PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Aproximación de Laplace de modelos geoestadísticos con distribución normal independiente y covarianza tapering

Tesis par optar por el grado académico de Maestro es Estadística

que presenta:

Yuri Vladimir Plasencia Lapa

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 *Aproximación de Laplace de modelos geoestadísticos con distribución normal independiente y covarianza tapering*, del autor Yuri Vladimir Plasencia Lapa, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 13 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 14/08/2024.
- 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, 14 de agosto de 2024

<table>
<tr><td colspan="2">Apellidos y nombres de la asesora:
Quiroz Cornejo Zaida Jesús</td></tr>
<tr><td>DNI: 43704124</td><td rowspan="2">Firma:</td></tr>
<tr><td>ORCID: https://orcid.org/0000-0003-3821-0815</td></tr>
</table>

Dedicatoria

A mis padres por su apoyo incondicional en todo momento de mi vida y por darme ejemplo de paciencia, perseverancia, resiliencia y amor por el prójimo.

Agradecimentos

Expreso mis agradecimientos a la profesora Zaida Quiroz por la orientación, paciencia y compromiso para la elaboración de este trabajo.

A todos los profesores por su retroalimentación en cada fase presentada.

A los profesores de la maestría en estadística por sus enseñanzas y su disposición para ayudar a cada pregunta o duda que tuve.

Resumen

Los modelos geoestadísticos gaussianos son útiles cuando los datos siguen una distribución normal. Sin embargo, cuando la distribución de los datos es simétrica pero hay presencia de observaciones atípicas entonces se debe asumir una distribución simétrica con colas más pesadas. Por otro lado, en estos modelos geoestadísticos cuando se tienen muchos datos, el principal problema para la inferencia radica en la matriz de covarianza asociada al modelo. En este contexto, esta tesis se ha centrado en extender modelos geoestadísticos usando la distribución normal independiente a través de una función de covarianza tapering o reducida. La función de reducción en la covarianza permite que la matriz de covarianza sea dispersa, característica muy útil en bases de datos grandes. Para la estimación a través de inferencia clásica se propone usar la aproximación de Laplace, para ello se implementó la inferencia en C++ y R a través del Template Model Buider (TMB). Se realizaron estudios de simulación para demostrar la correcta implementación del modelo y las bondades del modelo propuesto. Finalmente, se aplica el modelo para estudiar la distribución espacial del material particulado en Estados Unidos, variable útil para evaluar el nivel de contaminación del aire.

Palabras-clave: Aproximación de Laplace, distribución normal independiente, estadística espacial, geoestadística, tapering.

Abstract

Gaussian geostatistical models are useful when data follow a normal distribution. However, when the data distribution is symmetric but there are atypical observations, then it should be assumed a symmetric distribution with heavier tails. On the other hand, in these geostatistical models when there is a lot of data, the main problem for inference lies in the covariance matrix associated with the model. In this context, this thesis has focused on extending geostatistical models using the independent normal distribution through a covariance tapering function. The covariance tapering function allows the covariance matrix to be sparse, a very useful feature in large databases. For the estimation through classical inference, it is proposed to use the Laplace approximation, thus the inference was implemented in C++ and R through Template Model Builder (TMB). A simulation study was carried out in order to demonstrate the correct implementation as well as benefits of the proposed model. Finally, the model is used to study the spatial distribution of particulate matter in the United States, a useful variable to evaluate the level of air pollution.

Keywords: geostatistics, independent normal distribution, Laplace approximation, spatial statistics, tapering.

Índice general

Capítulo 1

Introducción

1.1. Consideraciones Preliminares

Los datos espaciales contienen información acerca de la posición geográfica de una variable en estudio o datos recolectados en una o varias regiones geográficas típicamente en un espacio euclideano $\mathbb{R}^d$, $d \geq 2$. Un aspecto relevante de estos datos es que en una vecindad cercana las observaciones están correlacionadas entre sí, estos datos pueden ser económicos, epidemiológicos, médicos, imágenes. Asimismo, con los avances computacionales recientes y los repositorios de datos en servidores, la disponibilidad de datos espaciales está creciendo considerablemente, generando gran interés en la aplicación de modelos de estadística espacial en muchos campos de investigación, por ejemplo, en salud pública (Lawson, 2018), ciencias ambientales (Quiroz y Prates, 2018), ciencias sociales (Claramunt y Stewart, 2015), economía agrícola (Amiri y Gerdtham, 2011) entre otras áreas. Una buena referencia para introducirse en el estado de arte del modelamiento espacial son Kent y Mardia (2022), Gelfand et al. (2010), Cressie y Wikle (2011) y Banerjee et al. (2015).

En general los modelos espaciales se definen a través de procesos espaciales gaussianos, debido básicamente a las propiedades conocidas de la distribución normal. Sin embargo, en ciertos contextos, por ejemplo ante la presencia de observaciones atípicas, dicho supuesto de normalidad es violado. Un enfoque para lidiar con este problema es proponer modelos no gaussianos, donde se asume que el proceso espacial tiene una distribución no gaussiana, por ejemplo la distribución t-student, slash, entre otras. En particular, Da-Silva (2017) propuso un modelo geoestadístico no gaussiano usando la distribución normal independiente (Lange y Sinsheimer, 1993), obteniendo como casos particulares los modelos geoestadísticos normal, t-Student y slash. La inferencia para este modelo se realizó a través de el método de maximización de la esperanza (expectation-maximization, EM).

Por otro lado, existe una limitación en el modelamiento espacial cuando se tienen grandes bases de datos. En particular, los modelos espaciales involucran con frecuencia calcular la inversa y determinante de matrices de grandes dimensiones, cuya complejidad computacional se incrementa con el número de ubicaciones espaciales. Debido a ello, algunos investigadores han propuesto trabajar con un proceso gaussiano donde la matriz de covarianza o su inversa se aproximan por una matriz llena de ceros (dispersa). Para alcanzar este objetivo, algunos autores han propuesto usar una función de reducción (tapering) en la matriz de covarianza (Kaufman et al., 2008), procesos predictivos gaussianos (Banerjee et al., 2015), la aproximación de escala completa (FSA) que combina el proceso predictivo y tapering, la aproximación a un proceso gausiano usando ecuaciones diferenciales parciales (Lindgren y Lindström, 2011), un proceso gaussiano derivado de distribuciones condicionales y los vecinos más cercanos a cada ubicación, llamado proceso gaussiano de vecinos más cercanos (NNGP - Datta et al. (2016)), entre otras propuestas.

En este contexto, en esta tesis se propone extender el modelo geoestadístico no gaussiano usando la distribución normal independiente a través de dos enfoques, en primer lugar se propone usar la función de reducción (tapering) en la matriz de covarianza debido a su flexibilidad para modelar grandes bases de datos y en segundo lugar se propone realizar la estimación de los parámetros de este modelo usando inferencia clásica a través de la aproximación de Laplace. Para ello implementamos los modelos en C++ usando la librería Template model Builder (TMB, Kristensen et al. (2016)). La ventaja de esta implementación se demuestra en bases de datos más grandes.

1.2. Objetivos

El objetivo general de esta tesis es extender e implementar el modelo geoestadístico con distribución normal independiente usando covarianza tapering y realizar la inferencia clásica del modelo a través de la aproximación de Laplace.

Los objetivos específicos son:

- Proponer modelos geoestadísticos para variables respuesta con distribución normal independiente usando covarianza reducida. En particular con esta familia de modelos incluye los modelos geoestadísticos con distribuciones normal, t-student y slash.
- Implementar los modelos en C++ y R, para realizar la inferencia usando aproximación de Laplace a través de la librería Template model Builder (TMB).
- Simular datos y ajustar los modelos propuestos.

- Aplicar los modelos en datos reales de contaminación ambiental.

1.3. Organización del Trabajo

En el Capítulo 2, se presentan los conceptos teóricos de la distribución normal independiente, geoestadística e inferencia a través de la aproximación de Laplace. En el Capítulo 3, se presenta el modelo normal independiente para datos no gaussianos usando la función tapering en la matriz de covarianza y sus propiedades. En el Capítulo 4 presentamos un estudio de simulación. En el Capítulo 5 se presenta la aplicación del modelo. Finalmente, en el Capítulo 6 se discuten las conclusiones obtenidas en esta tesis.

Capítulo 2

Marco teórico

En esta sección se aborda la literatura estadística sobre la distribución normal independiente, se realiza una introducción a la geoestadística, en particular a las definiciones y propiedades de un proceso espacial gaussiano y la estimación usando inferencia clásica a través de aproximación de Laplace.

2.1. Distribución normal-independiente (NI)

La distribución normal-independiente es una familia de distribuciones simétricas con colas más pesadas que la distribución normal, por ejemplo la distribución t-student, power exponencial o slash (Lachos y Labra, 2014). Para derivar la función de densidad de probabilidad (fdp) de un vector aleatorio $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ con distribución normal independiente consideremos la siguiente representación estocástica:

$$\mathbf{Y} = \boldsymbol{\mu} + U^{-1/2}\mathbf{Z}, \tag{2.1}$$

donde $\boldsymbol{\mu}$ es un vector de medias de dimensión n, $\mathbf{Z}$ es un vector aleatorio de dimensión n con distribución normal, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, y U es una variable aleatoria positiva cuya función de distribución acumulada (fda) es $H(u)$, que depende del parámetro ν. Entonces $\mathbf{Y} \sim \text{NI}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ tiene una distribución normal independiente (NI).

De forma equivalente, si la distribución condicional de $\mathbf{Y}$ dado $U = u$ es normal multivariada con vector de medias $\boldsymbol{\mu}$ y matriz de covarianza $U^{-1}\boldsymbol{\Sigma}$ y U es una variable aleatoria positiva cuya función de distribución acumulada es $H(u)$ entonces $\mathbf{Y}$ tiene una distribución normal independiente (Lange y Sinsheimer, 1993). La función de densidad de probabilidad (fdp) conjunta de un vector aleatorio (v.a.) $\mathbf{Y}$ con distribución normal independiente,

$\mathbf{Y} \sim \text{NI}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, está dada por:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_0^\infty \frac{u^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{\frac{-u\delta}{2}} dH(u), \tag{2.2}$$

donde $\delta = (\boldsymbol{y} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})$.

El valor esperado y la covarianza del vector aleatorio $\mathbf{Y}$ con distribución normal independiente son, respectivamente:

$$\begin{aligned} E(\mathbf{Y}) &= \boldsymbol{\mu}, \\ Cov(\mathbf{Y}) &= E[U]^{-1}\mathbf{Z}. \end{aligned}$$

Las distribuciones que pertenecen a la familia de la distribución normal-independiente son:

(i) **Distribución Normal:** En este caso la distribución de U es degenerada en uno, luego la función de de densidad de $\mathbf{Y}$ es la siguiente:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}\delta}.$$

(ii) **Distribución t-student:** Si U tiene una distribución $Gamma(\nu/2, \nu/2)$, con fdp

$$h(\mu) = \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)} e^{-\frac{u\nu}{2}} u^{\frac{\nu}{2}-1},$$

donde $\Gamma(.)$ es la función gamma, entonces $\mathbf{Y}$ tiene una distribución $t-$student con ν grados de libertad, es decir la función de densidad de $\mathbf{Y}$ viene dada por:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\Gamma(\frac{\nu+n}{2})}{\Gamma(\frac{\nu}{2})(\nu\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \left[1 + \frac{\delta}{\nu}\right]^{-\frac{(\nu+n)}{2}}.$$

Cuando $\nu \to \infty$ la distribución $t-$student converge a una distribución normal como podemos ver en el lado izquierdo de la Figura 2.1.

(iii) **Distribución slash:** Si $U \sim beta(\nu, 1)$, entonces $\mathbf{Y}$ tiene una distribución Slash con fdp dada por:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{\nu}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} G\left(\frac{n}{2} + \nu, \frac{\delta}{2}\right),$$

donde $G\left(\frac{n}{2} + \nu, \frac{\delta}{2}\right) = \int_0^1 u^{(n/2)+\nu-1} e^{-u\delta/2} du$ representa una función gamma incompleta. Conforme $\nu \to \infty$, la distribución Slash converge a una normal (Figura 2.1).

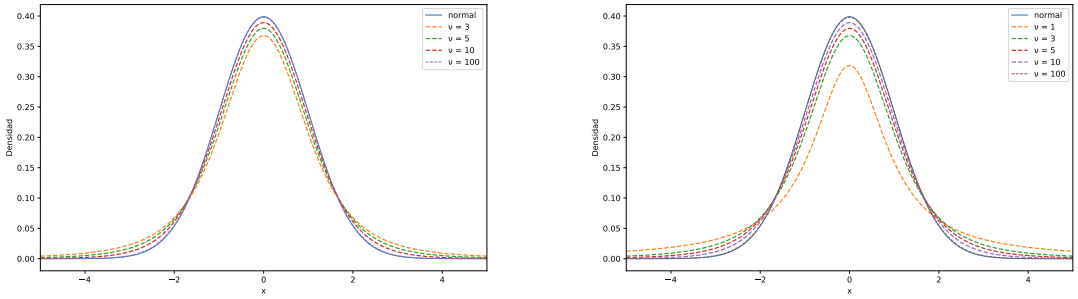


Figura 2.1: Funciones de densidad para distribuciones normales independientes: Distribución t-Student (izquierda) y distribución slash (derecha) para diferentes valores de ν.

2.2. Proceso espacial

Formalmente, podemos definir un campo aleatorio, $\{w(s) : s \in D\}$ sobre un dominio espacial continuo D en $\mathbb{R}^d$, donde $d \geq 2$ es la dimensión. Un proceso espacial es un proceso estocástico, que se caracteriza porque cualquier realización del proceso espacial $\mathbf{w} = (w(s_1), \ldots, w(s_n))^\top$, para cualquier conjunto finito de locales $s_i \in D$, tiene la siguiente función de distribución acumulada (fda) conjunta:

$$F_{\mathbf{w}}(w_1, \ldots, w_n) = P[w(s_1) \leq w_1, \ldots, w(s_n) \leq w_n],$$

donde $w(s_i)$ y $w(s_j)$ no son independientes. Por ejemplo cuando F es la fda conjunta de una normal multivariada, denominamos a este proceso, proceso espacial gaussiano.

Un proceso espacial es estacionario de segundo orden, o es débilmente estacionario, si:

- i) La media del proceso espacial es constante:

$$E(w(s)) = \mu, \quad \forall s \in D,$$

- ii) La función de covarianza $C(.)$ del proceso espacial depende únicamente del vector de separación $\mathbf{h}$ entre dos locales s_i y $s_i + \mathbf{h}$, es decir:

$$Cov(w(s_i), w(s_i + \mathbf{h})) = C(\mathbf{h}), \quad \forall s_i, s_{i+\mathbf{h}} \in D.$$

2.2.1. Variograma y función de covarianza

Si $\{w(s)\}$ es un proceso espacial estacionario intrínseco, la media es constante y el variograma representa la medida de variabilidad de los datos espaciales y se define como:

$$2\gamma(h) = V(w(s+h) - w(s)),$$

donde $h = \|\boldsymbol{h}\|$ es la distancia entre dos locales y $\gamma(h)$ es la función de semivariograma. Note que un proceso espacial estacionario de segunda orden implica que un proceso espacial presente estacionariedad intrínseca.

La relación que hay entre el variograma y la función de covarianza está dada por:

$$\begin{aligned} 2\gamma(h) &= V(w(s+h) - w(s)) \\ &= V(w(s+h)) + V(w(s)) - 2\mathrm{C}ov(w(s+h), w(s)) \\ &= \mathrm{C}(0) + \mathrm{C}(0) - 2\mathrm{C}(h) \\ &= 2(\mathrm{C}(0) - \mathrm{C}(h)), \end{aligned}$$

de donde se tiene que,

$$\gamma(h) = \mathrm{C}(0) - \mathrm{C}(h).$$

El semivariograma tiene tres componentes:

- Efecto pepita o *nugget effect* (τ^2): representa a la variabilidad no espacial, por ejemplo debido a errores de medición, es definida por

$$\lim_{\mathbf{h}\to 0^+} \gamma(h) = \tau^2.$$

- Meseta parcial o *partial sill* (σ^2): es la varianza marginal debido a la variabilidad puramente espacial, y se define por

$$\lim_{h\to\infty} \gamma(h) = \tau^2 + \sigma^2.$$

- Alcance o range (r): es la distancia a partir del cual se considera que no hay autocorrelación espacial significativa entre las observaciones.

Estas componentes se muestran en la Figura 2.2. El efecto pepita es la altura donde comienza la línea negra en el eje de las ordenadas aproximadamente 0.2, el rango es la distancia

donde se estabiliza la varianza, aproximadamente en 7 y la meseta parcial es aproximadamente 0.55 (es decir cuando se estabiliza la varianza) menos el nugget.

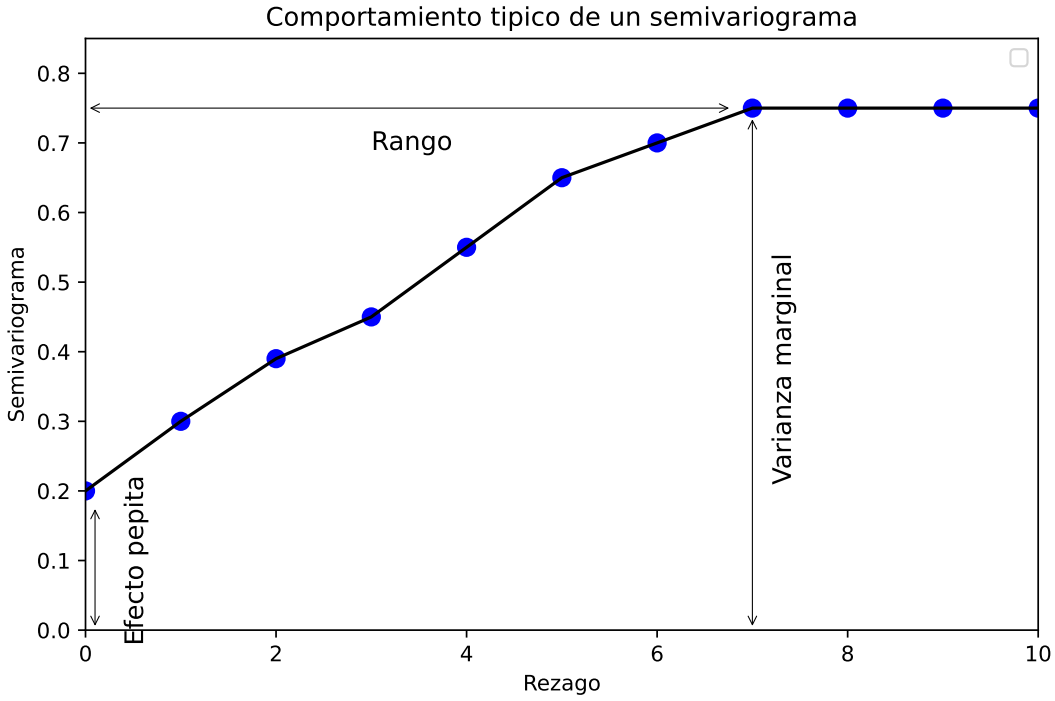


Figura 2.2: Componentes del variograma. Semivariograma teórico (línea azul), variograma empírico (puntos azules). Fuente:Kent y Mardia (2022).

Por simplicidad, sea la distancia $h = \boldsymbol{h}$. A continuación se definen algunos semivariogramas teóricos y sus asociadas funciones de covarianza teóricas:

- ***Esférico:***

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left\{ \frac{3h}{2r} - \frac{1}{2}(\frac{h}{r})^3 \right\}; & \text{si } 0 < h \leq r, \\ \tau^2 + \sigma^2; & \text{si } h \geq r \\ 0; & \text{en otro caso.} \end{cases}$$

$$\mathrm{C}(h) = \begin{cases} \tau^2 + \sigma^2; & \text{si } h = 0 \\ \sigma^2 \left(1 - \frac{3h}{2r} + \frac{1}{2}(\frac{h}{r})^3\right); & \text{si } 0 < h < r, \\ 0; & \text{en otro caso.} \end{cases}$$

- ***Exponencial:***

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left(1 - e^{-\frac{h}{r}}\right); & \text{si } h > 0 \\ 0; & \text{en otro caso.} \end{cases}$$

$$\mathrm{C}(h) = \begin{cases} \tau^2 + \sigma^2; & \text{si } h = 0 \\ \sigma^2 e^{-\frac{h}{r}}; & \text{si } h > 0 \end{cases} \tag{2.3}$$

- ***Gaussiano:***

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left(1 - e^{-(\frac{h}{r})^2}\right); & \text{si } h > 0 \\ 0; & \text{en otro caso} \end{cases}$$

$$\mathrm{C}(h) = \begin{cases} \tau^2 + \sigma^2; & \text{si } h = 0 \\ \sigma^2 e^{-\left(\frac{h}{r}\right)^2}; & \text{si } h > 0 \end{cases}$$

- ***Matérn:***

$$\gamma(h) = \begin{cases} \tau^2 + \sigma^2 \left[1 - (2^{\nu-1}\Gamma(\nu))^{-1} \left(\frac{h}{r}\right)^{\nu} K_\nu \left(\frac{h}{r}\right)\right]; & \text{si } h > 0 \\ 0; & \text{en otro caso.} \end{cases}$$

$$\mathrm{C}(h) = \begin{cases} \tau^2 + \sigma^2; & \text{si } h = 0 \\ \sigma^2 (2^{\nu-1}\Gamma(\nu))^{-1} \left(\frac{h}{r}\right)^{\nu} K_\nu \left(\frac{h}{r}\right); & \text{si } h > 0 \end{cases}$$

donde $\nu > 0$ es un parámetro que controla la suavización del campo aleatorio, $\Gamma(.)$ es la función gamma, mientras que K_ν es la función de Bessel modificada de orden ν.

La función de covarianza Matérn es un caso general, pues cuando $\nu = 1/2$ es una función de covarianza exponencial (ecuación (2.3)) y cuando $\nu \to \infty$ es una función de covarianza gaussiana.

2.2.2. Proceso espacial gaussiano

Un proceso espacial gaussiano es un proceso estocástico, que se caracteriza porque la realización del proceso para cualquier conjunto finito de locales $s_i \in D$, la cual tiene una función de distribución de probabilidad conjunta es gaussiana, es decir, $\mathbf{w}(s) = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^\top \sim N(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$, donde $\mathbf{C}(\boldsymbol{\theta})$ es una matriz de covarianza definida a través de una función de covarianza válida como las definidas en la sección anterior, la cual depende de un conjunto de parámetros desconocidos $\boldsymbol{\theta}$. Por ejemplo si se usa la función de covarianza exponencial, $\boldsymbol{\theta} = (\tau^2, \sigma^2, r)$.

2.2.3. Matriz reducida o Tapering

Para una "distancia taper" γ mayor que el alcance o range, es decir $\gamma > r$, se define $K_{taper}(h,\gamma)$ como una función de covarianza llamada función de covarianza "tapering" $K_{taper}(h,\gamma)$, donde para una distancia h se tiene que $K_{taper}(h,\gamma) = 0$ cuando h es mayor que la "distancia taper", es decir cuando $h \geq \gamma$. De acuerdo a Kaufman et al. (2008), podemos obtener una función de covarianza reducida válida, multiplicando una función de covarianza válida $C(h)$ por esta función de covarianza tapering, tal que:

$$C_{taper}(h,\gamma) \quad = \quad C(h)K_{taper}(h,\gamma), \quad h > 0. \tag{2.4}$$

Cabe resaltar que la matriz de covarianza definida por esta función de covarianza reducida (tapering) $C_{taper}(h,\gamma)$ es una definida positiva y es dispersa, es decir está llena de ceros. Furrer et al. (2006) define algunas funciones tapering, reparametrizadas para tener soporte en $[0,\gamma)$:

- **Esférica:** válida para ν <0.5,

$$C_{taper}(h,\gamma) = \left(1 - \frac{h}{\gamma}\right)_+^2 \left(1 + \frac{h}{2\gamma}\right).$$

- **Wendland$_1$**: válida para ν <1.5,

$$C_{taper}(h,\gamma) = \left(1 - \frac{h}{\gamma}\right)_+^4 \left(1 + 4\frac{h}{\gamma} + 3\frac{h^2}{\gamma^2} + \frac{3h^3}{4\gamma^3}\right).$$

- **Wendland$_2$** : válida para ν <2.5,

$$C_{taper}(h,\gamma) = \left(1 - \frac{h}{\gamma}\right)_+^6 \left(1 + 6\frac{h}{\gamma} + \frac{41h^2}{3\gamma^2} + 12\frac{h^3}{\gamma^3} + 5\frac{h^4}{\gamma^4} + \frac{5h^5}{6\gamma^5}\right).$$

donde $x_+ = max\{0, x\}$.

En la práctica la distancia taper γ es la distancia máxima hasta la cual deliberadamente asumimos que puede haber autocorrelación espacial entre los locales, por ello para definir su valor el investigador debe usar su conocimiento previo sobre la variable en estudio o usar herramientas de análisis exploratorio como el variograma.

2.2.4. Modelo geoestadístico

Un modelo geoestadístico clásico gaussiano tiene la siguiente forma:

$$Y(s) = \mathbf{x}^\top(s)\boldsymbol{\beta} + w(s) + \varepsilon(s),$$

donde $s \in D \subseteq \mathbb{R}^n$, $Y(s)$ es la variable dependiente de dimensión $n \times 1$, $\mathbf{x}(s)$ es el vector de covariables, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión, $\varepsilon(s)$ es el error aleatorio por ejemplo errores de medición, se asume que $\varepsilon(s) \overset{iid}{\sim} N(0, \tau^2)$, y $w(\mathbf{s})$ es un efecto aleatorio espacial definido por un proceso espacial gaussiano. El vector $\mathbf{w}(s) = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))^\top$ modela la autocorrelación espacial y tiene la siguiente distribución:

$$\mathbf{w}(s) \sim N(\mathbf{0}, C(\boldsymbol{\theta})).$$

Luego la distribución de $\mathbf{Y} = (Y(s_1), Y(s_2), \ldots, Y(s_n))^\top$, es $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$, con matriz de covarianza definida por $\Sigma_{\mathbf{Y}} = C(\boldsymbol{\theta}) + \tau^2\mathbf{I}_n$, donde $C(\boldsymbol{\theta}) = [C(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\theta})]_{i,j=1}^n$.

2.3. Inferencia clásica usando la aproximación de Laplace

Se define $f(\mu, \theta)$ como el negativo del logaritmo de la función de verosimilitud que depende del vector de variables aleatorias latentes $\mu \in \mathbb{R}^n$ y los parámetros $\theta \in \mathbb{R}^m$. El estimador de máxima verosimilitud, maximiza la función de verosimilitud marginal:

$$L(\theta) = \int_{\mathbb{R}^n} e^{-f(\mu,\theta)} d\mu.$$

Definamos $\hat{\mu}(\theta)$ como:

$$\hat{\mu}(\theta) = \arg\min_{\mu} f(\mu, \theta).$$

La aproximación de Laplace para la verosimilitud marginal es

$$L^*(\theta) = (2\pi)^{\frac{n}{2}} |H(\theta)|^{-\frac{1}{2}} e^{-f(\hat{\mu},\theta)},$$

donde $H(\theta)$ la hessiana de $f(\mu, \theta)$ evaluada en $\hat{\mu}(\theta)$, es decir,

$$H(\theta) = f''_{\mu\mu}(\hat{\mu}(\theta), \theta),$$

donde $f''_{\mu\mu}$ es la segunda derivada con respecto a μ. El cálculo de esta hessiana es complicado, pero puede obtenerse a través de diferenciación numérica.

El valor estimado de θ maximiza el logaritmo de la aproximación de Laplace $L^*(\theta)$, o en su defecto minimiza el negativo del logaritmo de la aproximación de Laplace:

$$-\log L^*(\theta) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}|H(\theta)| + f(\hat{\mu}, \theta).$$

El estimador $\hat{\theta}$ se pueden obtener usando métodos numéricos conocidos, por ejemplo Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno (BFGS).

La librería Template model Builder (TMB) en R resuelve estos problemas usando técnicas y softwares de desarrollo reciente, véase para más detalles Kristensen et al. (2016).

Capítulo 3

Modelo geoestadístico NI usando tapering

En la literatura actual, es común asumir procesos gaussianos para modelar efectos aleatorios espaciales, debido a su facilidad de enfoque, sin embargo este supuesto puede ser cuestionado ante la presencia de datos atípicos. Este problema se puede abordar asumiendo que el efecto aleatorio es no gaussiano implicando que la variable de respuesta también sea no gaussiana y siga la misma distribución no gaussiana.

3.1. Modelo geoestadístico NI

Este modelo fue propuesto por Da-Silva (2017). Para el conjunto de datos de dimensión n se define el vector aleatorio $\mathbf{Y} = (Y(s_1), \ldots, Y(s_n))^\top$. El modelo no-gaussiano a tratar en esta sección es definido por:

$$Y(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \epsilon(\mathbf{s}),$$

donde $\mathbf{X}\boldsymbol{\beta}$ es el componente determinístico, y se asume que el proceso $\boldsymbol{\epsilon} = (\epsilon(s_1), \ldots, \epsilon(s_n))^\top$ es un proceso cuya distribución proviene de una distribución normal independiente, es decir,

$$\boldsymbol{\epsilon} \sim \mathcal{NI}(\mathbf{0}, \boldsymbol{\Sigma}, \nu),$$

donde la matriz de covarianza de $\boldsymbol{\epsilon}$ es dada por:

$$\boldsymbol{\Sigma} = \tau^2\mathbf{I} + \sigma^2\mathbf{R}(\phi), \tag{3.1}$$

donde τ^2 es el efecto pepita , σ^2 es la varianza marginal, R es la función de una correlación y ϕ es el alcance o range. A este modelo se le llama modelo geoestadístico normal independiente (GNI).

Por tanto la distribución de $\mathbf{Y}$ es una normal independiente,

$$\mathbf{Y} \sim \mathcal{NI}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu).$$

Luego, usando la representación estocástica de la ecuación (2.1), se tiene que

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + U^{-1/2}\mathbf{Z},$$

donde $\mathbf{Z}$ es un vector aleatorio de dimensión n con distribución $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ y U es una variable aleatoria positiva que depende del parámetro ν.

De esta forma, los modelos geoestadísticos obtenidos son:

(i) Normal: Si U es una variable aleatoria degenerada en 1, obtenemos el modelo geoestadístico normal.

(ii) T-student: Si $U \sim gamma(\nu/2, \nu/2)$ obtenemos el modelo geoestadístico t-student(ν).

(iii) Slash: Si $U \sim beta(\nu, 1)$ obtenemos el modelo geoestadístico slash(ν).

De la ecuación (2.2) la fdp de $\mathbf{Y} \sim \mathcal{NI}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}, \nu)$ es dada por:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int_0^\infty \frac{u^{\frac{n}{2}}}{(2\pi)^{\frac{n}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{\frac{-u\delta}{2}} dH(u), \tag{3.2}$$

donde $H(u)$ es la función de distribución acumulada de U y $\delta = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top {\boldsymbol{\Sigma}_\mathbf{1}}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$.

En particular, asumiremos que la función de covarianza es una función exponencial (Ecuación (2.3)). Luego la matriz de covarianza para una función de covarianza exponencial queda definida por:

$$\Sigma(h) = \begin{cases} \tau^2 + \sigma^2; & \text{si } h = 0 \\ \sigma^2 e^{-\frac{h}{\phi}}; & \text{si } h > 0 \end{cases}$$

donde $\phi > 0$, h es la distancia Euclideana entre dos locales s y s', es decir, $h = \|\mathbf{s} - \mathbf{s}'\|$. Luego la ecuación (3.1) queda definida por $\boldsymbol{\Sigma} = \tau^2\mathbf{I} + \sigma^2\mathbf{R}(\phi) = \tau^2\mathbf{I} + \sigma^2 \exp(-\frac{\mathbf{D}}{\phi})$, donde $\mathbf{D}$ es la matriz de distancias.

Tanto para realizar la inferencia como para obtener predicciones se necesitará obtener la inversa de $\boldsymbol{\Sigma}$ lo cual es computacionalmente costoso. En la literatura hay dos propuestas que

intentan reducir el tiempo de la estimación, una es aproximar la matriz de covarianza a una dispersa (llena de ceros), la otra propuesta es obtener la inversa de la matriz de covarianza (precisión) de tal forma que esta por definición sea dispersa. En esta tesis se abordará el primer enfoque.

3.2. Modelo geoestadístico NI-tapering

El modelo geoestadístico NI presentado en la sección anterior se puede extender aproximando la matriz de covarianza por una matriz dispersa para mejorar la eficiencia en la estimación. Kaufman et al. (2008) propuso obtener una matriz de covarianza dispersa usando tapering, es decir multiplicando los elementos de la matriz de covarianza original $\mathbf{\Sigma}$ por una matriz de covarianza tapering $\mathbf{K}_\gamma$, luego los elementos de la matriz de covarianza reducida (tapering) se obtienen de la siguiente función de covarianza:

$$\Sigma_{tap}(h,\gamma) = K_\gamma(h) \times \Sigma(h),$$

donde γ es la distancia taper.

Siguiendo Furrer et al. (2006), como se usa una función de covarianza exponencial donde ν=0.5, se considera $K_\gamma(h)$ del tipo $\mathbf{Wendland_1}$, válida para ν <1.5, definida por:

$$K_\gamma(h) = \left(1 - \frac{h}{\gamma}\right)_+^4 \left(1 + 4\frac{h}{\gamma} + 3\frac{h^2}{\gamma^2} + \frac{3h^3}{4\gamma^3}\right) \tag{3.3}$$

donde $x_+ = max\{0, x\}$. Luego, la matriz de covarianza "tapering" del modelo es dada por: $\mathbf{\Sigma}_{tap} = \mathbf{K}_\gamma \times \mathbf{\Sigma}$. Reemplazando $\mathbf{\Sigma}$ por $\mathbf{\Sigma}_{tap}$ en los modelos geoestadísticos de la sección anterior, se obtienen los modelos geoestadísticos normal, t-student y slash "con tapering".

3.3. Inferencia

El vector de parámetros a estimar $\boldsymbol{\theta} = (\sigma^2, \tau^2, \phi, \boldsymbol{\beta})$ incluye los parámetros de varianza marginal, efecto pepita, decaimiento y vector de coeficientes de regresión. Así la función de log-verosimilitud puede ser definida como el logaritmo de la ecuación (3.2), tal que:

$$l(\boldsymbol{\theta}, \mathbf{y}) = \log\left(\int_0^\infty (2\pi)^{-n/2} u^{n/2} \left|\mathbf{\Sigma_1}\right|^{-1/2} e^{-\frac{u\delta}{2}} dH(u)\right), \tag{3.4}$$

donde para el modelo geoestadístico NI la matriz de covarianza es $\mathbf{\Sigma_1} = \tau^2\mathbf{I} + \sigma^2 \exp\left(-\frac{\mathbf{D}}{\phi}\right)$ y para el modelo NI usando covarianza tapering $\mathbf{\Sigma_1} = \mathbf{\Sigma}_{tap} = \mathbf{K}_\gamma \times [\tau^2\mathbf{I} + \sigma^2 \exp\left(-\frac{\mathbf{D}}{\phi}\right)]$,

donde $\mathbf{K}_\gamma$ queda definida por la ecuación (3.3).

En particular, dependiendo de la distribución que asume U, se obtienen las siguientes funciones de log-verosimilitud (ecuación 3.4) de los modelos geoestadísticos sin tapering y con tapering:

(i) Normal:

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\boldsymbol{\Sigma_1}| - \frac{\delta}{2}.$$

(ii) T-Student:

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \log C_t - \frac{1}{2}\log|\boldsymbol{\Sigma_1}| - \left(\frac{\nu+n}{2}\right)\log\left(1+\frac{\delta}{\nu}\right),$$

$$\text{donde } C_t = \frac{\Gamma((\nu+n)/2)}{\Gamma(\nu/2)(\nu\pi)^{n/2}}.$$

(iii) Slash:

$$l(\boldsymbol{\theta}, \boldsymbol{y}) = \log C_s + \log\nu - \frac{1}{2}\log|\boldsymbol{\Sigma_1}| - \left(\frac{n}{2}+\nu\right)\log\frac{\delta}{2} + \log F_S(1),$$

$$\text{donde } C_s = \frac{\Gamma(n/2+\nu)}{(\nu\pi)^{n/2}}.$$

3.3.1. Inferencia clásica usando aproximación de Laplace

La estrategia de estimación de los parámetros podría ser vía la maximización de la función de log-verosimilitud, usando algoritmos numéricos en caso no haya una solución analítica. Una alternativa es el uso del algoritmo EM. No obstante muchos de estos algoritmos presentan algunos problemas para su implementación.

Otra alternativa es maximizar el logaritmo de la función verosimilitud usando la aproximación de Laplace. En esta tesis, este método será usado para la estimación de los parámetros. Por definición podemos escribir la función de verosimilitud en función de la variable latente U, tomando la forma:

$$L(\boldsymbol{\theta}; y) = f_{\boldsymbol{Y}}(\boldsymbol{y}) = \int f_{\boldsymbol{Y},U}(\boldsymbol{y}, u)du = \int f_{\boldsymbol{Y}|U}(\boldsymbol{Y}|U=u)f_U(u)du, \tag{3.5}$$

donde $f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u) \sim N(\boldsymbol{X\beta}, u^{-1}\boldsymbol{\Sigma_1})$ y $f_U(u)$ es la fdp de la distribución de mistura, es decir, de las distribuciones degenerada en 1, gamma$(\nu/2, \nu/2)$ o beta$(\nu, 1)$.

La función de verosimilitud de la ecuación (3.5) puede ser reescrita como:

$$L(\boldsymbol{\theta}; y) = \int \exp(-f(u, \boldsymbol{\theta}))du, \tag{3.6}$$

considerando

$$\begin{aligned} f(u, \boldsymbol{\theta}) &= -\log[f_{\boldsymbol{Y}|U}(\boldsymbol{y}|u) f_U(u)] \\ &= -\log[f_U(u)] - \log[f_{\boldsymbol{Y}|U}(y|u)]. \end{aligned}$$

El estimador de máxima verosimilitud, maximiza la función de verosimilitud marginal definida por la ecuación (3.6). Definamos $\hat{u}(\boldsymbol{\theta})$ como:

$$\hat{u}(\boldsymbol{\theta}) = \arg \min_{u} f(u, \boldsymbol{\theta}).$$

La aproximación de Laplace para la verosimilitud marginal es

$$L^*(\boldsymbol{\theta}) = (2\pi)^{\frac{n}{2}} |H(\boldsymbol{\theta})|^{-\frac{1}{2}} e^{-f(\hat{u}, \boldsymbol{\theta})},$$

donde $H(\boldsymbol{\theta})$ la hessiana de $f(u, \boldsymbol{\theta})$ evaluada en $\hat{u}(\boldsymbol{\theta})$, es decir,

$$H(\boldsymbol{\theta}) = f''_{uu}(\hat{u}(\boldsymbol{\theta}), \boldsymbol{\theta}),$$

donde f''_{uu} es la segunda derivada con respecto a u. El cálculo de esta hessiana es complicado, pero puede obtenerse a través de diferenciación numérica.

El valor estimado de $\boldsymbol{\theta}$ maximiza el logaritmo de la aproximación de Laplace $L^*(\boldsymbol{\theta})$, o en su defecto minimiza el negativo del logaritmo de la aproximación de Laplace:

$$-\log L^*(\boldsymbol{\theta}) = -\frac{n}{2} \log(2\pi) + \frac{1}{2} |H(\boldsymbol{\theta})| + f(\hat{u}, \boldsymbol{\theta}).$$

El estimador $\hat{\boldsymbol{\theta}}$ se pueden obtener usando métodos numéricos conocidos, por ejemplo Newton-Raphson, Broyden-Fletcher-Goldfarb-Shanno (BFGS). La inferencia para todos los modelos geoestadísticos gaussiano, t-student y slash, tanto sin tapering y con tapering, se implementó en C++ usando esta aproximación de Laplace a través de una interfaz que brinda el paquete *Template Model Buider* (*TMB*) en R. Para mayor información sobre TMB ver Kristensen et al. (2016).

3.4. Predicción

Sea $Y = (Y(s_1), \ldots, Y(s_n))^\top$ el vector aleatorio observado y sea la variable aleatoria $Y_0 = Y(\mathbf{s}_0)$ en el local $\mathbf{s}_0$, sobre la cual se desea realizar la predicción definida por $E[y_0|\mathbf{y}]$.

3.4.1. Modelo geoestadístico gaussiano

Para el modelo geoestadístico normal, se sabe que la distribución de $(Y_0, \mathbf{Y})$ es normal multivariada,

$$\begin{pmatrix} \mathbf{Y}_0 \\ \boldsymbol{Y} \end{pmatrix} \sim N_{(1+n)}\left(\begin{pmatrix} \mu_0 \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

Debido a las propiedades conocidas para la distribución normal multivariada se tiene que:

$$\begin{aligned} E[y_0|\mathbf{y}] &= \mu_0 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \\ V[y_0|\mathbf{y}] &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

3.4.2. Modelo geoestadístico t-Student

Si $(y_p, \mathbf{y}_0)$ tiene una distribución t-student multivariada, se tiene que

$$\begin{pmatrix} \mathbf{Y}_0 \\ \boldsymbol{Y} \end{pmatrix} \sim tM_{(1+n)}\left(\begin{pmatrix} \mu_0 \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \nu\right)$$

donde ν son los grados de libertad.

Ding (2016) y Roth (2013) muestran que $f(y_0|\mathbf{y})$ es una distribución t-student y se caracteriza por

$$\begin{aligned} E[y_0|\mathbf{y}] &= \mu_0 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \\ \Sigma_{[y_0|\mathbf{y}]} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ V[y_0|\mathbf{y}] &= \frac{\nu + (\mathbf{y} - \boldsymbol{\mu})'\Sigma_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu})}{\nu + n}\Sigma_{[y_0|\mathbf{y}]} \\ \nu_{[y_0|\mathbf{y}]} &= \nu + n. \end{aligned}$$

3.4.3. Modelo geoestadístico slash

La distribución condicional $f(y_0|\mathbf{y})$ es

$$\begin{aligned} f(y_0|\mathbf{y}) &= \int f(y_0, u|\mathbf{y})du \\ &= \int f(y_0|\mathbf{y}, u) f_U(u)du. \end{aligned} \tag{3.7}$$

Como no se tiene forma cerrada para esta distribución, pero la distribución de $f(y_0|\mathbf{y}, u)$ es normal con la siguiente media y varianza:

$$\begin{aligned} E[y_0|\mathbf{y}, u] &= \mu_0 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \\ V[y_0|\mathbf{y}], u &= u^{-1}\left(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right), \end{aligned}$$

entonces se pueden obtener muestras de esta distribución y usar Monte Carlo en la ecuación (3.7) para estimar la predicción de y_0.

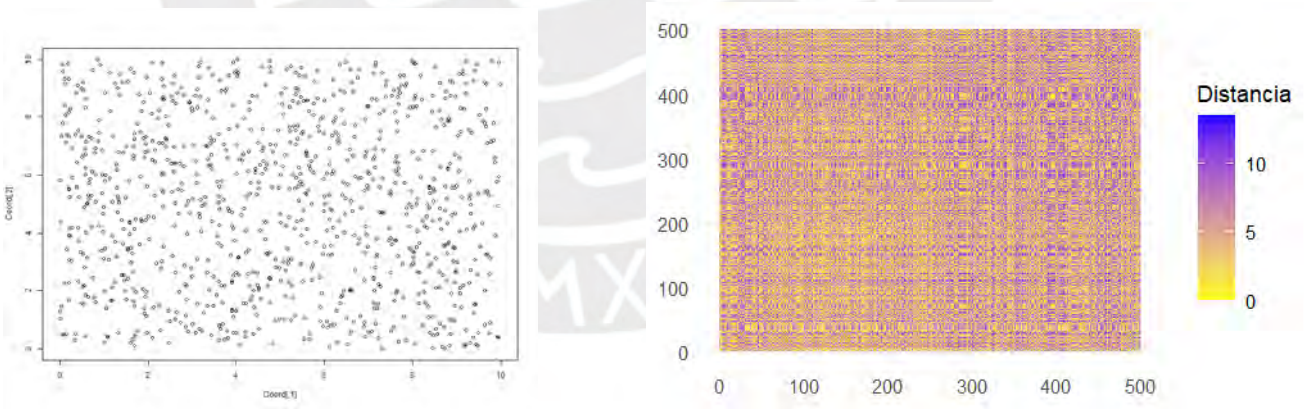
Capítulo 4

Estudio de Simulación

Se simularon datos a partir de los modelos geoestadísticos NI (nomal, beta y gamma) definidos en el capítulo 3. Posteriormente se ajustaron los modelos geoestadístico NI usando el proceso gaussiano (Full-GP) y el modelo geoestadístico NI-tapering (cov-taper). Se estimaron los parámetros mediante inferencia clásica usando la aproximación de Laplace.

4.1. Simulación de datos del modelo geoestadístico NI

Primero generamos $n = 500$ locales aleatoriamente en un espacio de $[0, 10] \times [0, 10]$, luego obtenemos la matriz de distancias entre todos los puntos (Figura 4.1). Para simular los datos



(a) Generación aleatoria de 500 puntos. (b) Matriz de distancias.

Figura 4.1: Generación aleatoria de ubicaciones o sitios

definimos los valores de los parámetros como sigue: β_0 =0.5, $\beta_1 = 2$, $\phi = 2$, τ^2 =0.1, y el efecto espacial con $\sigma^2 = 1$. Se genera datos de la covariable a partir de una normal estándar, es decir, $x_i \sim N(0, 1)$, y se calcula la media de $Y(s_i)$ dada por $\beta_0 + \beta_1 x_i$. También se calcula la matriz de covarianza $\mathbf{\Sigma}$ a partir de la función de covarianza exponencial, tal que $\Sigma(h) = \sigma^2 e^{-\frac{h}{\phi}}$. Se simulan datos de $U \sim f_U(u)$ para las distribuciones: i) $U \sim$ gamma$(\nu/2, \nu/2)$ y ii) $U \sim$

beta$(\nu, 1)$. Luego para el datos u simulado de la distribución de U, se calcula $\mathbf{\Sigma}_Y = u^{-1}\mathbf{\Sigma}$. En base a etas definiciones se simularon datos para diferentes escenarios a partir de los modelos geoestadísticos normal (u=1), t-student (U es una gamma) con $\nu = 3, 5, 10$ y Slash (U es una beta) con $\nu = 3, 5, 10$. Así en cada escenario se simulan n datos del vector aleatorio $Y = (Y_1, \ldots, Y_n)^\top$ a partir de una distribución $N(\boldsymbol{X}\beta, \mathbf{\Sigma}_Y)$, que es equivalente a simular datos del vector aleatorio $Y \sim NI(\mathbf{X}\beta, \mathbf{\Sigma}, \nu)$.

4.2. Estimación

Para estudiar la eficiencia de los modelos geoestadísticos NI originales y usando el tapering se ajustaron los siguientes modelos en cada escenario:

i) Full-GP: El modelo geoestadístico NI usando la matriz de covarianza completa. Es decir se estimaron los parámetros a partir del modelo original.

ii) Cov-taper-7: El modelo geoestadístico NI-taper (cov-taper) usando la aproximación tapering definido en la sección 3.2. Como se simularon los datos en el cuadrado $[0, 10] \times [0, 10]$, el rango $\phi = 2$ es decir el rango es pequeño, por ello se asumió que la distancia tapering es $\gamma = 7$, es decir un tapering mucho mayor que el verdadero rango.

Luego se estimaron todos los parámetros mediante inferencia clásica, específicamente usando la aproximación de Laplace a través de la implementación de los modelos en TMB. Con la finalidad de evaluar la bondad de ajuste de los modelos se corrieron 100 réplicas para los siete escenarios bajos los dos modelos geoestadísticos NI y NI-taper. Los resultados obtenidos se muestran en la siguiente sección.

4.3. Resultados

La Figura 4.2 muestra diagramas de caja de los valores estimados de β_0 para todos los escenarios y modelos ajustados. En general vemos que se ha recuperado bien el valor original del parámetro igual a 0.5, tanto con tapering (cov-taper 7) y sin tapering (Full-GP), el resultado es similar en ambos casos. En el caso de datos simulados del modelo geoestadístico normal se tienen menos outliers, y el modelo geoestadístico t-student con $\nu = 3$ es decir cuando se tienen colas más pesadas, es donde se tienen más outliers, mostrando que para este modelo es más difícil la recuperación del β_0.

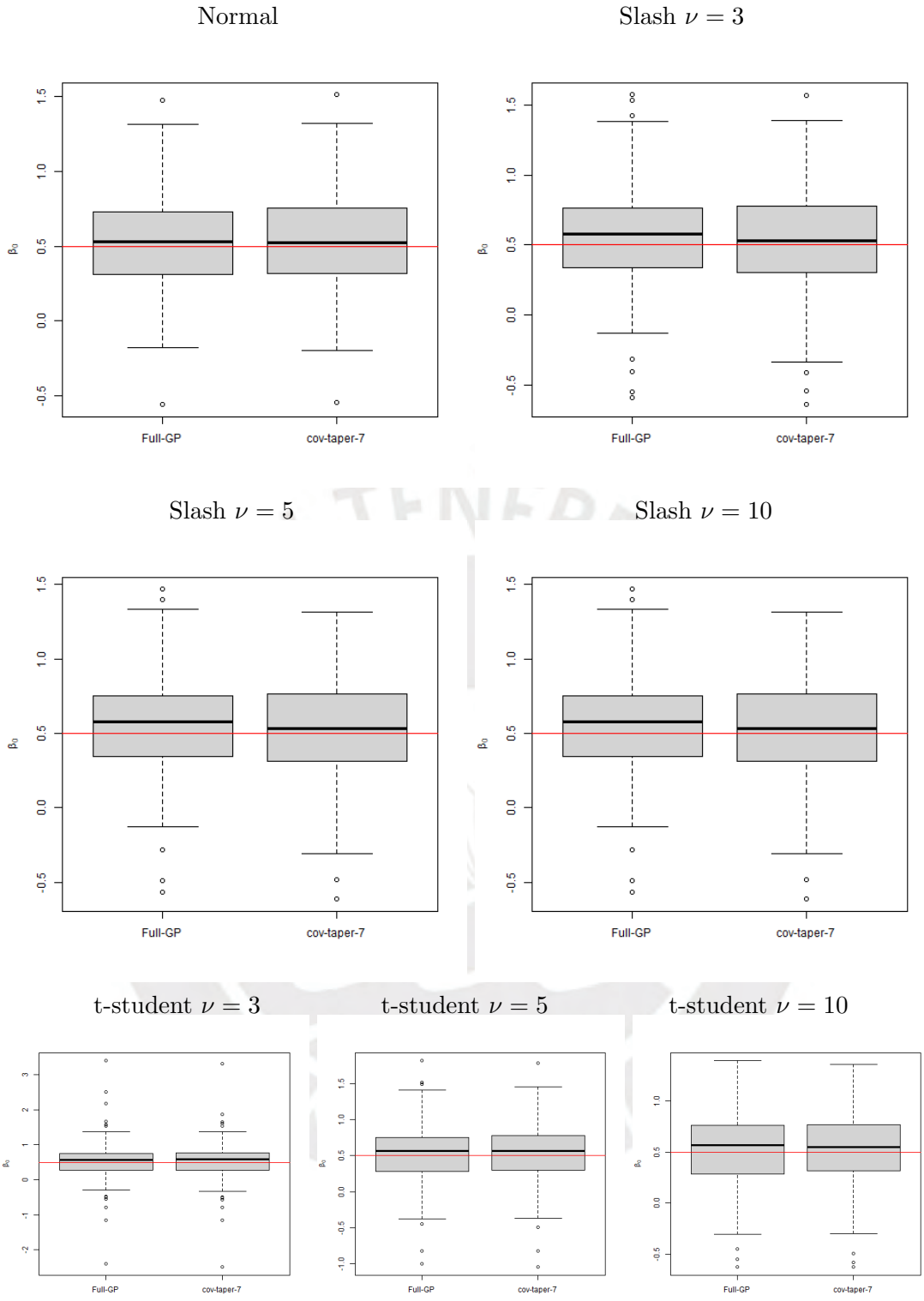


Figura 4.2: Diagramas de caja de 100 estimaciones de β_0 con distribuciones normal, slash y t-student.

La Figura 4.3 muestra diagramas de caja de los valores estimados de β_1 para todos los escenarios y modelos ajustados. En general vemos que se ha recuperado bien el valor original del parámetro igual a 2, tanto con tapering (cov-taper 7) y sin tapering (Full-GP), el resultado

muy similar en ambos casos y para todos los modelos.

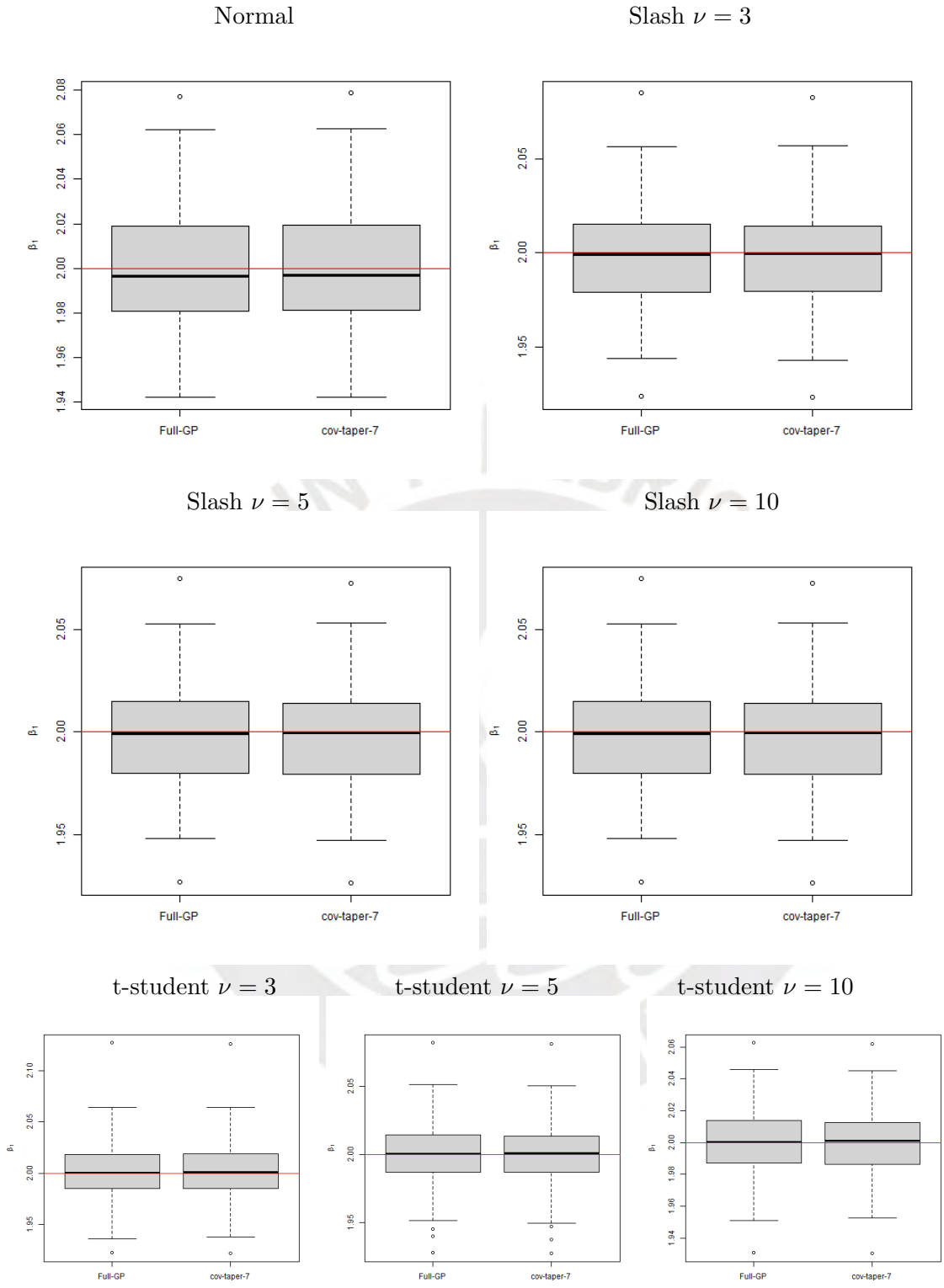


Figura 4.3: Diagramas de caja de 100 estimaciones de β_1 con distribuciones normal, slash y t-student.

La Figura 4.4 muestra diagramas de caja de los valores estimados de ϕ para todos los escenarios y modelos ajustados. En general vemos que se ha recuperado bien el valor ori-

ginal del parámetro igual a 2, tanto con tapering (cov-taper 7) y sin tapering (Full-GP), sin embargo para este parámetro se observa que con tapering hay mayor variabilidad en las estimaciones de ϕ bajo todos los escenarios.

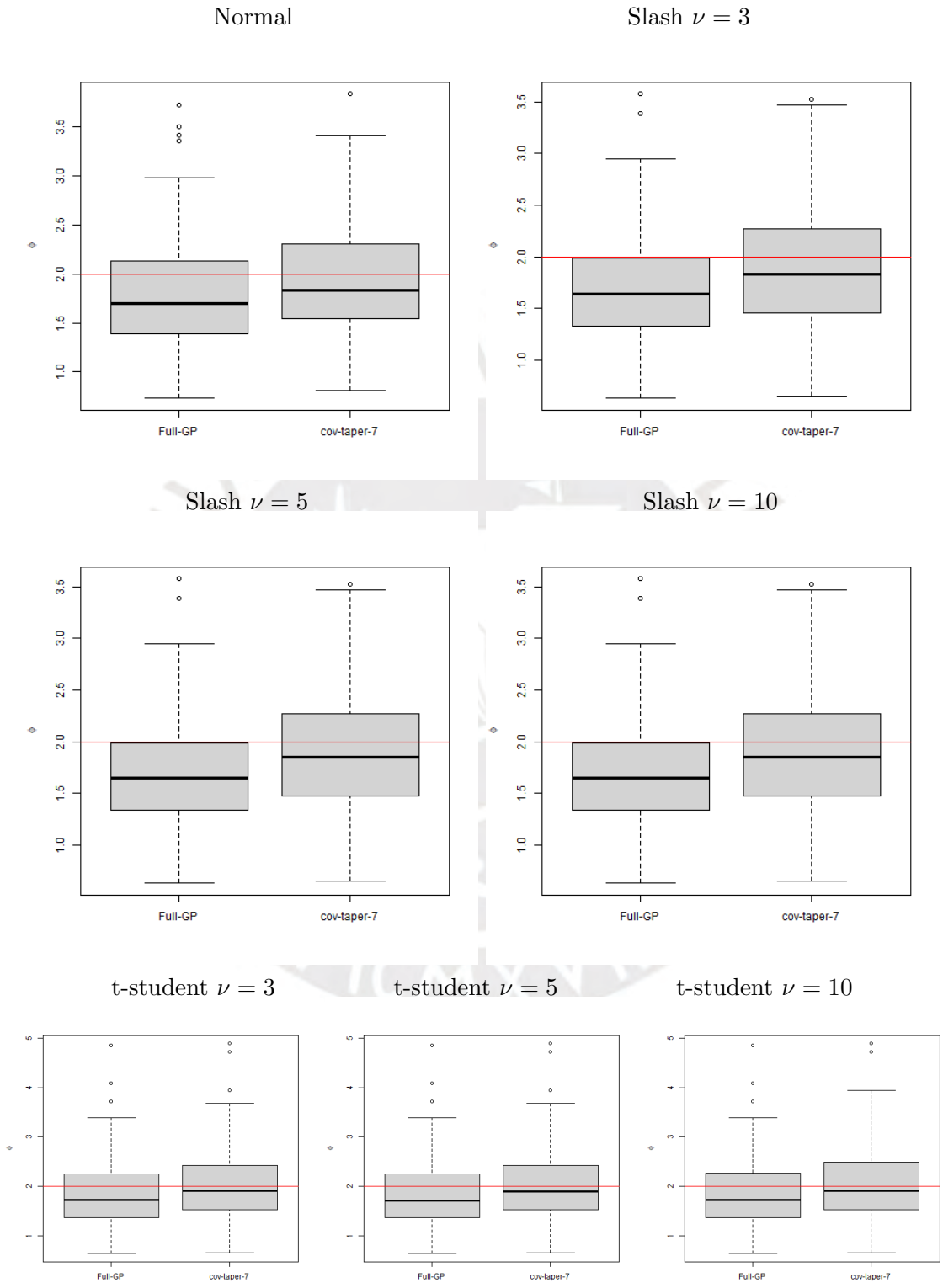


Figura 4.4: Diagramas de caja de 100 estimaciones del rango (ϕ) con distribuciones normal, slash y t-student.

Los diagramas de caja de las estimaciones de σ^2 se muestran en la figura 4.5. Podemos observar que la varianza marginal es ligeramente más subestimada por los modelos con tapering (cov-taper-7).

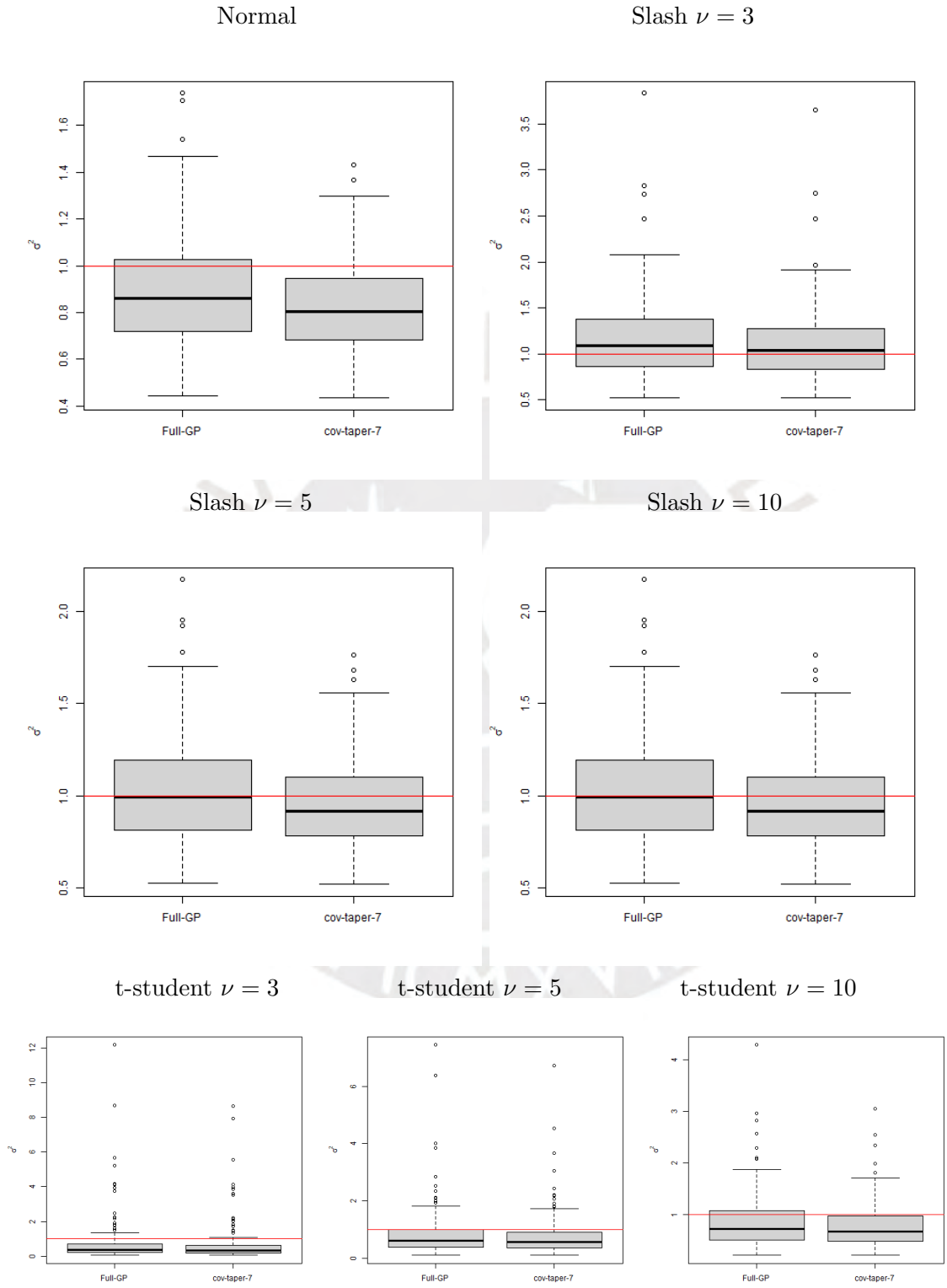


Figura 4.5: Diagramas de caja de 100 estimaciones de la varianza marginal (σ^2) con distribuciones normal, slash y t-student.

Además las estimaciones presentan mayor variabilidad bajo el modelo geoestadístico t-student, siendo mayor esta variabilidad cuando el parámetro ν es menor, es decir cuando las colas son más pesadas. Los resultados obtenidos para τ^2 se muestran en la figura 4.6.

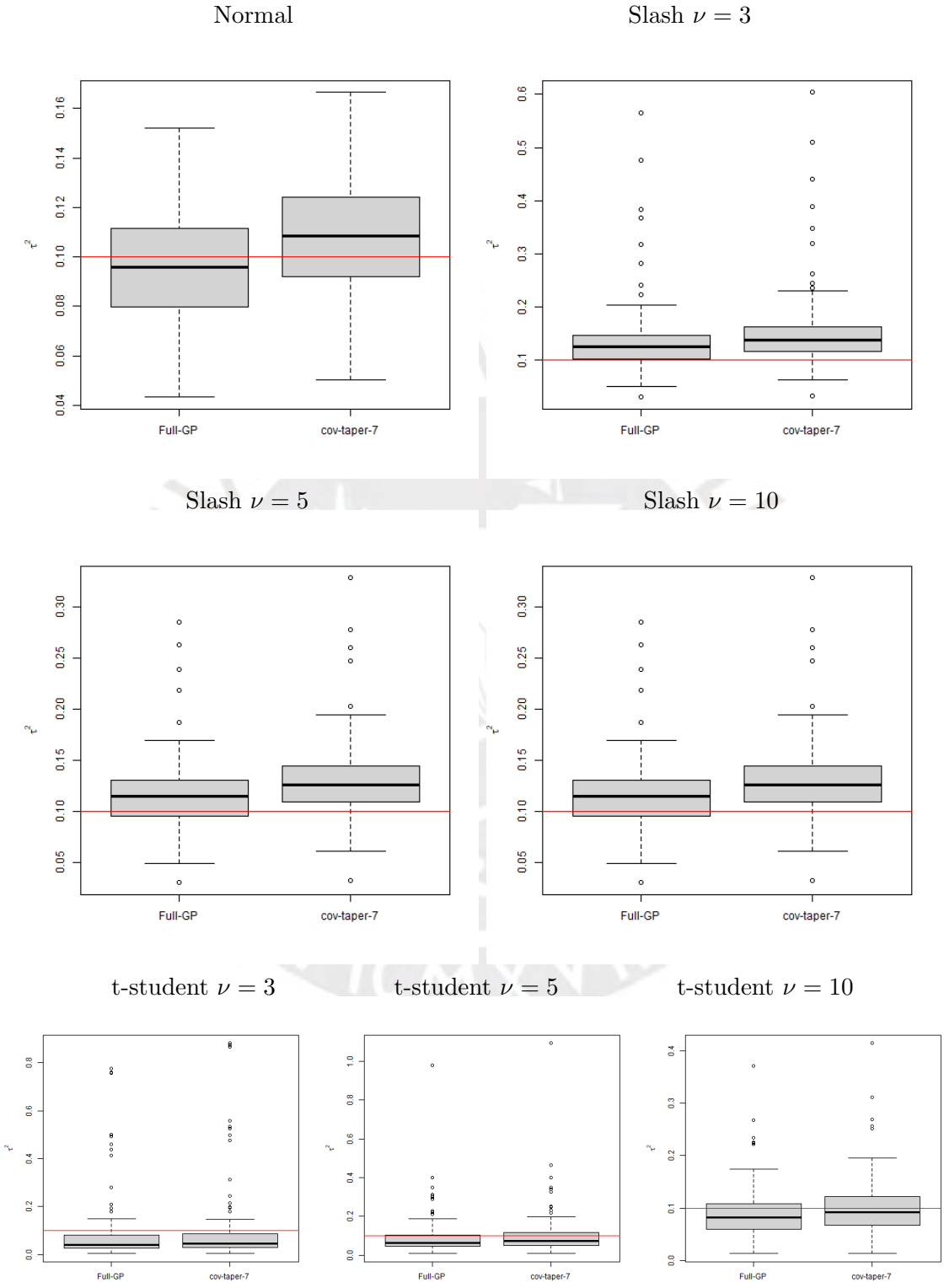


Figura 4.6: Diagramas de caja de 100 estimaciones del efecto pepita (τ^2) con distribuciones normal, slash y t-student.

En general vemos que se ha sobreestimado ligeramente el efecto pepita cuando se usa el tapering (cov-taper 7) pues su valor estimado es ligeramente mayor que 1 en todos los escenarios. Nuevamente también existe mayor variabilidad en las estimaciones de este parámetro para el modelo t-student.

En la tabla 4.1 se puede observar el error cuadrático medio (MSE, por sus siglas en inglés) de la estimación de los parámetros en cada escenario ajustando el modelo Full-GP, mientras que en la tabla 4.2 observamos el error cuadrático medio ajustando el modelo cov-taper-7. Entre los modelos geoestadísticos con diferentes distribuciones, se observa que los modelos gaussianos tienen menor MSE, seguidos por los modelos slash. Los modelos t-student son los que presentan mayor MSE. Comparando los modelos Full-GP y cov-taper-7, según los resultados observamos valores similares para β_0 y β_1. Como se esperaba usando la distancia tapering se tiene un MSE mayor para el rango ϕ, y además se tiene un MSE ligeramente menor para σ^2, τ^2.

Cuadro 4.1: Error cuadrático medio usando el modelo Full-GP.

Modelo	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$
Normal	1.04	0.57	1.68	1.05	1.10
t-student($\nu = 3$)	1.48	3.60	1.70	1.36	1.10
t-student($\nu = 5$)	1.48	3.60	1.70	1.36	1.10
t-student($\nu = 10$)	1.51	1.04	1.69	1.06	1.10
Slash($\nu = 3$)	1.03	0.73	1.58	1.02	1.10
Slash($\nu = 5$)	1.04	0.65	1.63	1.01	1.10
Slash($\nu = 10$)	1.04	0.65	1.63	1.01	1.10

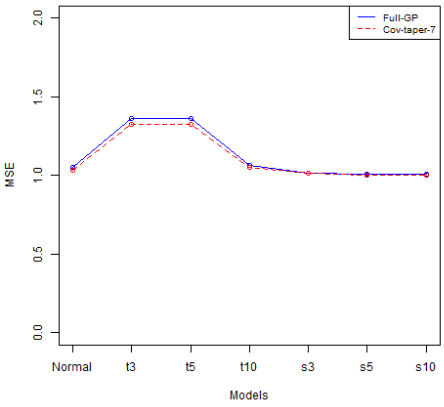
Cuadro 4.2: Error cuadrático medio usando el modelo cov-taper-7.

Modelo	$\hat{\phi}$	$\hat{\sigma}^2$	$\hat{\tau}^2$	$\hat{\beta}_0$	$\hat{\beta}_1$
normal	1.25	0.60	1.65	1.03	1.10
t-student($\nu = 3$)	1.90	2.67	1.68	1.32	1.10
t-student($\nu = 5$)	1.90	2.67	1.68	1.32	1.10
t-student($\nu = 10$)	1.94	0.93	1.67	1.05	1.10
Slash($\nu = 3$)	1.32	0.68	1.54	1.01	1.10
Slash($\nu = 5$)	1.33	0.64	1.60	1.00	1.10
Slash($\nu = 10$)	1.33	0.64	1.60	1.00	1.10

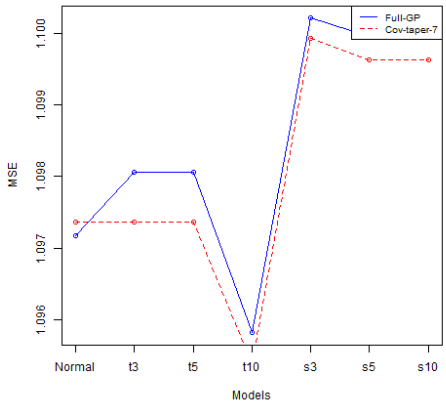
Gráficamente los MSE se resumen en la figura 4.7. El MSE en cada gráfico está en el orden siguiente: normal, t-student($\nu = 3$), t-student($\nu = 5$), t-student($\nu = 10$), slash($\nu = 3$), slash($\nu = 5$) y slash($\nu = 10$). Las líneas rojas son el valor del MSE del modelo cov-taper-7, mientras que la línea azul del Full-GP. Observamos que el MSE para ambos modelos Full-GP y con tapering son parecidos, complementando a lo argumentado sobre los valores estimados de los parámetros, es decir con tapering se obtienen resultados parecidos al modelo con la

matriz de covarianza orginal.

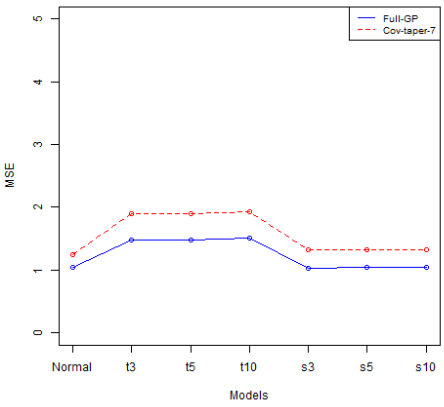
MSE de β_0



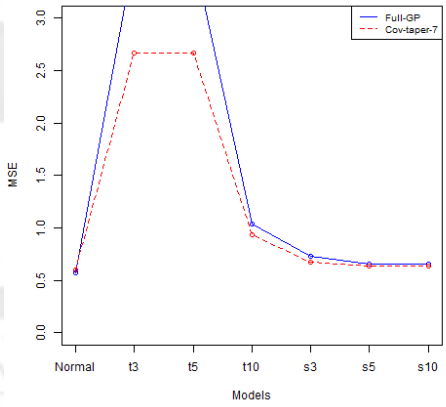
MSE de β_1



MSE de ϕ



MSE de σ^2



MSE de τ^2

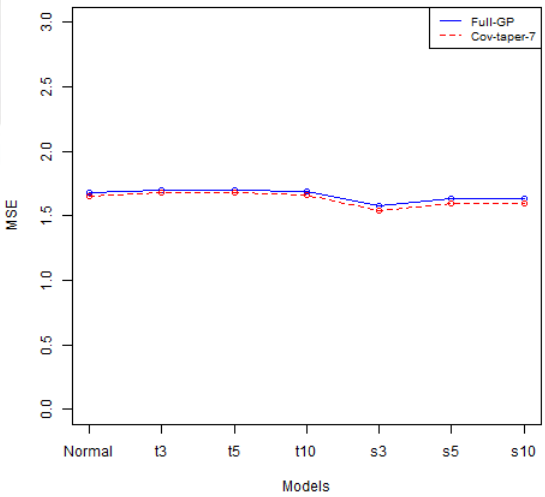


Figura 4.7: Error cuadrático medio de los parámetros ajustando el modelo Full-GP (línea azul) y cov-taper-7 (línea roja) según los modelos gaussiano (normal), t-student con $\nu = 3, 5, 10$ (t_3, t_5, t_{10}) y slash con $\nu = 3, 5, 10$ (s_3, s_5, s_{10}).

4.3.1. Eficiencia computacional

En la tabla 4.3 podemos observar los tiempos promedio de estimación de las 100 réplicas para cada modelo sin usar la covarianza tapering y usando la covarianza tapering. La Figura 4.8 muestra diagramas de caja de los tiempos de estimación para de las 100 réplicas para cada modelo. En general observamos que se tiene una mejor eficiencia en términos de tiempos para los modelos t-student con 5 y 10 grados de libertad y para todos los modelos slash con los tres grados de libertad. Posiblemente se notaría una diferencia sustancial en tiempo computacional si se aumenta el tamaño de muestra y si se optimiza el código implementado en C++.

Cuadro 4.3: Tiempo promedio (en segundos) de estimación para las 100 réplicas de cada modelo.

tipo	modelo	tiempo (seg.)
Full-GP	normal	17.60
Full-GP	t3	32.39
Full-GP	t5	26.28
Full-GP	t10	24.56
Full-GP	slash3	44.11
Full-GP	slash5	43.38
Full-GP	slash10	44.13
cov-taper-7	normal	29.24
cov-taper-7	t3	33.56
cov-taper-7	t5	25.39
cov-taper-7	t10	23.30
cov-taper-7	slash3	43.67
cov-taper-7	slash5	41.19
cov-taper-7	slash10	41.53

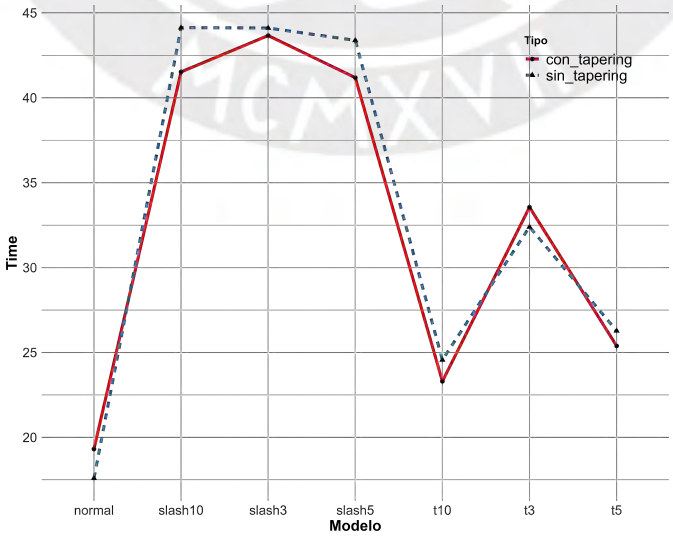


Figura 4.8: Tiempo de estimación promedio de las 100 réplicas de los modelos simulados.

Finalmente se corrió una nueva simulación del modelo slash con 10 grados de libertad pero con distintos tamaños de muestra $n = 750, 1000, 1250, 1500$ para evaluar el tiempo de estimación y usando diferentes valores de distancia tapering, específicamente $\gamma = 3, 5, 7$. para evaluar como impacta la distancia tapering en términos de tiempo de estimación. Los valores de los parámetros son los mismos que los definidos en la simulación anterior. Además evaluar la eficiencia de los modelos usando la covarianza tapering con respecto al modelo sin usar covarianza tapering, también se corrió el modelo slash con 10 grados de libertad sin usar covarianza tapering, es decir el modelo Full-GP. La Figura 4.9 muestra los tiempos de estimación para todos los escenarios, se observa que cuando aumenta el tamaño de muestra, el tiempo de estimación es mayor y sobre todo aumenta en mayor medida para el modelo sin tapering, es decir para el Full-GP. Además se observa que cuando se usa una distancia tapering menor, en este caso igual a 3, el tiempo de estimación siempre es menor con respecto a los modelos ajustados con una distancia tapering mayor.

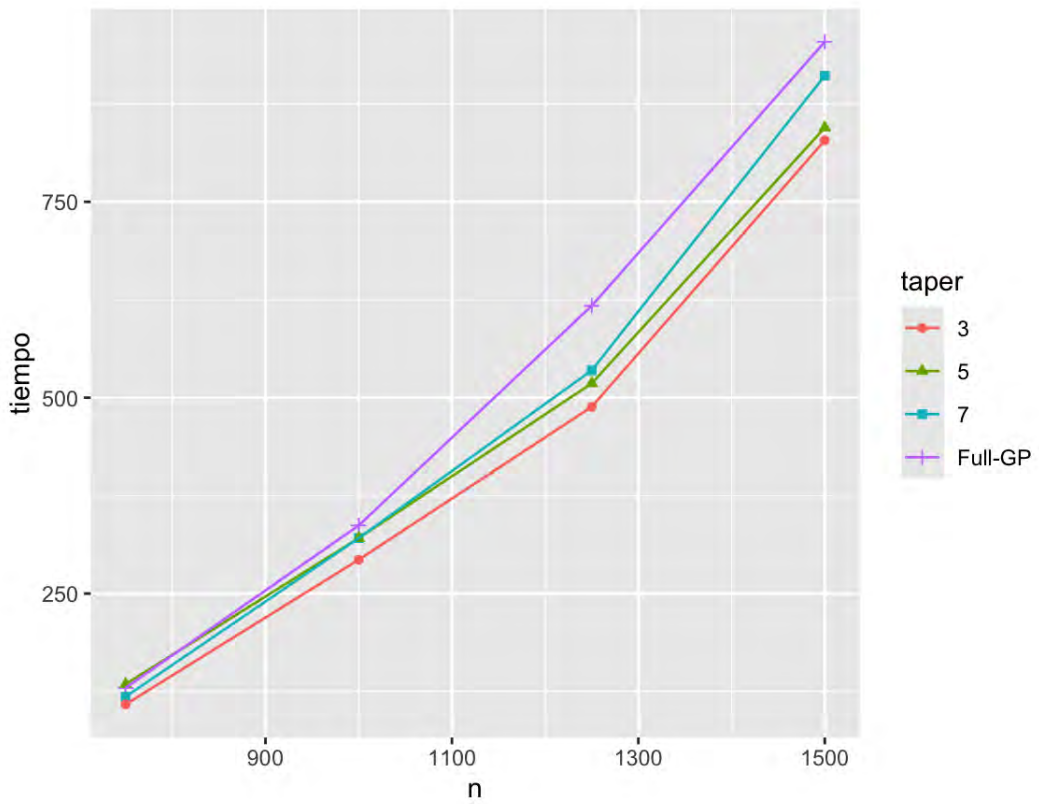


Figura 4.9: Tiempo de estimación de los parámetros para el modelo slash con 10 grados de libertad, bajo diferentes número de tamaño de muestra (n) y distancias tapering igual a $\gamma = 3$ (línea roja), $\gamma = 5$ (línea verde) y $\gamma = 7$ (línea azul). La línea morada representa el tiempo de estimación del modelo Full-GP para los diferentes tamaños de muestra.

Capítulo 5

Aplicación

En el presente capítulo se aplicó el modelo geoestadístico NI y NI con taper a datos de contaminación del aire en los Estados Unidos, estos datos se encuentran disponibles en el sistema de calidad de aire (EPA, por sus siglas en inglés).

La fuente de contaminación más importante son: el gas compuesto de 3 átomos de oxígeno que daña el ozono (O_3), Materia Particulada de 2.5 micrómetros o menor de diámetro ($PM_{2.5}$) y Materia Particulada de 10 micrómetros o menor de diámetro (PM_{10}). Tanto el $PM_{2.5}$ como el PM_{10} dañan la salud de las personas, los principales síntomas palpables son irritación en los ojos, la lengua y la garganta. Estas partículas pueden ser transportadas a través del polvo, el polen, humo de los cigarrillos, etc. Esta investigación se centra en analizar en $PM_{2.5}$ dado que afecta de forma directa a la salud de las personas. Los datos de la contaminación del aire disponibles corresponden a datos promedio anuales de contaminación ambiental (PAC) del aire por partículas de 2.5 micrómetros de diámetro ($PM_{2.5}$) en Estados Unidos del año 2005 al 2014. En el cuadro 5.1 se puede observar el valor mínimo, promedio y máximo de contaminación por $PM_{2.5}$ en estos años.

Cuadro 5.1: Promedio de contaminación ambiental (PAC) del aire por $PM_{2.5}$ en Estados Unidos del año 2005 al 2014.

	Year	n	minPAC	maxPAC	meanPAC
1	2005	1071	3.18	22.87	12.63
2	2006	983	3.22	20.78	11.47
3	2007	941	3.39	22.59	11.89
4	2008	941	3.80	23.16	10.72
5	2009	931	3.42	22.34	9.60
6	2010	914	2.50	17.94	9.84
7	2011	830	2.94	21.60	9.68
8	2012	797	3.37	40.77	9.10
9	2013	776	3.13	22.30	8.85
10	2014	755	2.42	21.77	8.81

Se observa que esta variable puede tomar valores entre 2.42 (valor mínimo en 2013) y 40.77 (valor máximo en 2012). Si analizamos por periodo, observamos que este promedio de contaminación se ha reducido en los últimos años, como podemos observar en la Figura 5.1, pero los valores máximos han aumentado en contraste al promedio.

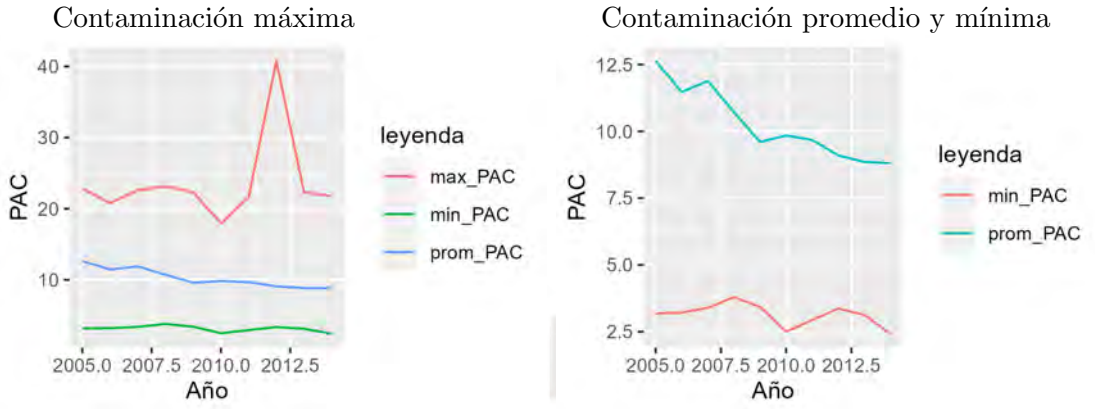


Figura 5.1: Series de tiempo de contaminación máxima (maxPAC), mínima (minPAC) y promedio (meanPAC).

5.1. Descripción de los datos

En esta tesis se analizan los datos de $PM_{2.5}$ en el año 2014. En el cuadro 5.2 observamos que los valores de contaminación varían entre 2.42 y 21.77.

Cuadro 5.2: Estadística descriptiva de $PM_{2.5}$, año 2014.

	Estadística	Valor
1	Min.	2.42
2	1st Qu.	7.53
3	Median	8.93
4	Mean	8.81
5	3rd Qu.	10.00
6	Max.	21.77

La distribución espacial de la variable $PM_{2.5}$ en EEUU en el año 2014 se observa en la Figura 5.2. Vemos que la contaminación ambiental debido a $PM_{2.5}$ se concentra principalmente en las grandes urbes de EEUU, en el extremo oeste y este donde en particular se concentra la población y la actividad económica.

La forma de la distribución de los datos de $PM_{2.5}$ se muestra a través del histograma de la Figura 5.3. Se pueden observar colas más pesadas que en una distribución normal, lo cual es un indicio de que una distribución no gaussiana podría ser más adecuada para modelar estos datos.

Figura 5.2: Datos de $PM_{2.5}$ en EEUU en el año 2014.

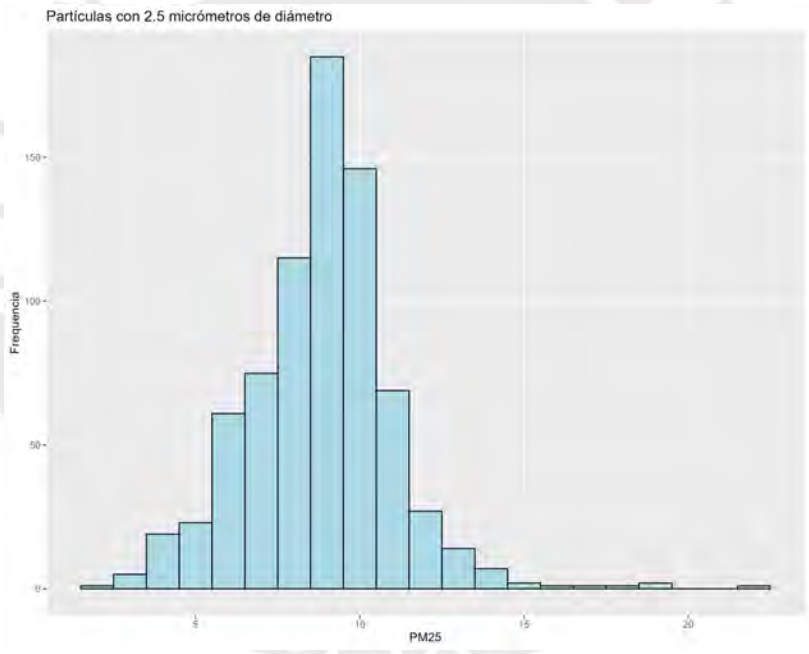


Figura 5.3: Histograma de los datos de $PM_{2.5}$.

5.2. Modelamiento de material particulado PM_{25}

Sea Y_i la variable aleatoria que representa el $PM_{2.5}$ en la estación locals$_i$ donde $i = 1, 2, \ldots, n = 755$. La base de datos se dividió en dos grupos: base de entrenamiento (700 datos seleccionados al azar) y una base de validación (55 datos restantes). Se asume que $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$ sigue una distribución normal independiente, esto es,

$$\mathbf{Y} \sim \mathcal{NI}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, v).$$

De esta forma, podemos modelar los datos a través de una distribución simétrica con colas más pesadas que la distribución normal.

Los media de la distibución en cada local, $\mu(\mathbf{s_i}) = \mathbf{X}(\mathbf{s_i})\boldsymbol{\beta}$, queda especificada como:

$$\mu(s_i) = \beta_0 + \beta_1 \times latitud(s_i).$$

En particular, asumiremos que la función de covarianza es una función exponencial, luego la matriz de covarianza queda definida por $\mathbf{\Sigma} = [\tau^2\mathbf{I} + \sigma^2 e^{-\frac{\boldsymbol{D}}{\phi}}]$, donde $\mathbf{D}$ es la matriz de distancias. La Figura 5.4 muestra la matriz de distancias $\boldsymbol{D}$ para los datos en estudio.

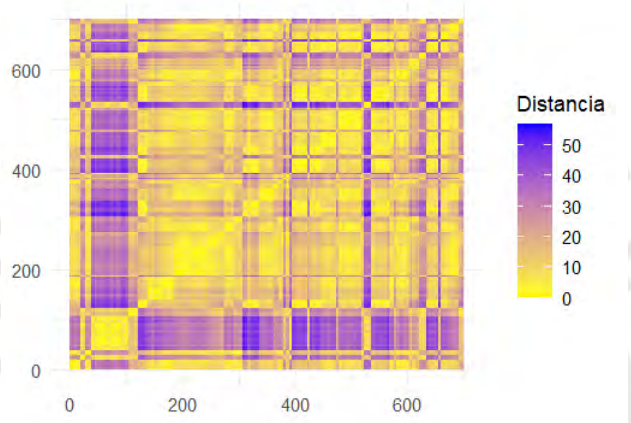


Figura 5.4: Matriz de distancias entre todas las observaciones.

Se propone estimar los parámetros ajustando los modelos geoestadísticos NI, es decir los modelos geoestadísticos normal, t-student con $\nu = 3, 5, 10$ y Slash con $\nu = 3, 5, 10$. Estos modelos se ajustaron en los 700 datos de la base de entrenamiento usando la matriz de covarianza completa y usando una distancia tapering en la matriz de covarianza, es decir se ajustaron los siguientes modelos:

i) Full-GP: El modelo geoestadístico NI usando la matriz de covarianza completa. Es decir se estimaron los parámetros a partir del modelo original.

ii) Cov-taper-7: El modelo geoestadístico NI-taper (cov-taper) usando la aproximación tapering definida en la sección 3.2. Como la distancia máxima entre los locales más lejanos es 50 grados se asumió que la distancia tapering es $\gamma = 7$. Este valor es razonable pues cuando se ajustó el modelo Full-GP bajo las diferentes distribuciones (como se verá más adelante), el range ϕ resultó ser menor a 2 grados.

Luego se estimaron todos los parámetros mediante inferencia clásica, específicamente usando la aproximación de Laplace a través de la implementación de los modelos en TMB. Los resultados obtenidos se muestran en la siguiente sección.

5.3. Resultados

En el cuadro 5.3 se puede observar el el criterio de información Akaike (AIC) y tiempo de estimación de los parámetros bajo los modelos geoestadísticos ajustados. Los valores AIC también se muestran en la Figura 5.5, mientras que tiempos de estimación también se muestran en la Figura 5.6.

Según estos resultados se observa que los valores del AIC son menores para los modelos Full-GP, bajo las diferentes distribuciones, obviamente este resultado es esperado, pues en los modelos FUll-GP, se considera toda la información espacial. Si observamos por tipo de modelo, los modelos Full-GP usando la distribución slash tienen el menor AIC = 2544.51, seguido del modelo normal (AIC=2544.70).

Por otro lado, los modelos cov-taper-7 siguen el patrón observado para los modelos Full-GP. En particular, el modelo cov-taper-7 que resultó ser el mejor es el modelo slash con $\nu = 10$ con el menor AIC = 2559.23, seguido del modelo slash con $\nu = 5$ y $\nu = 3$, respectivamente.

Con respecto a la eficiencia computacional, como era de esperarse los modelos gaussianos son los más eficientes. En particular, los modelos t-student resultaron correr más rápido que los modelos slash tanto para el Full-GP como para el cov-taper-7. Además los modelos slash cov-taper-7 son más eficientes que los modelos full-GP. Mientras que los modelos t-student cov-taper-7 toman más tiempo que sus respectivos modelos competencia sin tapering. Cabe resaltar que si se tuvieran más datos se notaría más la ventaja de usar el tapering en términos de eficiencia computacional.

Cuadro 5.3: AIC y tiempo (en segundos) de estimación para cada modelo ajustado.

tipo	modelo	AIC	tiempo (segundos)
Full-GP	normal	2544.70	23.52
Full-GP	t3	2547.34	35.85
Full-GP	t5	2547.22	32.36
Full-GP	t10	2546.75	45.47
Full-GP	slash3	**2544.51**	97.39
Full-GP	slash5	**2544.51**	97.35
Full-GP	slash10	**2544.51**	97.06
cov-taper-7	normal	2561.84	23.78
cov-taper-7	t3	2564.48	51.30
cov-taper-7	t5	2564.36	58.92
cov-taper-7	t10	2563.89	47.91
cov-taper-7	slash3	2561.64	78.25
cov-taper-7	slash5	2560.62	88.62
cov-taper-7	slash10	**2559.23**	82.34

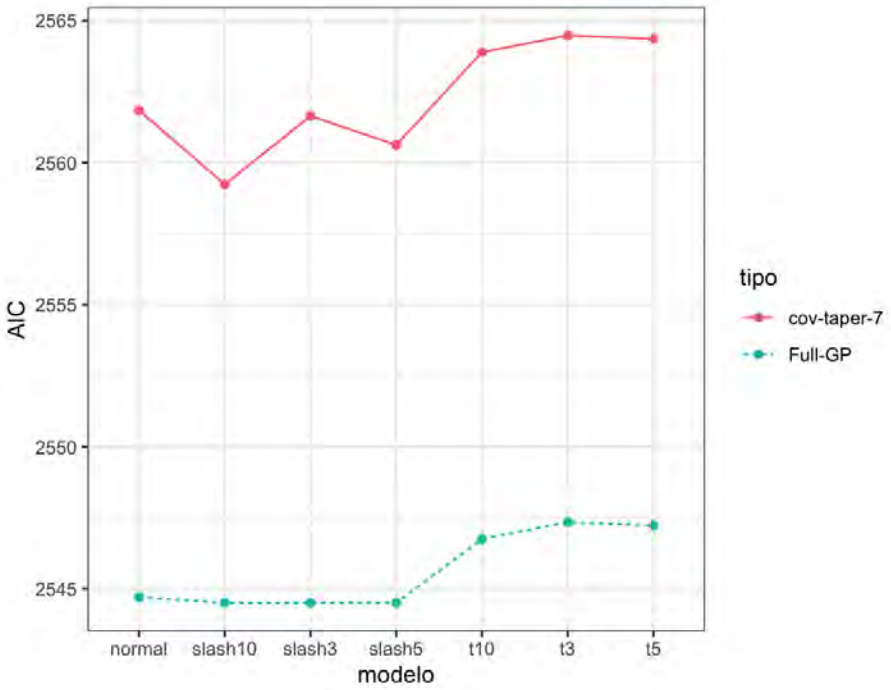


Figura 5.5: Criterio de información AIC de los modelos ajustados.

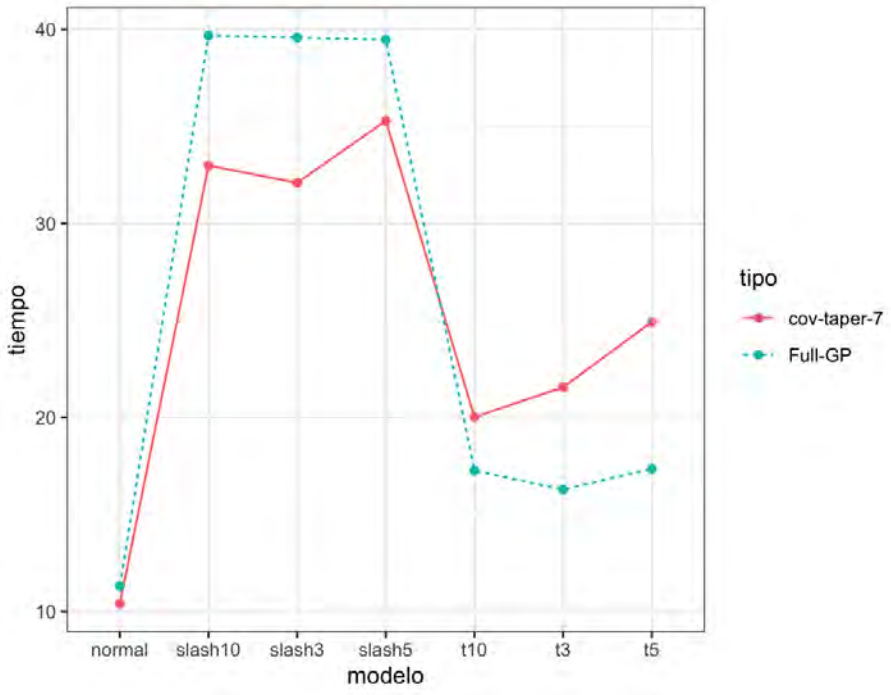


Figura 5.6: Tiempo de estimación de los modelos ajustados.

Una comparación de los valores estimados de los parámetros con tapering (cov-taper-7) y sin tapering (Full-GP) se puede observar en el cuadro 5.4. En general las estimaciones tienen valores son similares para los modelos normal y slash con diferentes grados de libertad. Todos los modelos tienen valores estimados para los parámetros β_0, β_1, y ϕ. Las mayores diferencias

se observan para la varianza marginal σ^2 y el efecto pepita τ^2 entre los modelos t-student con respecto a los modelos normal y slash. Sin embargo como se concluyó antes, los modelos slash son mejores en términos del AIC, seguido del modelo normal.

Cuadro 5.4: Valores estimados (desviación estándar - desv.est.) de los parámetros para los modelos ajustados.

Modelo	Parámetro	Full-GP (desv.est.)	cov-taper-7 (desv.est.)
Normal	β_0	12.157 (1.854)	12.143 (1.471)
Normal	β_1	-0.118 (0.048)	-0.113 (0.038)
Normal	ϕ	1.806 (0.335)	1.745 (0.326)
Normal	σ^2	4.021 (0.540)	3.825 (0.436)
Normal	τ^2	0.757 (0.103)	0.744 (0.103)
t3	β_0	12.157 (1.854)	12.143 (1.471)
t3	β_1	-0.118 (0.048)	-0.113 (0.038)
t3	ϕ	1.806 (0.335)	1.745 (0.326)
t3	σ^2	1.341 (1.904)	1.275 (1.809)
t3	τ^2	0.252 (0.358)	0.248 (0.352)
t5	β_0	12.157 (1.854)	12.143 (1.470)
t5	β_1	-0.118 (0.048)	-0.113 (0.038)
t5	ϕ	1.806 (0.335)	1.745 (0.326)
t5	σ^2	1.609 (1.997)	2.295 (1.892)
t5	τ^2	0.303 (0.376)	0.446 (0.369)
t10	β_0	12.157 (1.854)	12.143 (1.471)
t10	β_1	-0.118 (0.048)	-0.113 (0.038)
t10	ϕ	1.806 (0.335)	1.745 (0.326)
t10	σ^2	0.804 (1.666)	3.060 (1.569)
t10	τ^2	0.151 (0.314)	0.595 (0.309)
Slash(3)	β_0	12.157 (1.854)	12.143 (1.471)
Slash(3)	β_1	-0.118 (0.048)	-0.113 (0.038)
Slash(3)	ϕ	1.806 (0.335)	1.745 (0.326)
Slash(3)	σ^2	4.021 (0.540)	3.825 (0.436)
Slash(3)	τ^2	0.757 (0.103)	0.743 (0.103)
Slash(5)	β_0	12.157 (1.854)	12.143 (1.471)
Slash(5)	β_1	-0.118 (0.048)	-0.113 (0.038)
Slash(5)	ϕ	1.806 (0.335)	1.745 (0.326)
Slash(5)	σ^2	4.021 (0.540)	3.825 (0.436)
Slash(5)	τ^2	0.757 (0.103)	0.744 (0.103)
Slash(10)	β_0	12.157 (1.854)	12.143 (1.471)
Slash(10)	β_1	-0.118 (0.048)	-0.113 (0.038)
Slash(10)	ϕ	1.806 (0.335)	1.745 (0.326)
Slash(10)	σ^2	4.021 (0.540)	3.825 (0.436)
Slash(10)	τ^2	0.757 (0.103)	0.744 (0.103)

Según los valores estimados del modelo slash con $\nu = 10$ grados de libertad, a mayor longitud, es decir más hacia la costa este de Estados Unidos, entonces la media de $PM_{2.5}$ se reduce en 0.118 unidades. Según el rango, aproximadamente el nivel de contaminación de un local depende significativamente del nivel de contaminación de otro local que se encuentra a

una distancia máxima de 1.806 grados. Y como la varianza marginal espacial σ^2 es mucho mayor que el efecto pepita τ^2 en efecto en los datos hay evidencia de variabilidad espacial, finalmente como el efecto pepita es pequeño, resta poca variabilidad no estructurada.

En la Figura 5.7 se comparan los valores reales de la contaminación $PM_{2.5}$ con las estimaciones bajo los diferentes modelos Full-GP ajustados. Observamos que en general tienen relación similar los valores estimados y reales en todos los modelos.

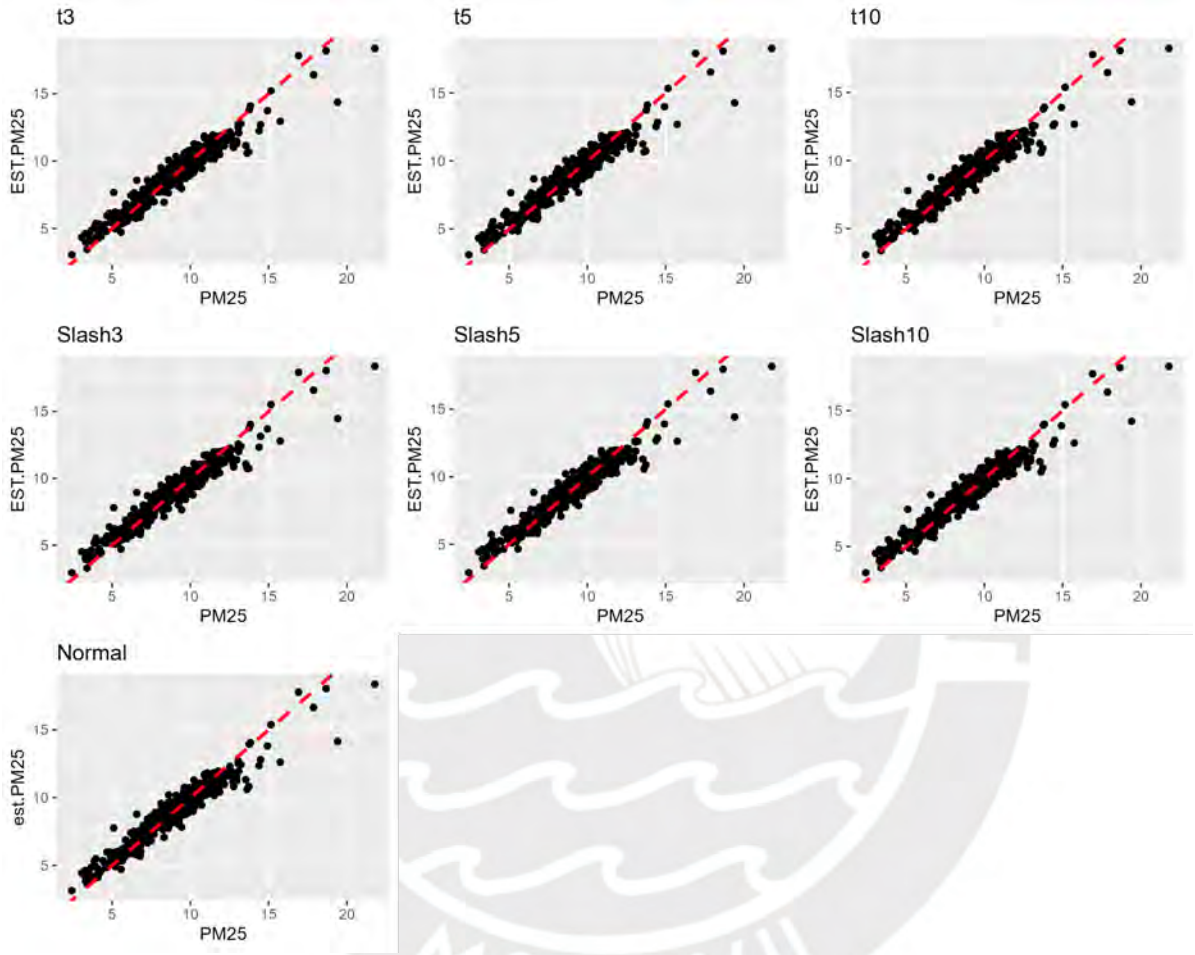


Figura 5.7: Comparación de valores originales de $PM_{2.5}$ versus las estimaciones obtenidas de los modelos Full-GP ajustados. La línea roja representa $x = y$.

De forma similar en la Figura 5.8 se comparan los valores reales de la contaminación $PM_{2.5}$ con las estimaciones bajo los diferentes modelos con tapering (cov-taper-7). Observamos que tienen relación similar los valores estimados y reales en todos los modelos.

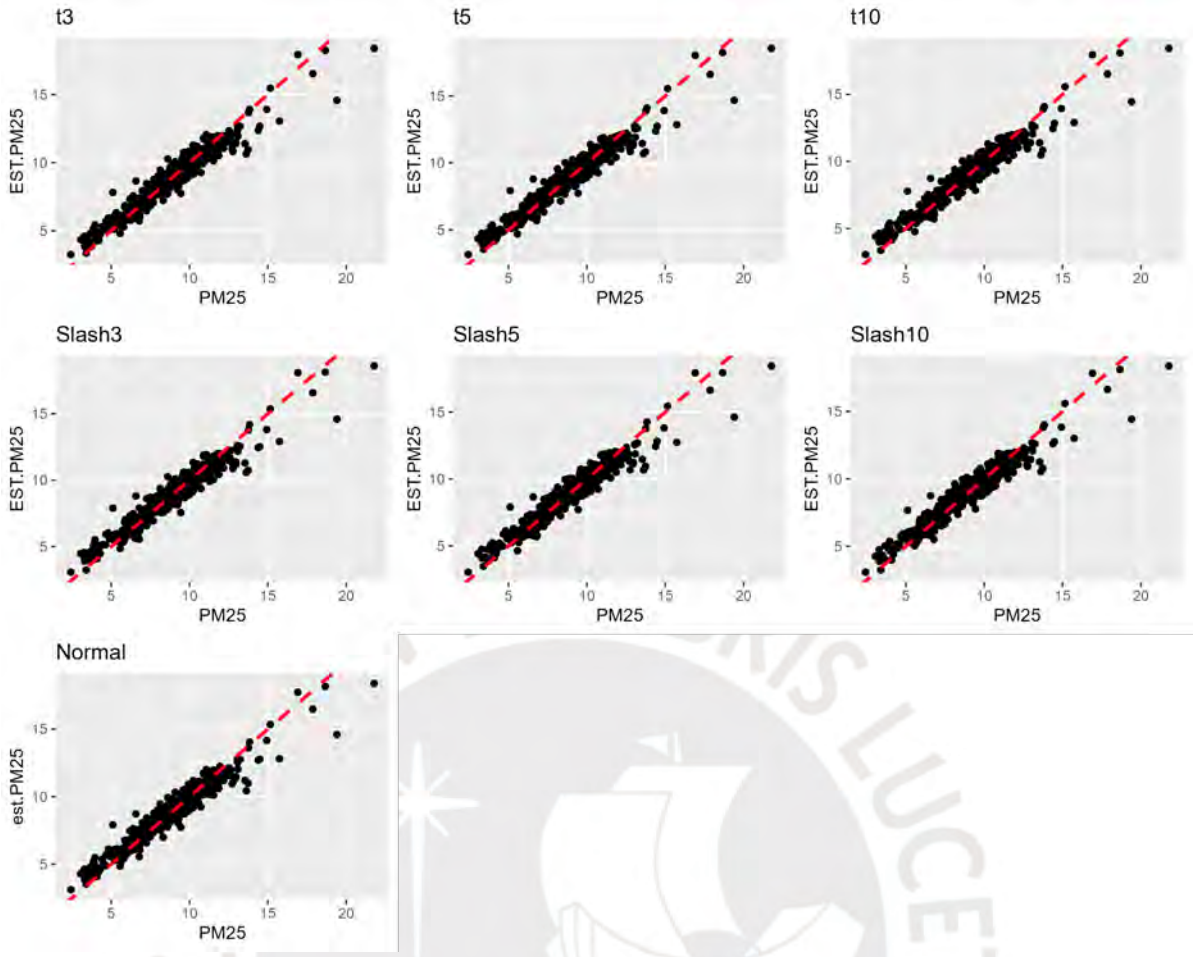


Figura 5.8: Comparación de valores originales de $PM_{2.5}$ versus las estimaciones obtenidas de los modelos con tapering (cov-taper-7). La línea roja representa $x = y$.

La Figura 5.9 muestra mapas de las estimaciones de la variable respuesta para algunos modelos Full-GP (panel izquierdo). Se observa que la estimación de la variable los modelos normal, t-student(10) y slash(10) es muy parecida a los datos originales. La Figura 5.9 también muestra mapas de las estimaciones de la variable respuesta para algunos modelos usando covarianza tapering (cov-taper-7), en el panel del lado derecho. Se observa que la estimación de la variable los modelos normal, t-student(10) y slash(10) es muy parecida a los datos originales.

5.3.1. Predicción

El cuadro 5.5 muestra el error cuadrático medio de predicción (MSP) de $PM_{2.5}$, es decir de Y_i, en base a la base de datos de validación según los modelos ajustados Full-GP y usando la distancia tapering (cov-taper-7). El modelo Full-GP con el menor valor de MSP es el modelo geoestadístico de la distribución t-student ($\nu = 3$). seguido del modelo t-student ($\nu = 5$). Este resultado muestra que por lo menos un modelo con distribución normal no sería adecuado

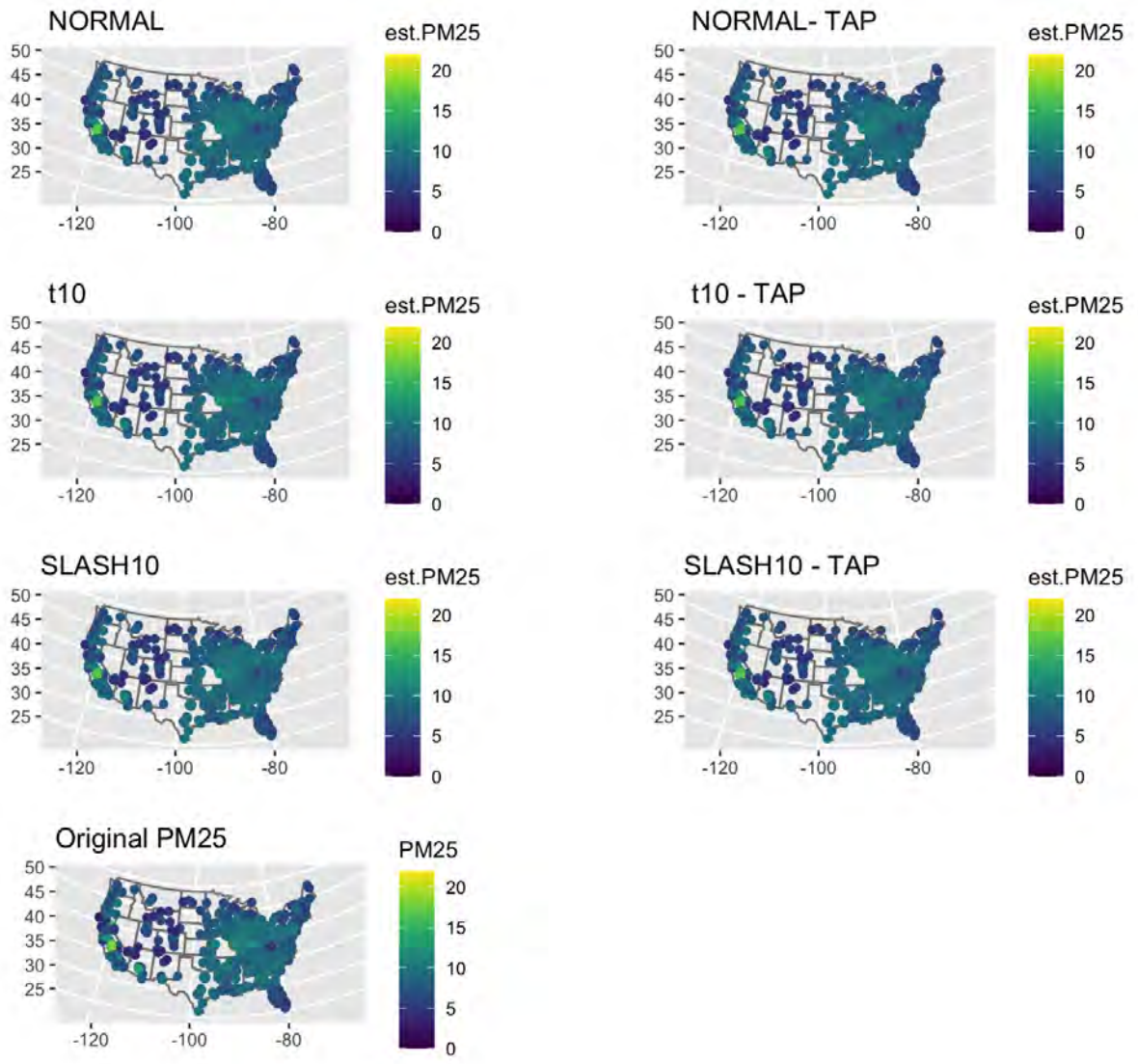


Figura 5.9: Mapa de la variable $PM_{2.5}$ original (Original PM25), y mapas de estimación de $PM_{2.5}$ para los modelos Full-GP (izquierda) y cov-taper-7 (derecha) con distribuciones gaussiana (NORMAL), t-student ($\nu = 10$, t10) y slash ($\nu = 10$, SLASH10).

para estos datos. Por otro lado, de los modelos cov-taper-7, el mejor modelo geoestadístico de la distribución t-student ($\nu = 5$).

En la Figura 5.10 se comparan los valores reales de la contaminación $PM_{2.5}$ con las predicciones bajo los diferentes modelos Full-GP ajustados. Un resultado similar se muestra en la Figura 5.11 donde se comparan los valores reales de la contaminación $PM_{2.5}$ con las predicciones pero bajo los diferentes modelos cov-taper-7 ajustados. En general se observa que los resultados de los modelos Full-GP son similares a los obtenidos usando la covarianza tapering. Y por otro lado aunque no se ve una diferencia crucial entre las distribuciones usando el MSP se concluyó que el modelo normal no sería el más adecuado para estos datos, mostrando la eficiencia de los modelos geoestadísticos no normales para estos datos.

Cuadro 5.5: MSP para cada modelo ajustado.

Tipo	modelo	MSP
Full-GP	normal	1.246
Full-GP	t3	**1.147**
Full-GP	t5	1.305
Full-GP	t10	1.215
Full-GP	slash3	1.298
Full-GP	slash5	1.239
Full-GP	slash10	1.311
cov-taper-7	normal	1.232
cov-taper-7	t3	1.293
cov-taper-7	t5	**1.192**
cov-taper-7	t10	1.258
cov-taper-7	slash3	1.240
cov-taper-7	slash5	1.272
cov-taper-7	slash10	1.246

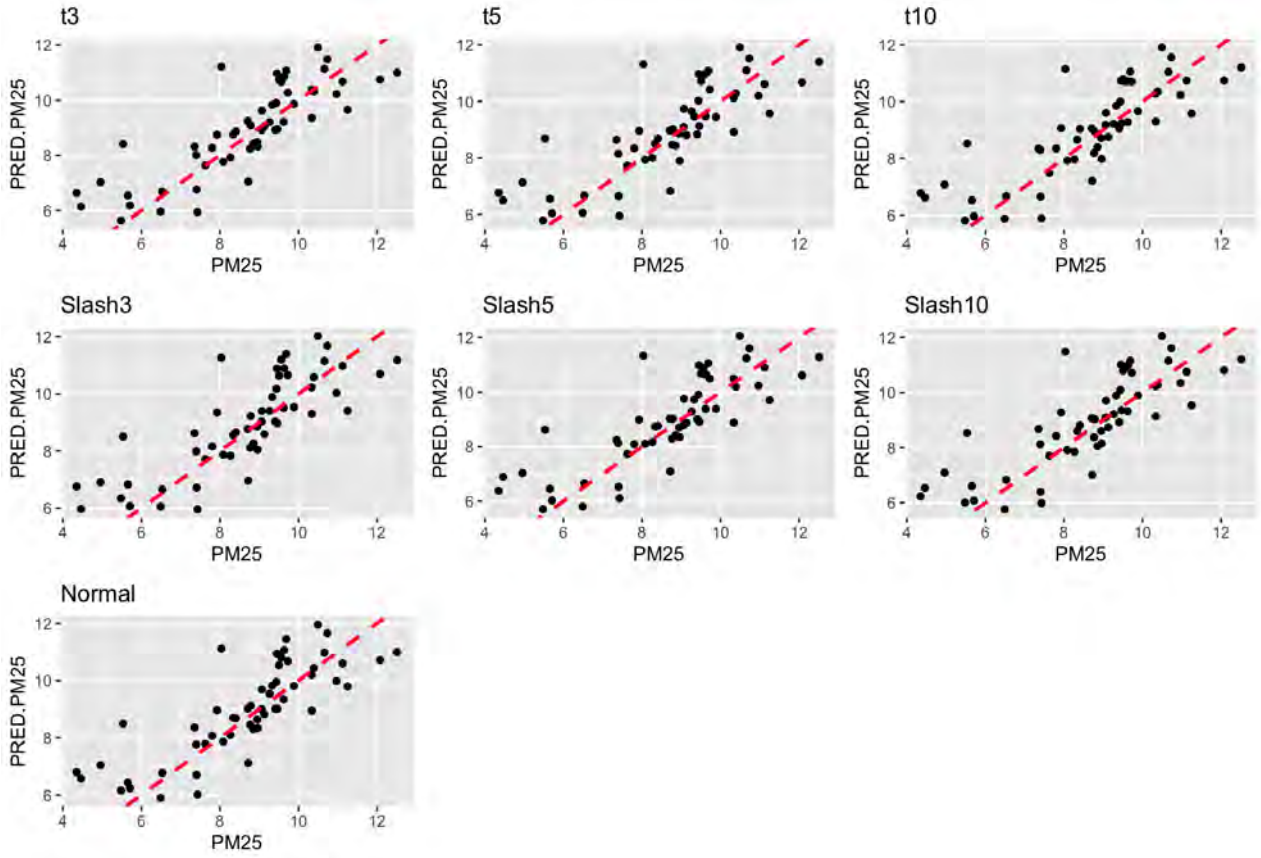


Figura 5.10: Comparación de valores originales de $PM_{2.5}$ versus las predicciones obtenidas de los modelos Full-GP ajustados. La línea roja representa $x = y$.

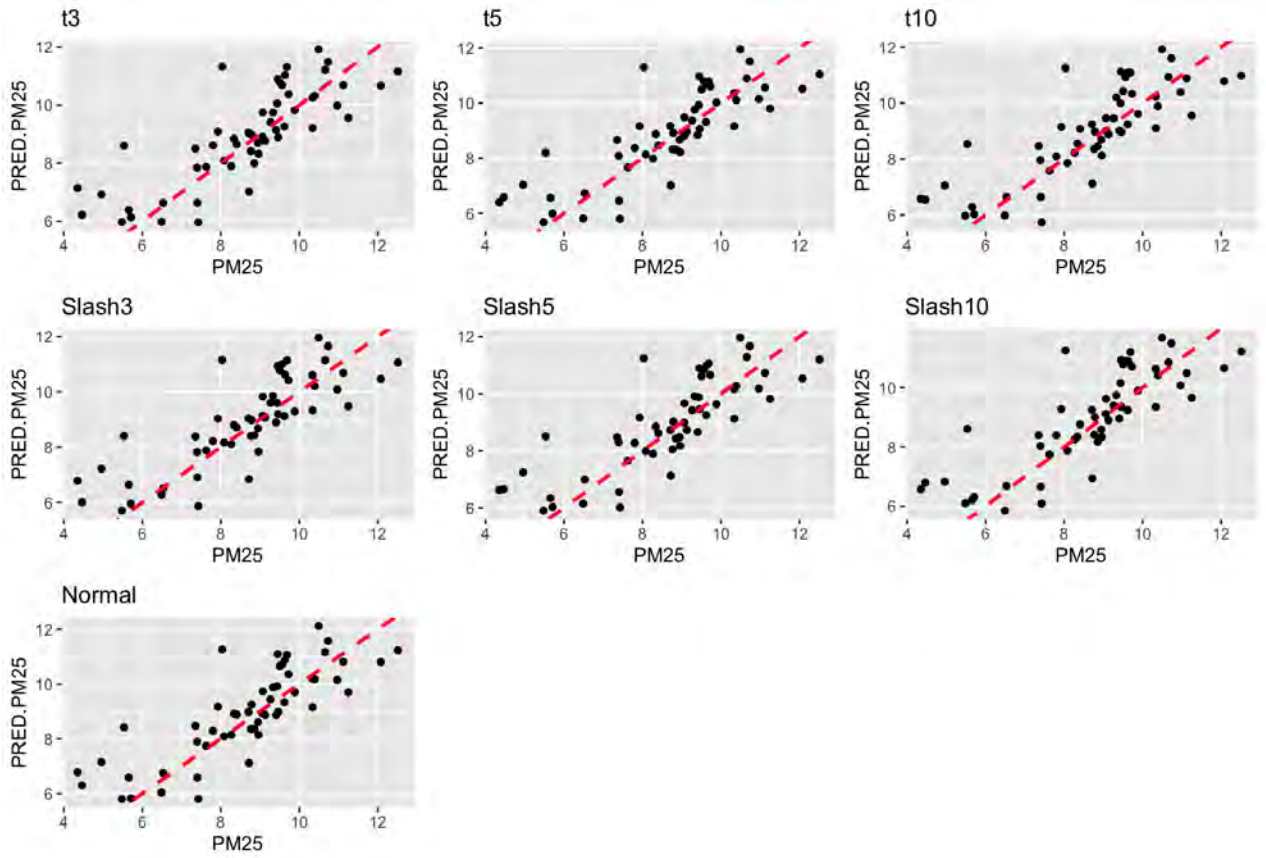


Figura 5.11: Comparación de valores originales de $PM_{2.5}$ versus las predicciones obtenidas de los modelos cov-taper-7 ajustados. La línea roja representa $x = y$.

Capítulo 6

Conclusiones

6.1. Conclusiones

Esta tesis extiende el modelo geoestadístico normal independiente definido por Da-Silva (2017). Como contribuciones puntuales: i) se extiende el modelo usando funciones tapering, para construir matrices de covarianza dispersas, ii) se propone e implementa la inferencia clásica a través de la aproximación de Laplace, y iii) se aplicaron los modelos para analizar datos reales de contaminación ambiental.

Se han realizado simulaciones, utilizando la covarianza completa y con tapering. Para mejorar la eficiencia y evitar problemas numéricos se programaron los modelos propuestos en C++, y se usó aproximación de Laplace a través de la librería en R Template Model Builder. En general se han recuperado los parámetros satisfactoriamente. Observando que en general el tapering afecta ligeramente la estimación de los parámetros espaciales y del nugget.

Por otro lado, se ha utilizado información de contaminación del aire en los Estados Unidos en el 2014, específicamente de material particulado de 2.5 micrómetros de diámetro. Se ajustaron los modelos propuestos, observándose que en efecto las distribuciones no gaussianas se ajustaron mejor a estos datos. Además se han obtenido las predicciones de la variable en estudio de forma satisfactoria, obteniéndose valores parecidos a los originales tanto usando la matriz de covarianza completa como usando el tapering.

6.2. Sugerencias para investigaciones futuras

- Aplicar los modelos propuestos a bases de datos más grandes donde se debería apreciar mejor la contribución del enfoque propuesto.

- La aplicación realizada en esta tesis nos sugiere extensiones a otros campos como pre-

dicción de clima, datos de salud y ciencias sociales como impactos tanto de políticas y teóricas.

- La posibilidad de hacer extensión a modelos bayesianos, modelos espaciales con campos aleatorios de Markov, es decir con matrices de precisión dispersas.

Apéndice A: Códigos en C++

En esta sección se presentan los códigos implementados para tres modelos usando covarianza tapering: normal, t-student con $\nu = 3$ y slash con $\nu = 3$. De forma similar se implementaron los modelos con grados de libertad $\nu = 5$ y $\nu = 10$.

Normal con covarianza tapering

```
// Spatial  NI model, with exponentially decaying correlation function
#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
{
  DATA_INTEGER(n);
  DATA_VECTOR(y);
  DATA_MATRIX(X);
   DATA_MATRIX(Ktaper);
  DATA_MATRIX(dd);
  PARAMETER(a);
  PARAMETER(logsigma2);
  PARAMETER(logtau2);
 PARAMETER_VECTOR(b);

  using namespace density;
  int i,j;
```

```
  vector<Type> residual(n);
  residual = y - X*b;

  matrix<Type> cov(n,n);
  for (i=0;i<n;i++)
  {
    cov(i,i)= (exp(logtau2)+exp(logsigma2));
    for ( j=0;j<i;j++)
    {
      cov(i,j)=exp(logsigma2)*exp(-dd(i,j)/a)*Ktaper(i,j);
      // Exponentially decaying correlation
      cov(j,i)=cov(i,j);
    }
  }

  MVNORM_t<Type> neg_log_density(cov);
  Type res = neg_log_density(residual);

// reports on transformed parameters
  ADREPORT (exp(logsigma2));
  ADREPORT(exp(logtau2));

  return res;

}
```

T-student con nu=3 y covarianza tapering

```
// Spatial  NI model, with exponentially decaying correlation function

#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
```

```
{
  DATA_INTEGER(n);
  DATA_VECTOR(y);
  DATA_MATRIX(X);
  DATA_MATRIX(dd);
  DATA_MATRIX(Ktaper);
   PARAMETER(a);
  PARAMETER(logsigma2);
  PARAMETER(logtau2);
  PARAMETER_VECTOR(b);
  PARAMETER(u);

  using namespace density;
  int i,j;

  vector<Type> residual(n);
  residual = y - X*b;

  matrix<Type> cov(n,n);
  for (i=0;i<n;i++)
  {
    cov(i,i)= (exp(logtau2)+exp(logsigma2))/u;
    for ( j=0;j<i;j++)
    {
      cov(i,j)=exp(logsigma2)*exp(-dd(i,j)/a)*Ktaper(i,j)/u;
      // Exponentially decaying correlation
      cov(j,i)=cov(i,j);
    }
  }

  MVNORM_t<Type> neg_log_density(cov);
  Type res = neg_log_density(residual);
   res -= -lgamma(Type(1.5))+ (Type(0.5))*log(u) -(Type(1.5)*u) +
   Type(1.5)*log(Type(1.5));
```

```
// reports on transformed parameters
  ADREPORT (exp(logsigma2));
  ADREPORT(exp(logtau2));

  return res;

}
```

Slash con nu=3 y covarianza tapering

```
// Spatial  NI model, with exponentially decaying correlation function

#include <TMB.hpp>

template<class Type>
Type objective_function<Type>::operator() ()
{
  DATA_INTEGER(n);
  DATA_VECTOR(y);
  //DATA_VECTOR(X);
  DATA_MATRIX(X);
  DATA_MATRIX(dd);
  DATA_MATRIX(Ktaper);
  PARAMETER(a);
  PARAMETER(logsigma2);
  PARAMETER(logtau2);
  PARAMETER_VECTOR(b);
  PARAMETER(logitu);

  using namespace density;
  int i,j;

  vector<Type> residual(n);
```

```
  residual = y - X*b;

  matrix<Type> cov(n,n);
  for (i=0;i<n;i++)
  {
    cov(i,i)= (exp(logtau2)+exp(logsigma2))/invlogit(logitu);
    for ( j=0;j<i;j++)
    {
      cov(i,j)=exp(logsigma2)*exp(-dd(i,j)/a)*Ktaper(i,j)/invlogit(logitu);
       // Exponentially decaying correlation
      cov(j,i)=cov(i,j);
    }
  }

  MVNORM_t<Type> neg_log_density(cov);
  Type res = neg_log_density(residual);
res -= lgamma(Type(4)) - lgamma(Type(3)) - lgamma(Type(1))  +
Type(2)*log(invlogit(logitu)); //+ Type(beta-1)*log(1-u);

// reports on transformed parameters
  ADREPORT (exp(logsigma2));
  ADREPORT(exp(logtau2));
  ADREPORT(invlogit(logitu));

  return res;

}
```

Bibliografía

Amiri, A. y Gerdtham, U.-G. (2011). Relationship between exports, imports and economic growth in France: evidence from cointegration analysis and granger causality with using geostatistical models, *Technical report*, University Library of Munich, Germany.

Banerjee, S., Carlin., B. P. y Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data*, CRC Press.

Claramunt, C. y Stewart, K. (2015). Special issue on spatio-temporal theories and models for environmental, urban and social sciences: where do we stand?, *Spatial Cognition and Computation* **15**(2): 61–67.

Cressie, N. y Wikle, C. K. (2011). *Statistics for spatial Data*, Wiley Series.

Da-Silva, D. M. (2017). *Modelos espaciais com distribuição normal independente*, Master's thesis, Universidade Federal de Minas Gerais.

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.

Ding, P. (2016). On the conditional distribution of themultivariate t distribution, *Journal of Multivariate Analysis* **70**(3): 293–295.

Furrer, R., Genton, M. G. y Nychka, D. (2006). Covariance tapering for interpolationof large spatial datasets, *Journal of Computational and Graphical Statistics* .

Gelfand, A., Digle, P., Fuentes, M. y Guttorp, P. (2010). *Handbook of spatial statistics*, Chapman and Hall/CRC Handbooks of Modern Statistical Methods.

Kaufman, C. G., Schervish, M. J. y Nychka, D. W. (2008). Covariance tapering for likelihood based estimation in large spatial data sets, *Journal of the American Statistical Association* **103**(484): 1545–1555.

Kent, J. T. y Mardia, K. V. (2022). *Spatial Analysis*, Wiley Series in Probability and Statistics.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H. y Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation, *Journal of Statistical Software.* **70**(5): 1–21.

Lachos, V. H. y Labra, F. V. (2014). Multivariate skew-normal/independent distributions: properties and inference, *Pro Mathematica, 28, 56 (2014), 11-53* .

Lange, K. y Sinsheimer, J. S. (1993). Normal/independent distributions and their applications in robust regression, *Taylor and Francis, Ltd. on behalf of the American Statistical Association* **2**(2): 175–198.

Lawson, A. (2018). *Bayesian Disease Mapping: Hierarchical Modeling in Spatial Epidemiology*, CRC Press.

Lindgren, F.; Rue, H. y Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The SPDE approach, *Journal of the Royal Statistical Society. Series B. Statistical Methodology.* **73**(4): 423–498.

Quiroz, Z. C. y Prates, M. O. (2018). Bayesian spatio-temporal modelling of anchovy abundance through the SPDE approach, *Spatial Statistics* **28**: 236–256.

Roth, M. (2013). On the multivariate t distribution, *Technical Report 3059*, Linköping Universityoping University.