretan de manera similar a los de American Robin. Por ejemplo, el promedio de la abundancia de Mourning dove cuando todas las covariables y efectos aleatorios son cero es $\exp(7.122) = 1238.926$. Cuando el número de vehículos aumenta, aumenta la abundancia de aves de ambas especies. En particular, cuando la covariable $Log(Veh+1)$ aumenta en una unidad, la cantidad de aves Mourning Dove aumenta en $(\exp(0.092) - 1)100\,\% = 3.76\,\%$, mientras que la cantidad de aves American Robin aumenta en $(\exp(0.095) - 1)100\,\% = 9.97\,\%$.

Por otro lado, la media a posteriori del rango efectivo ($2/\phi_k$) implica que la cantidad de aves Mourning Dove y American Robin tienen una autocorrelación espacial significativa compartida; hasta una distancia igual a 3.33 grados es decir hasta 369.963 km para la especie American Robin. Mientras que la cantidad de aves Mourning Dove tienen una autocorrelación espacial específica significativa hasta una distancia igual a 7.936 grados es decir hasta 881.689 km. Esto significa que las aves american robin están ligeramente más agregadas que las mourning dove. El rango de dependencia espacial es de corto alcance para la abundancia de American Robin, lo que podría indicar que la dependencia espacial es más local. La varianza espacial es considerable para ambas especies, lo que indica que existe una variación espacial significativa en la distribución espacial de la abundancia de aves. En particular, hay mayor varianza marginal espacial par la abundancia de aves American Robin. En efecto hay una observación de abundancia de aves mucho mayor que el resto de valores como se verá más adelante.

En la figura 5.5 se muestran las estimaciones de medias a posteriori de efectos aleatorios espaciales (arriba) y variables de respuesta (abajo).

Asimismo, podemos observar en la Fig. 5.6 donde se compara la variable de respuesta verdadera con sus estimaciones de media a posteriori y nos muestra que la variable de respuesta

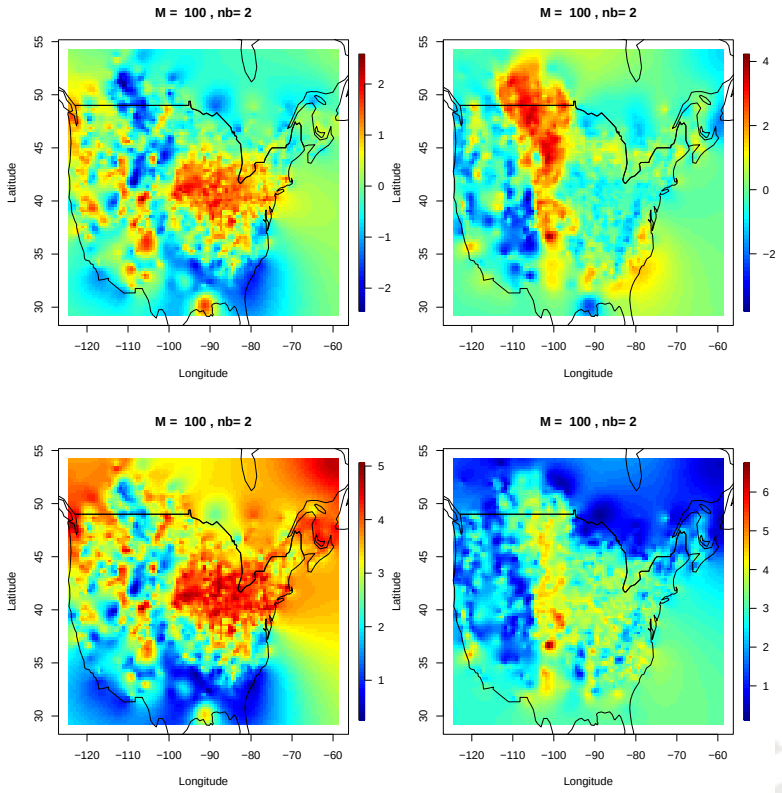

Figura 5.5: BlockNNGP Regular ($M = 100$, $nb = 2$) para datos de BBS con el modelo III. Arriba: Estimaciones de media a posteriori de efectos aleatorios w_1 (izquierda) y w_2 (derecha). Abajo: Estimaciones de las medias a posteriori y_1 (izquierda) e y_2 (derecha).

se recupera bien.

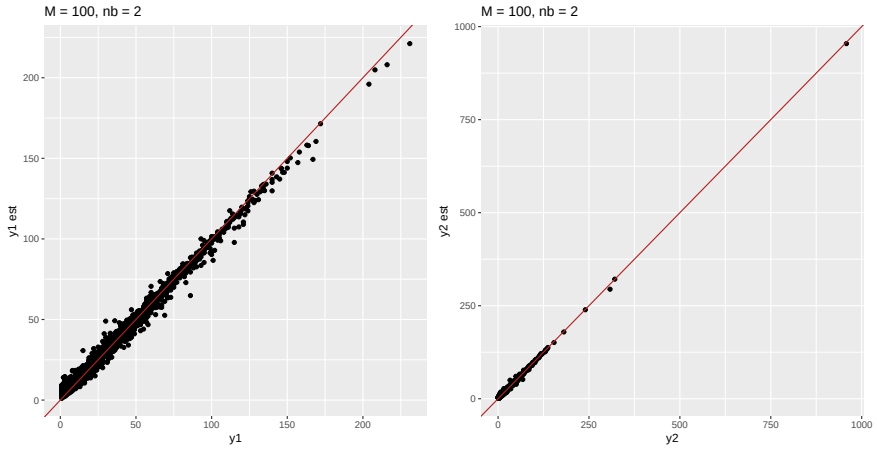

Figura 5.6: Block-NNGP Regular ($M = 100$, $nb = 2$) para datos de BBS con el modelo III. Comparación de los datos originales y_1 (izquierda), y_2 (derecha) y la media a posteriori $y.est$, respectivamente.

5.2.1. Capacidad predictiva

En cuanto a la predicción, se ajustaron los modelos, usando 500 datos de validación y 2000 datos para la estimación. Cada modelo fue evaluado con las diferentes configuraciones de bloques y número de vecinos, se reportaron métricas de capacidad predictiva (MSP) como se detalla en la tabla 5.5. En la Fig. 5.7 compara los valores de la variable respuesta (datos de prueba no usados para la estimación) con la predicción estimada a través del modelo blockNNGP irregular (25-2). En general, podemos observar que las predicciones son razonables dada la complejidad del modelo multivariado, esto quiere decir que el modelo logra hacer predicciones que se alinean adecuadamente con los datos de prueba.

Cuadro 5.5: Aplicación 2. Resultados de los modelos **NNGP**, **blockNNGP - regular** y **blockNNGP - irregular**. Resumen de las métricas de capacidad predictiva.

Modelos	Bloques	Número Vecinos	MSP Y_1	MSP Y_2	RMSP Y_1	RMSP Y_2
NNGP	—	nb = 10	607.967	444.829	24.657	24.091
	—	nb = 30	618.541	438.159	24.870	20.932
	—	nb = 60	611.886	437.5	24.736	20.917
BlockNNGP Regular	M = 25	nb = 2	**603.667**	602.323	24.57	24.542
	M = 25	nb = 4	**610.316**	580.442	24.705	24.092
	M = 49	nb = 2	621.777	**507.937**	24.935	22.537
	M = 49	nb = 4	617.472	**485.206**	24.849	22.027
	M = 100	nb = 2	638.581	581.506	25.272	24.114
	M = 100	nb = 4	630.962	577.631	25.119	24.034
	M = 100	nb = 6	630.049	581.991	24.101	24.124
BlockNNGP Irregular	M = 25	nb = 2	**620.32**	**457.255**	24.906	21.384
	M = 25	nb = 4	**618.28**	547.149	24.865	23.391
	M = 50	nb = 2	622.9	473.796	24.958	21.768
	M = 50	nb = 4	634.88	**459.16**	25.197	21.428
	M = 100	nb = 2	635.623	484.136	25.212	22.003
	M = 100	nb = 4	624.424	561.356	24.988	23.693
	M = 100	nb = 6	645.099	521.244	25.399	22.831

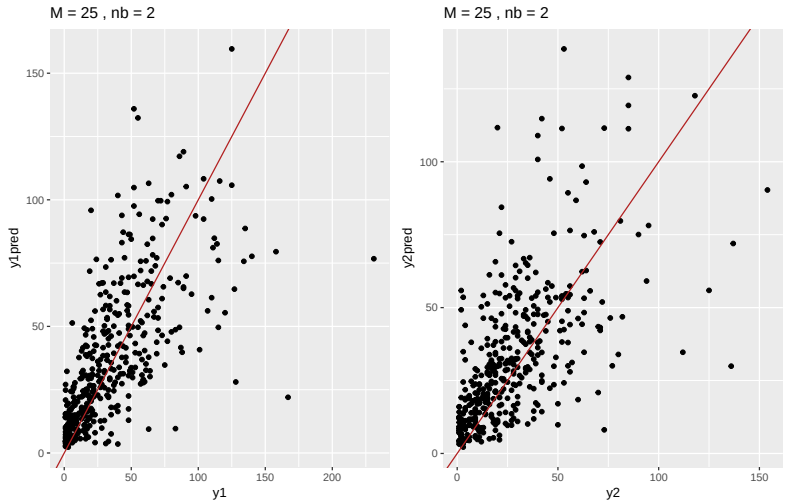

Figura 5.7: Block-NNGP Irregular ($M = 25$, $nb = 2$) para datos de BBS con el modelo III. Comparación de los datos originales y_1 (izquierda), y_2 (derecha) y la media a posteriori $y.est$, respectivamente.

Capítulo 6

Conclusiones

Esta investigación presenta una introducción al modelo blockNNGP multivariado para datos geoestadísticos utilizando INLA, aprovechando que el blockNNGP es un GMRF se aproximó el proceso gaussiano (GP) original. En particular, el blockNNGP multivariado ofrece un mejor rendimiento que el NNGP cuando el rango especial es amplio (largas distancias). Aunque no existe un número fijo de bloques y vecinos óptimos, se puede determinar una configuración adecuada considerando el tiempo de ejecución y la eficiencia estadística.

Los modelos se implementaron en el R-INLA, lo cual proporciona un gran beneficio computacional, permitiendo obtener estimaciones precisas en minutos. La combinación de INLA y blockNNGP multivariado facilita una inferencia rápida para observaciones gaussianas y no gaussianas, como se discute en las secciones 4.1 y 4.2 respectivamente. Se considera que el blockNNGP multivariado dentro de INLA es una alternativa eficaz para modelar grandes conjuntos de datos, preservando las características de los datos y siendo fácil de implementar para los profesionales. Mediante datos simulados y aplicaciones prácticas, hemos demostrado que el blockNNGP proporciona una aproximación precisa con poca complejidad computacional y tiempos reducidos para modelos geostadísticos.

El estudio de especies de aves es un excelente indicador de la biodiversidad y la productividad. En esta investigación, nos centramos en las especies Mourning Dove y American Robin, las cuales nos permiten identificar amenazas como el calentamiento global y los cambios en el uso de la tierra.

Los resultados muestran cómo diferentes configuraciones de modelos geo-estadísticos pueden afectar la capacidad predictiva en datos de conteo con distribución de Poisson, proporcionando información valiosa para la selección y evaluación de modelos en aplicaciones.

Bibliografía

Banerjee, S., Carlin, B. P. y Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 2 edn, CRC Press.

Banerjee, S., Gelfand, A. E., Finley, A. O. y Sang, H. (2008). Gaussian Predictive Process Models for Large Spatial Data Sets, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **70**(4): 825–848.
URL: *https://academic.oup.com/jrsssb/article/70/4/825/7109503*

Chiles, J.-P. y Delfiner, P. (2009). *Geostatistics: Modeling Spatial Uncertainty*, John Wiley & Sons.

Cortés, J. L., Bautista, F., Delgado, C., Quintana, P., Aguilar, D., García, A., Figueroa, C. y Gogichaishvili, A. (2016). Spatial distribution of heavy metals in urban dust from Ensenada, Baja California, Mexico, *Revista Chapingo Serie Ciencias Forestales y del Ambiente* **23**(1): 47–60.
URL: *https://revistas.chapingo.mx/forestales/article/view/r.rchscfa.2016.02.005*

Datta, A., Banerjee, S., Finley, A. O. y Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets, *Journal of the American Statistical Association* **111**(514): 800–812.
URL: *https://www.tandfonline.com/doi/full/10.1080/01621459.2015.1044091*

Dinges, A. J., Szymanski, M. L. y Parent, C. J. (2022). Effects of weather and landscape use on mourning dove population trends in North Dakota, *Wildlife Society Bulletin* **46**(4): e1346.
URL: *https://wildlife.onlinelibrary.wiley.com/doi/10.1002/wsb.1346*

Fabijańczyk, P., Zawadzki, J. y Wojtkowska, M. (2016). Geostatistical study of spatial correlations of lead and zinc concentration in urban reservoir. Study case Czerniakowskie Lake, Warsaw, Poland, *Open Geosciences* **8**(1).
URL: *https://www.degruyter.com/document/doi/10.1515/geo-2016-0043/html*

Fuglstad, G.-A., Simpson, D., Lindgren, F. y Rue, H. (2019). Constructing Priors that Penalize the Complexity of Gaussian Random Fields, *Journal of the American Statistical Association* **114**(525): 445–452.
URL: *https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1415907*

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. y Rubin, D. B. (2014). Bayesian Data Analysis, *CRC Press, Taylor & Francis Group.* **3**.
URL: *http://www.stat.columbia.edu/ gelman/book/BDA3.pdf*

Gneiting, T., Kleiber, W. y Schlather, M. (2010). Matérn Cross-Covariance Functions for Multivariate Random Fields, *Journal of the American Statistical Association* **105**(491): 1167–1177.
URL: *https://www.tandfonline.com/doi/full/10.1198/jasa.2010.tm09420*

Green, P. J. y Sibson, R. (1978). Computing Dirichlet Tessellations in the Plane, *The Computer Journal* **21**(2): 168–173.
URL: *https://academic.oup.com/comjnl/article-lookup/doi/10.1093/comjnl/21.2.168*

Grekousis, G. (2020). *Spatial Analysis Methods and Practice: Describe – Explore – Explain through GIS*, 1 edn, Cambridge University Press.
URL: *https://www.cambridge.org/core/product/identifier/9781108614528/type/book*

Griffith, E. H., Sauer, J. R. y Royle, J. A. (2010). Traffic Effects on Bird Counts on North American Breeding Bird Survey Routes, *The Auk* **127**(2): 387–393.
URL: *https://academic.oup.com/auk/article/127/2/387-393/5148370*

Handcock, M. S. y Stein, M. L. (1993). A Bayesian Analysis of Kriging, *Technometrics* **35**(4): 403–410.
URL: *http://www.tandfonline.com/doi/abs/10.1080/00401706.1993.10485354*

Handcock, M. S. y Wallis, J. R. (1994). An Approach to Statistical Spatial-Temporal Modeling of Meteorological Fields, *Journal of the American Statistical Association* **89**(426): 368–378.
URL: *https://www.tandfonline.com/doi/full/10.1080/01621459.1994.10476754*

Lindgren, F., Rue, H. y Lindström, J. (2011). An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **73**(4): 423–498.
URL: *https://academic.oup.com/jrsssb/article/73/4/423/7034732*

Martino, S. y Riebler, A. (2019). Integrated Nested Laplace Approximations (INLA). arXiv:1907.01248 [stat].
URL: *http://arxiv.org/abs/1907.01248*

Matheron, G. (1963). Principles of geostatistics, *Economic Geology* **58**(8): 1246–1266.

Matérn, B. (1960). *Spatial Variation*, Vol. 36 of *Lecture Notes in Statistics*, Springer New York, New York, NY.
URL: *http://link.springer.com/10.1007/978-1-4615-7892-5*

Myers, D. E. (1983). Estimation of linear combinations and co-kriging, *Journal of the International Association for Mathematical Geology* **15**(2): 633–637.

Palmí-Perales, F., Gómez-Rubio, V., Bivand, R. S., Cameletti, M. y Rue, H. (2023). Bayesian inference for multivariate spatial models with inla, *The R Journal* **15**: 172–190. https://doi.org/10.32614/RJ-2023-068.

Quiroz, Z. C., Prates, M. O., Dey, D. K. y Rue, H. (2023). Fast Bayesian inference of block Nearest Neighbor Gaussian models for large data, *Statistics and Computing* **33**(2): 54.
URL: *https://link.springer.com/10.1007/s11222-023-10227-1*

Ramos, L., González, M. J. y Hernández, L. M. (1999). Sequential Extraction of Copper, Lead, Cadmium, and Zinc in Sediments from Ebro River (Spain): Relationship with Levels Detected in Earthworms, *Bulletin of Environmental Contamination and Toxicology* **62**(3): 301–308.
URL: *http://link.springer.com/10.1007/s001289900874*

Rue, H. y Held, L. (2005). *Gaussian Markov random fields: theory and applications*, Monographs on statistics and applied probability, Chapman & Hall/CRC, Boca Raton.

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.
URL: *https://academic.oup.com/jrsssb/article/71/2/319/7092907*

Rue, H., Riebler, A., Sørbye, S. H., Illian, J. B., Simpson, D. P. y Lindgren, F. K. (2017). Bayesian Computing with INLA: A Review, *Annual Review of Statistics and Its Application* **4**(1): 395–421.
URL: *https://www.annualreviews.org/doi/10.1146/annurev-statistics-060116-054045*

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region, *Economic Geography* **46**: 234.
URL: *https://www.jstor.org/stable/143141?origin=crossref*

Waller, L. A. y Gotway, C. A. (2004). *Applied spatial statistics for public health data*, Wiley series in probability and statistics, John Wiley & Sons, Hoboken, N.J.