ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ciencias Naturales y Matemáticas

Inferencia de datos espaciales con respuestas censuradas o faltantes usando ecuaciones diferenciales parciales estocásticas (SPDE)

Proyecto Integrador

Previo a la obtención del Título de:
Ingeniero en Estadística

Presentado por:
PABLO ADRIÁN ZÚÑIGA GUAMÁN

Guayaquil - Ecuador

2025

Dedicado a mis padres, Betty y Reinaldo. No hubo ni habrá motivación más grande que pensar en ellos, mi razón de ser.

Agradecimientos

A Dios, por su propósito en mí.

A mis padres y hermanos, por el sacrificio, apoyo y amor incondicional.

A mis profesores por las enseñanzas y conocimientos impartidos.

A mi tutora Katherine, por su guía y aporte en la elaboración de este trabajo.

A mis amigos, por las historias y vivencias que quedan marcadas en la memoria.

Declaración Expresa

"Los derechos de titularidad y explotación, me corresponden conforme al reglamento de propiedad intelectual de la institución. Zúñiga Guamán Pablo Adrián, doy mi consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual."

Firmado electrónicamente por:
PABLO ADRIAN ZUNIGA GUAMAN

Zúñiga Guamán Pablo Adrián

Evaluadores

Firmado electrónicamente por:
SANDRA LORENA
GARCIA BUSTOS

Ph.D. García Bustos Sandra Lorena

Profesor de Materia

Firmado electrónicamente por:
KATHERINE ANDREINA
LOOR VALERIANO

Ph.D. Loor Valeriano Katherine

Tutor de proyecto

Resumen

En escenarios realistas, cuando estamos en presencia de datos censurados o faltantes, la estadística espacial enfrenta desafíos. No tratar estos datos de forma adecuada compromete la precisión e introduce sesgo en las estimaciones. Este problema es frecuente en aplicaciones donde las limitaciones de los instrumentos de medición o la falta de reportes generan datos incompletos.

Si bien existen enfoques frecuentistas y bayesianos para abordar esta dificultad, su aplicación en grandes volúmenes de datos sigue siendo un reto debido al elevado costo computacional asociado al manejo de matrices de varianzas y covarianza que son densas. Para afrontar esta problemática, este estudio propone un modelo espacial basado en verosimilitud, diseñado específicamente para gestionar respuestas censuradas o faltantes mediante la aproximación de campos aleatorios gaussianos (GRF) por campos aleatorios gaussianos de Markov (GMRF), obtenidos como solución de ecuaciones diferenciales parciales estocásticas (SPDE). Además, se emplea el algoritmo Stochastic Approximation Expectation-Maximization (SAEM) para la estimación eficiente de los parámetros del modelo en presencia de datos censurados o faltantes. Este algoritmo combina métodos de aproximación estocástica y maximización, permitiendo obtener estimaciones de máxima verosimilitud mediante una exploración eficiente del espacio de parámetros latentes. Su estructura iterativa evita la necesidad de calcular integrales de alta dimensión de manera explícita, lo que lo hace especialmente adecuado para modelos con datos incompletos. Esta estrategia, junto con la representación de un GRF a través de un GMRF, reduce la carga computacional al introducir esparcidad en las matrices involucradas, permitiendo una modelización más óptima en términos de costo computacional y precisión en la estimación de parámetros.

Palabras claves: Algoritmo EM, Algoritmo SAEM, Datos censurados, Datos Faltantes, Esparcidad, SPDE.

Abstract

In real-world applications, spatial statistics encounter significant challenges when handling censored or missing data. Inadequate data treatment can compromise accuracy and introduce biases into estimates. This problem frequently arises in scenarios marked by limitations of measurement instruments or reporting gaps that result in incomplete datasets.

While both frequentist and Bayesian methodologies are available to address these challenges, applying them to large datasets remains problematic due to the significant computational cost associated with managing dense variance-covariance matrices. To overcome this problem, we propose a spatial likelihood-based model specifically designed to handle censored or missing responses. This model approximates Gaussian Random Fields (GRF) with Gaussian Markov Random Fields (GMRF), which are derived as a solution of a Stochastic Partial Differential Equations (SPDE). Furthermore, we employ the Stochastic Approximation Expectation-Maximization (SAEM) algorithm for efficient parameter estimation in scenarios involving censored or missing data. This algorithm combines stochastic approximation and maximization techniques, which enables the computation of maximum likelihood estimates through efficient exploration of the latent parameter space. Its iterative framework eliminates the need for explicit calculations of high-dimensional integrals, making it particularly suitable for models with incomplete data. This approach, in conjunction with the GMRF representation, significantly reduces the computational burden by introducing sparsity into the involved matrices, allowing for more optimal modeling in terms of computational cost and accuracy in parameter estimation.

Keywords: Censored data, EM algorithm, Missing data, SAEM algorithm, Sparsity, SPDE.

Índice de figuras

Índice de cuadros

Lista de abreviaturas y siglas

AIC	Akaike information criterion
BIC	Bayesian information criterion
CRAN	Comprehensive R Archive Network
DMQ	Distrito Metropolitano de Quito
EM	Expectation-Maximization
ESPOL	Escuela Superior Politécnica del Litoral
GMRF	Gaussian Markov Random Field
GRF	Gaussian Random Field
INLA	Integrated Nested Laplace Approximation
MC	Monte Carlo
MCEM	Monte Carlo EM
MSE	Mean Squared Error
SAEM	Stochastic Approximation EM
SPDE	Stochastic Partial Differential Equation

Índice general

1 Introducción

La estadística espacial es una rama esencial de la Estadística, que permite describir una variedad de modelos y métodos para el análisis de datos georreferenciados. Su relevancia radica en la capacidad de integrar la localización geográfica como un componente clave del análisis, permitiendo no solo estudiar cómo varía un fenómeno en función de distintas condiciones, sino también identificar dónde ocurren dichos cambios. En muchos estudios, es común suponer que el proceso espacial subyacente es un campo aleatorio gaussiano (Gaussian Random Field, GRF), debido a las propiedades matemáticas convenientes que ofrecen los GRF, como la determinación completa de su distribución mediante la media y la función de covarianza (Hristopulos, 2020). Esto implica que cualquier subconjunto finito del campo sigue una distribución conjunta gaussiana, definida únicamente por estos parámetros, lo que simplifica el modelado y la inferencia en análisis espaciales.

Los datos espaciales pueden presentarse en múltiples formatos, dependiendo de la naturaleza del fenómeno bajo estudio. Podemos encontrar datos areales (areal data), donde las observaciones se asocian a regiones geográficas discretas; datos de patrones de puntos (point pattern data), que representan la presencia o ausencia de eventos en localizaciones específicas; datos puntuales (point data), caracterizados por su alta resolución y precisión geográfica; datos espacio-temporales, que incorporan una dimensión temporal para modelar fenómenos dinámicos; y, datos funcionales espaciales, que describen valores continuos a través del espacio. Una discusión más detallada sobre tipos de datos espaciales está disponible en Cressie (2015) y Gaetan *et al.* (2010).

La estadística espacial ha encontrado aplicaciones en áreas tan diversas como en el estudio de patrones de precipitación y otros eventos climáticos (Wilks, 2011; Juntto & Paatero, 1994), el mapeo de enfermedades y la distribución de factores de riesgo (Lawson, 2013), el análisis del mercado inmobiliario y la valoración de propiedades (Anselin, 2013; McIlhatton *et al.*, 2016), el monitoreo de la calidad del aire (Brody *et al.*, 2004), y optimización del uso de recursos en estudios de suelo (Griffin, 2010), por mencionar algunas. Estas aplicaciones no solo destacan su versatilidad, sino también la importancia de llevar en consideración la correlación espacial como un elemento fundamental para la comprensión y modelado de fenómenos complejos.

Sin embargo, su aplicabilidad se ve limitada por dos situaciones adversas que se abordarán en este trabajo. Actualmente, los datos se generan constantemente, a diario, por minuto o incluso por segundo, lo que resulta en grandes volúmenes de

datos disponibles para su análisis. Esto ha generado un creciente interés por desarrollar nuevos modelos que, por un lado, sean capaces de capturar la complejidad inherente a los datos y, por otro, resulten computacionalmente viables. En el caso de los modelos espaciales, uno de los principales desafíos radica en gestionar las matrices de varianzas y covarianzas, que suelen ser densas. Esto implica un alto costo computacional debido a la complejidad algorítmica de las operaciones involucradas. Esta dificultad, conocida en la literatura como el "Big n problem" (Banerjee *et al.*, 2008), ha dado lugar a diversas propuestas, las cuales son revisadas exhaustivamente en Sun *et al.* (2012). Una estrategia eficiente y de creciente popularidad es la aproximación del campo aleatorio gaussiano mediante un campo aleatorio gaussiano de Markov (Gaussian Markov Random Field, GMRF), el cual se obtiene como solución de una ecuación diferencial parcial estocástica (Stochastic Partial Differential Equation, SPDE) (Lindgren *et al.*, 2011). Esta aproximación permite reducir la complejidad computacional al introducir esparcidad en las matrices involucradas.

Otro desafío común surge cuando los datos son censurados (parcialmente observados) o faltantes. Los datos censurados suelen deberse a limitaciones en los instrumentos de medición, mientras que los datos faltantes aparecen cuando ciertos valores no han sido registrados. La presencia de este tipo de datos, si no se considera adecuadamente, puede comprometer tanto la precisión como la robustez de los modelos espaciales. Por ejemplo, en el monitoreo de contaminantes ambientales, las mediciones suelen estar limitadas por la sensibilidad de los instrumentos (Sun *et al.*, 2012). De manera similar, en el mapeo de enfermedades, la falta de reportes precisos genera datos faltantes, afectando la inferencia estadística y la validez de los resultados obtenidos debido al incremento del sesgo y varianza de los estimadores (Lawson, 2006). Aunque se han desarrollado enfoques prometedores para abordar estas dificultades, como los métodos basados en verosimilitud presentados por Ordoñez *et al.* (2018) o enfoques en el contexto bayesiano, como el trabajo de Fridley & Dixon (2007), estas metodologías todavía enfrentan limitaciones importantes. En particular, no están optimizadas para manejar grandes volúmenes de datos, lo que restringe su aplicabilidad en escenarios prácticos.

En el trabajo recientemente presentado por Sahoo *et al.* (2024), se propone un modelo bayesiano espacial, diseñado para manejar respuestas censuradas a la izquierda, logrando además una eficiente gestión de grandes volúmenes de datos. Inspirados por esta contribución y ante la ausencia de un modelo basado en verosimilitud con características similares, esta tesis propone un modelo espacial para lidiar con respuestas censurados o faltantes, que permite considerar múltiples tipos de censura (izquierda, derecha o intervalar) y que, al mismo tiempo, sea eficaz para tratar con grandes

conjuntos de datos al aproximar el GRF a través de un GMRF que es obtenido como solución de una SPDE. Enfrentar esta problemática es esencial para garantizar la robustez y extender aún más la aplicabilidad de los modelos espaciales en estudios del mundo real, donde la complejidad y las restricciones de los datos suelen representar desafíos significativos.

1.1 Objetivos

1.1.1 Objetivo General

Proponer un modelo espacial basado en verosimilitud, que integre la aproximación de campos aleatorios gaussianos (GRF) por campos aleatorios gaussianos de Markov (GMRF), obtenidos como solución de ecuaciones diferenciales parciales estocásticas (SPDE), junto con el algoritmo Stochastic Approximation Expectation-Maximization (SAEM), para estimar parámetros de manera eficiente en presencia de respuestas censurados o faltantes, reduciendo la carga computacional y evitando sesgos en las estimaciones.

1.1.2 Objetivos Específicos

- Evaluar las limitaciones computacionales asociadas a los enfoques tradicionales en estadística espacial al manejar grandes volúmenes de datos censurados o faltantes, con énfasis en el manejo de matrices de covarianza densas.
- Implementar un modelo espacial que combine la representación de GRF mediante GMRF (vía SPDE) con el algoritmo SAEM, optimizando la gestión de datos incompletos mediante aproximaciones estocásticas.
- Validar la eficiencia computacional y precisión del modelo propuesto frente a métodos tradicionales, cuantificando la reducción en complejidad algorítmica y la precisión en la estimación de parámetros bajo escenarios realistas de censura o ausencia de datos.

1.2 Disposición del trabajo

Este trabajo se encuentra dividido en los siguientes capítulos.

Capítulo 2 aborda el modelo de regresión lineal espacial para datos completamente observados. Al mismo tiempo que se describe brevemente al algoritmo EM y SAEM. Así como también se describen los datos que serán analizados posteriormente.

Capítulo 3 presenta el modelo de regresión lineal espacial para respuestas censuradas y faltantes. Se aborda también la aproximación de este proceso a través de un GMRF que es solución de una ecuación diferencial parcial estocástica. Se muestra una aproximación de la matriz de información observada obtenida a través del método propuesto por Louis (1982). Finalizamos el capítulo mostrando las expresiones para predecir el estado del proceso en localidades no observadas.

Capítulo 4 presenta los resultados de los estudios de simulación diseñados para examinar las propiedades asintóticas de las estimaciones obtenidas mediante el modelo propuesto. Además, se evalúa el efecto del número de nodos empleados en la estimación de los parámetros. Finalmente, se incluye el análisis de un conjunto de datos reales, cuya descripción se proporciona brevemente en el Capítulo 2.

Capítulo 5 presenta las conclusiones obtenidas, resume la producción técnica realizada y las orientaciones para trabajos futuros.

2 Preliminares

En este capítulo se introduce el modelo espacial de manera general, acompañado de las definiciones fundamentales para la construcción de la matriz de varianzas y covarianzas asociada. Asimismo, se ofrece una descripción de los algoritmos Expectation-Maximization (EM) y Stochastic Approximation EM (SAEM), siendo este último utilizado en capítulos posteriores para estimar los parámetros del modelo. Finalmente, se describe el conjunto de datos que será utilizado como caso de estudio real para ilustrar la aplicación del modelo propuesto.

2.1 Modelo Espacial

Los datos espaciales, se caracterizan por estar geográficamente referenciados, estos pueden interpretarse como una realización del proceso estocástico espacial continuo

$$\{\mathbf{Y}(\mathbf{s}) : \mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d\} \tag{2.1}$$

caracterizado por un índice $\mathbf{s}$ que varía de forma continua en una región fija $\mathcal{D} \subset \mathbb{R}^d$, comúnmente se utiliza $d = 2$, mientras que, $Y(\mathbf{s})$ denota el estado del proceso en la localización $\mathbf{s}$.

El proceso estocástico definido en (2.1) es un campo aleatorio gaussiano (GRF) si, para cualquier $n \geq 1$ y para cada subconjunto de localizaciones $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, el vector $\mathbf{Y}(\mathcal{S}) = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n))^\top$, sigue una distribución normal multivariada con vector de medias $\boldsymbol{\mu} = (\mu(\mathbf{s}_1), \mu(\mathbf{s}_2), \ldots, \mu(\mathbf{s}_n))^\top$ y matriz de covarianzas $\boldsymbol{\Sigma}$ de orden $n \times n$, de modo que

$$\mathbf{Y}(\mathcal{S}) \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{2.2}$$

El modelo de regresión lineal espacial gaussiano es definido por:

$$Y(\mathbf{s}_i) = \mu(\mathbf{s}_i) + Z(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad i = 1, \ldots, n, \tag{2.3}$$

en que la media del proceso estocástico puede ser definida a través de funciones conocidas, que pueden estar indexadas en $\mathbf{s}_i \in \mathcal{S}$ y $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ son coeficientes de regresión desconocidos a ser estimados. Por otro lado, $Z(\mathbf{s}_i)$ y $\epsilon(\mathbf{s}_i)$ representan la parte estocástica del modelo, que pueden estar indexados en $\mathbf{s}_i$, y que consideraremos normalmente distribuidos e independientes, tal que $Z(\mathbf{s}_i) \sim N(0, \sigma_0^2)$ y $\epsilon(\mathbf{s}_i) \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$, en que σ_0^2 y σ_ϵ^2 representan la varianza del proceso espacial y del error, respectivamente.

Estos términos serán abordados nuevamente en la Sección 2.1.1.2. Usaremos $\overset{iid}{\sim}$ para denotar que una secuencia de variables es independiente e idénticamente distribuida.

La representación matricial del modelo espacial lineal definido en (2.3), es dada por

$$\mathbf{Y}(\mathcal{S}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}(\mathcal{S}) + \boldsymbol{\epsilon}(\mathcal{S}),$$

en que $\mathbf{X}$ es la matriz de covariables de dimensiones $n \times p$, $\boldsymbol{\beta}$ es el vector de coeficientes de regresión de dimensiones $p \times 1$, y $\mathbf{Z}(\mathcal{S})$ y $\boldsymbol{\epsilon}(\mathcal{S})$ son procesos gaussianos n-dimensionales, tal que $\mathbf{Z}(\mathcal{S}) \sim N_n(\mathbf{0}, \sigma_0^2\mathbf{R})$ y $\boldsymbol{\epsilon}(\mathcal{S}) \sim N_n(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n)$. Aquí, $\mathbf{R}$ es la matriz de correlación del proceso espacial de dimensiones $n \times n$, mientras que $\mathbf{I}_n$ denota una matriz identidad de dimensiones $n \times n$. Por simplicidad, nos referiremos a las variables aleatorias n-variadas $\mathbf{Y}(\mathcal{S})$, $\mathbf{Z}(\mathcal{S})$ y $\boldsymbol{\epsilon}(\mathcal{S})$ como $\mathbf{Y}$, $\mathbf{Z}$ y $\boldsymbol{\epsilon}$, respectivamente. Consecuentemente,

$$\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma_0^2\mathbf{R} + \sigma_\epsilon^2\mathbf{I}_n).$$

La construcción de la matriz $\mathbf{R}$ y conceptos adicionales que serán necesarios para su adecuada comprensión son mostrados en la siguiente Sección.

2.1.1 Función de Covarianza Espacial

La función de covarianza es un componente fundamental para modelar la dependencia entre observaciones en diferentes localizaciones. Esta función define la relación entre puntos espaciales en términos de proximidad y similitud, permitiendo capturar cómo las variaciones en el espacio están correlacionadas (Hristopulos, 2020). A continuación, presentamos algunas definiciones importantes sobre la función de covarianza.

Definición 1 (Función de Covarianza)**.** *La covarianza $\sigma_{ij} = \mathcal{C}(s_i, s_j)$ entre dos variables aleatorias espaciales está dada por:*

$$\mathcal{C}(s_i, s_j) = \mathbb{E}\left[(Y(s_i) - \mu(s_i))(Y(s_j) - \mu(s_j))\right] \tag{2.4}$$

con las siguientes propiedades

- $\mathcal{C}(s_i, s_j) = \mathcal{C}(s_j, s_i)$
- $\mathcal{C}(s_i, s_i) = Var(Y(s_i)) = \sigma_i^2$
- *Para una secuencia $\{Y(s_i)\}_{i=1}^n$ de variables aleatorias espaciales y una secuencia de números reales $\{a_i\}_{i=1}^n$, se cumple que $\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathcal{C}(s_i, s_j) \geq 0$.*

Por lo tanto, la representación matricial de la covarianza está dada por:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_n^2 \end{bmatrix}.$$

donde $\boldsymbol{\Sigma}$ *es una matriz cuadrada, simétrica, positiva definida y no singular.*

Definición 2 (Función de correlación)**.** *La correlación, entre dos variables aleatorias espaciales está definida por*

$$Corr(s_i, s_j) = \frac{\mathcal{C}(s_i, s_j)}{\sqrt{Var[Y(s_i)]Var[Y(s_j)]}} \tag{2.5}$$

De tal manera que

- $Corr(s_i, s_j) \in [-1, 1]$
- $Corr(s_i, s_j) = 1$ *, para todo* $i = j$

La matriz de correlación $\boldsymbol{R}$*, es expresada como*

$$\boldsymbol{R} = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\sigma_2} & \dots & \frac{\sigma_{1n}}{\sigma_1\sigma_n} \\ \frac{\sigma_{21}}{\sigma_2\sigma_1} & 1 & \dots & \frac{\sigma_{2n}}{\sigma_2\sigma_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_{n1}}{\sigma_n\sigma_1} & \frac{\sigma_{n2}}{\sigma_2\sigma_2} & \dots & 1 \end{bmatrix}$$

Adicionalmente, si $\mathcal{C}(s_i, s_i) = \sigma^2 > 0$*, para todo* $i = 1, 2, \dots, n$*. entonces* $\mathcal{C}(s_i, s_j) = \sigma^2 Corr(s_i, s_j)$*.*

Aunque el estudio de las funciones de covarianza abarca una amplia gama de conceptos y aplicaciones, como los mostrados en Cressie (2015), para nuestro objetivo resulta importante introducir las definiciones de función de covarianza estacionaria e isotrópica.

Definición 3 (Función de covarianza estacionaria)**.** *Dado un GRF, este se dice tener función de covarianza estacionaria si esta puede ser escrita como*

$$\mathcal{C}(s_i, s_j) = \mathcal{C}(\boldsymbol{h}) \tag{2.6}$$

con $\boldsymbol{h} = s_j - s_i \in \mathbb{R}^d$*, es decir, solo depende de la diferencia entre las localizaciones.*

Definición 4 (Función de covarianza isotrópica)**.** *Dado un GRF, este se dice tener función de covarianza isotrópica si*

$$\mathcal{C}(s_i, s_j) = \mathcal{C}(\|\boldsymbol{h}\|) \tag{2.7}$$

donde $\|\cdot\|$ *representa la distancia euclidiana, es decir, la covarianza espacial solo depende de la distancia entre las localizaciones.*

2.1.1.1 Función de Covarianza Matérn

La elección de una función de covarianza adecuada es crucial, ya que afecta directamente la precisión de la estimación, la predicción y la interpretación del modelo. Es natural pensar que para una función de covarianza estacionaria la correlación entre observaciones de dos localidades, s_i y s_j, disminuye a medida que la distancia euclidiana entre estas localizaciones $\|\boldsymbol{h}\| = \|s_j - s_i\|$ aumenta.

En la literatura, se han desarrollado diversos modelos ampliamente utilizados para describir la función de correlación espacial. Entre ellos, destacan los modelos exponenciales, gaussianos, esféricos, power-exponentiales y la familia de funciones de Matérn. Esta última es de especial interés debido a su flexibilidad para modelar diferentes grados de suavidad en los procesos espaciales y su aplicación en una amplia variedad de situaciones (Guttorp & Gneiting, 2006; Wang *et al.*, 2023). Dado su uso extensivo y sus propiedades favorables, en este estudio se considerará específicamente la función de Matérn. Para un GRF con función de correlación espacial estacionaria e isotrópica de la clase Matérn, esta es definida por

$$Corr(\|\boldsymbol{h}\|) = \begin{cases} \dfrac{2^{1-\nu}}{\Gamma(\nu)} \left(\dfrac{\|\boldsymbol{h}\|}{\phi}\right)^{\nu} \mathcal{K}_\nu \left(\dfrac{\|\boldsymbol{h}\|}{\phi}\right), & \|\boldsymbol{h}\| > 0 \\ 1, & \|\boldsymbol{h}\| = 0 \end{cases} \tag{2.8}$$

en que $\mathcal{K}_\nu(\cdot)$ es la función modificada de Bessel de segundo tipo, de orden ν, $\nu > 0$ es un parámetro de forma que determina la suavidad del proceso y $\phi > 0$ es el parámetro de escala también conocido como el "range parameter" que determina la distancia máxima a la cual las observaciones espaciales están correlacionadas.

2.1.1.2 Efecto Nugget

Por otro lado, el efecto nugget se refiere a la variabilidad que ocurre a escalas más pequeñas que la distancia mínima entre observaciones, comúnmente atribuida a errores de medición o a variaciones microescalares (Schabenberger & Gotway, 2017).

Sea $\sigma^2 > 0$ la varianza total del proceso de tal forma que $\sigma^2 = \sigma_0^2 + \sigma_\epsilon^2$, con $\sigma_0^2, \sigma_\epsilon^2 > 0$ donde σ_0^2 corresponde a la variabilidad espacial del proceso y σ_ϵ^2

es el efecto nugget, es fácil ver que la variabilidad total puede descomponerse de la siguiente forma

$$\sigma^2 = \gamma\sigma^2 + (1-\gamma)\,\sigma^2 \tag{2.9}$$

donde $\gamma = \left(\dfrac{\sigma_0^2}{\sigma_0^2+\sigma_\epsilon^2}\right) \in [0,1]$ representa la porción de la variabilidad atribuida al proceso espacial, y $1-\gamma$ la proporción de la variabilidad captada por el efecto nugget.

Luego, la matriz de varianzas y covarianzas del proceso estocástico definido en (2.3) es dada por

$$\boldsymbol{\Sigma} = \sigma_0^2\mathbf{R} + \sigma_\epsilon^2\mathbf{I}_n = \sigma^2\left[\gamma\mathbf{R} + (1-\gamma)\mathbf{I}_n\right], \tag{2.10}$$

en que cada entrada de la matriz $\mathbf{R}$ es calculada a través de la función de correlación de Matérn definida en (2.8).

La reparametrización de la matriz de varianzas y covarianzas descrita en (2.10) será de utilidad en el proceso de estimación de los parámetros del modelo ya que los parámetros σ^2 y γ pueden ser estimados consistentemente, como se menciona en Zhang (2004), lo cual no sucede al estimar σ_0^2 y σ_ϵ^2, vea Ordoñez *et al.* (2018) y Valeriano *et al.* (2021).

2.1.2 Función de verosimilitud

Definiendo $\boldsymbol{\Sigma} = \sigma^2[\gamma\mathbf{R} + (1-\gamma)\mathbf{I}_n]$, el vector de parámetros a estimar es $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \gamma, \sigma^2)$ y considerando que $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))^\top$ es una realización del proceso estocástico definido en (2.3), el logaritmo de la función de verosimilitud del modelo espacial lineal gaussiano es dado por:

$$\ell(\boldsymbol{\theta}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right). \tag{2.11}$$

Por lo tanto, para obtener las estimaciones de máxima verosimilitud, es necesario maximizar la función dada en (2.11) respecto al vector de parámetros $\boldsymbol{\theta}$. Como este modelo no es el principal interés de este trabajo, abordaremos nuevamente la función de verosimilitud en el Capítulo 3.

A continuación, se resume brevemente dos algoritmos ampliamente utilizados en la estimación de parámetros vía máxima verosimilitud ante la presencia de datos incompletos o ausentes.

2.2 Algoritmo EM y SAEM

El algoritmo Expectation-Maximization (EM), introducido por Dempster *et al.* (1977), es un método iterativo usualmente utilizado para la estimación de pará-

metros mediante máxima verosimilitud en modelos con presencia de datos censurados o faltantes. Este algoritmo utiliza el conjunto de datos completos $\mathbf{y} = (\mathbf{y}_o, \mathbf{y}_c)$, donde $\mathbf{y}_o$ representa los datos observados y $\mathbf{y}_c$ corresponde a los datos incompletos, y maximiza iterativamente la función de verosimilitud de los datos completos, $\ell_c(\boldsymbol{\theta}|\mathbf{y})$, hasta alcanzar la convergencia en un punto estacionario de la función de verosimilitud de los datos observados, $\ell(\boldsymbol{\theta}|\mathbf{y}_o)$. Este proceso se lleva a cabo mediante los siguientes dos pasos:

- **Paso E (Esperanza):** Dada la estimación de $\boldsymbol{\theta}$ en la k-ésima iteración $\widehat{\boldsymbol{\theta}}^{(k)}$, calculamos la esperanza de la verosimilitud de los datos completos condicionada a los datos observados y a la estimación actual de los parámetros, es decir,

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) = \mathbb{E}\left[\ell_c(\boldsymbol{\theta}|\mathbf{y}) \,\middle|\, \widehat{\boldsymbol{\theta}}^{(k)}, \mathbf{y}_o\right].$$

- **Paso M (Maximización):** Maximizamos la función $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ con respecto a $\boldsymbol{\theta}$, obteniendo así la nueva estimación del parámetro. La actualización se expresa como $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

El *Paso E* del algoritmo EM puede resultar inviable en algunos casos, particularmente cuando no es posible obtener una expresión analítica para la esperanza condicional calculada en este paso. Una alternativa, aunque poco eficiente, es utilizar el método de integración de Monte Carlo para aproximar la esperanza condicional, lo cual deriva en el algoritmo conocido como Monte Carlo EM (MCEM) y propuesto por Wei & Tanner (1990). Debido al costo computacional, de simular grandes cantidades de números aleatorios a partir de la distribución condicional, requerido por el algoritmo MCEM, Delyon *et al.* (1999) propuso el algoritmo Stochastic Approximation EM (SAEM), que introduce una aproximación estocástica durante el *Paso E*.

En cada iteración del algoritmo SAEM, se simulan observaciones de los datos incompletos a partir de su distribución condicional, lo que permite actualizar los parámetros a estimar. El *Paso E* en el algoritmo SAEM se expresa como:

- **Paso E (Esperanza):** Dado $\widehat{\boldsymbol{\theta}}^{(k)}$ la estimación de $\boldsymbol{\theta}$ en la k-ésima iteración

  1. Simular $\mathbf{q}^{(l,k)}$, para $l = 1, 2, \ldots, M$, a partir de la distribución condicional $f(\mathbf{q}|\widehat{\boldsymbol{\theta}}^{(k)}, \mathbf{y}_o)$. Luego, el nuevo vector de datos completos en la iteración (l, k) estará formado por $\mathbf{y}^{(l,k)} = (\mathbf{y}_o^\top, \mathbf{q}^{(l,k)\top})^\top$.
  2. Realizar la aproximación estocástica de $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ de la siguiente forma:

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \approx Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)}) + \lambda_k \left[\frac{1}{M}\sum_{l=1}^{M} \ell_c(\boldsymbol{\theta}; \mathbf{y}_o, \mathbf{q}^{(l,k)}) - Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k-1)})\right].$$

Donde λ_k es una secuencia de valores positivos que decrecen hacia 0, de tal manera que $\sum_{k=1}^{\infty} \lambda_k = \infty$ y $\sum_{k=1}^{\infty} \lambda_k^2 < \infty$.

- **Paso M (Maximización):** Maximizar $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$.

El algoritmo SAEM posee una propiedad de "memoria" controlada por el parámetro λ_k. Gracias a esta característica, solo es necesario realizar una pequeña cantidad de simulaciones M (se recomienda $M \leq 20$) para obtener resultados satisfactorios. Este aspecto diferencia notablemente a SAEM del algoritmo MCEM, el cual puede interpretarse como un caso particular donde $\lambda_k = 1$, lo que equivale a una "ausencia de memoria" en la aproximación estocástica.

La secuencia de valores λ_k puede definirse de la siguiente manera:

$$\lambda_k = \begin{cases} 1, & \text{si } 1 \leq k \leq \text{cW}, \\ \dfrac{1}{k - \text{cW}}, & \text{si } \text{cW} + 1 \leq k \leq W. \end{cases} \tag{2.12}$$

Donde M denota el número máximo de iteraciones y $c \in [0, 1]$ es un parámetro de corte que determina el porcentaje inicial de iteraciones "sin memoria". Para un análisis exhaustivo de las propiedades asociadas a esta elección de la secuencia de valores, se recomienda consultar Lavielle (2014), Kuhn & Lavielle (2005) y Delyon *et al.* (1999).

A continuación, se describe brevemente los datos geoespaciales que serán analizados en capítulos posteriores.

2.3 Datos: Incendios forestales

La provincia de Pichincha ha sido una de las más golpeadas por los incendios forestales en los últimos años, registrando un alarmante total de 2031 eventos de Nivel 1 entre 2022 y 2024. Estos incendios, no atribuibles a errores humanos, evidencian la urgente necesidad de implementar estrategias preventivas que salvaguarden vidas, protejan el ecosistema y optimicen el uso de los recursos disponibles.

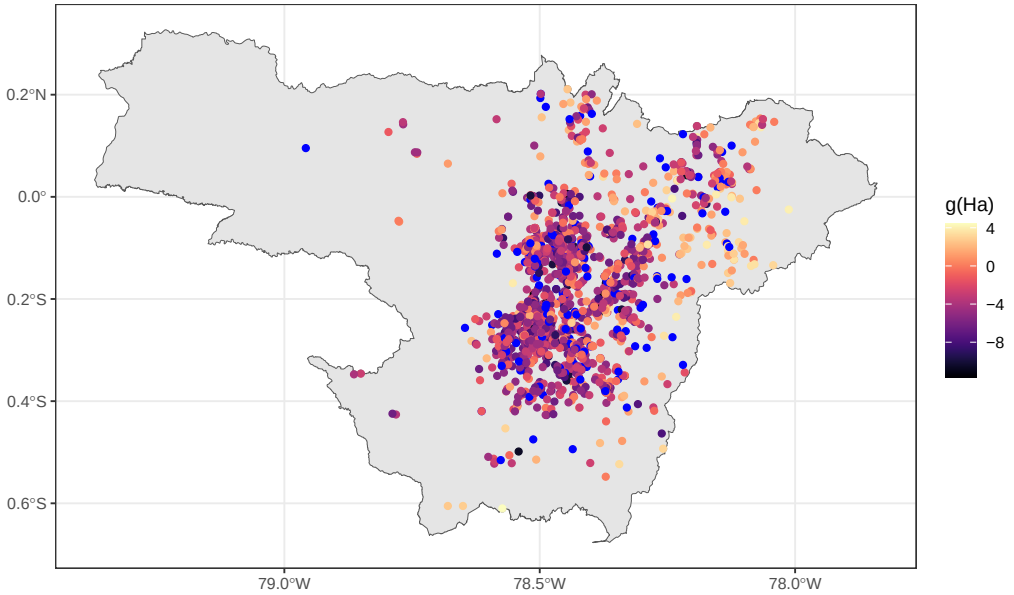
Para comprender mejor la magnitud de este problema, se dispone de un conjunto de datos proporcionado por la Secretaría Nacional de Riesgos del Ecuador, el cual documenta la extensión de cobertura vegetal afectada, medida en hectáreas (Ha). Cada incendio está georreferenciado con coordenadas precisas de latitud y longitud, permitiendo un estudio espacial detallado.

La Figura 1 presenta la distribución espacial de los niveles de afectación por incendios forestales en 2031 localidades de la provincia de Pichincha. Estos niveles corresponden a una transformación de la extensión de cobertura vegetal quemada, dada por la función:

$$g(\text{Ha}) = -10(\text{Ha}^{-0.1} - 1).$$

Se observa que las zonas centrales registran los menores niveles de afectación, mientras que la mayor concentración de impactos se encuentra en la frontera oeste.

Figura 1 – Distribución espacial de incendios forestales de nivel 1 en la provincia de Pichincha. Las ubicaciones con valores faltantes se representan a través de puntos azules.



Para evaluar el desempeño del método que propondremos en este trabajo (vea Capítulo 3), se seleccionaron aleatoriamente 10 % de las observaciones para ser tratadas como valores faltantes. Las localizaciones de estas observaciones están representadas en la Figura 1 mediante puntos en color azul.

3 Modelo espacial para respuestas censuradas usando SPDE

En este capítulo, se introduce el modelo de regresión espacial gaussiano con respuestas censuradas o faltantes y se desarrolla su formulación teórica. Se presenta la función de verosimilitud del modelo, destacando los desafíos computacionales asociados a la estimación de sus parámetros. Para mitigar estos costos, se emplea una aproximación de un campo aleatorio gaussiano (GRF) mediante un campo aleatorio gaussiano de Markov (GMRF), obtenido como solución de una ecuación diferencial parcial estocástica (SPDE). Esta estrategia permite una representación computacionalmente eficiente de procesos espaciales complejos.

Además, se derivan las expresiones para la estimación de los parámetros del modelo obtenidos a través del algoritmo SAEM, incorporando explícitamente la censura en los datos. También se examina la aproximación del error estándar asociado a estas estimaciones. Finalmente, se aborda la predicción en ubicaciones no observadas, integrando tanto la estructura espacial del proceso como la presencia de censura en los datos.

3.1 Modelo espacial para repuestas censuradas

El modelo espacial lineal gaussiano presentado en la Sección 2.1 asume que todos los valores correspondientes a una realización del proceso estocástico $\mathbf{Y}(\mathcal{S})$ son observados en las localidades $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Es decir, se dispone de todas las observaciones de $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))^\top$ de manera completa. No obstante, en muchas aplicaciones prácticas, algunas observaciones del proceso pueden estar censuradas o faltantes en determinadas localidades.

Sea $C_i \subseteq \mathbb{R}$ un intervalo, tal que la variable $Y(\mathbf{s}_i)$ no es observada si $y(\mathbf{s}_i) \in C_i$, para algunas localidades $\mathbf{s}_i \in \mathcal{S}$. En su lugar, observamos V_i y δ_i para cada localidad $\mathbf{s}_i$, tal que $\delta_i = 1$, si $y(\mathbf{s}_i) \in C_i$ y $\delta_i = 0$ en caso contrario, es decir, δ_i es el indicador de censura; y, V_i es dado por:

$$V_i = \begin{cases} c_i, & y(\mathbf{s}_i) \in C_i \text{ (censurado)} \\ y(\mathbf{s}_i), & y(\mathbf{s}_i) \notin C_i \text{ (observado)} \end{cases} \tag{3.1}$$

Por lo tanto, C_i describe la región de censura. La forma de C_i indicará el tipo de censura a considerar, que puede ser a la izquierda, a la derecha o intervalar, en cuyo

caso C_i asume la forma $(-\infty, c_i)$, (c_i, ∞) o (c_{i1}, c_{i2}), respectivamente. La constante $c_i \in \mathbb{R}$ es conocida como el límite de detección o límite de censura. Para el caso de valores faltantes (missing data) aleatorios, también denominados "missing at random" (MAR), Valeriano *et al.* (2021) sugiere considerar $C_i = \mathbb{R}$.

El modelo definido por las ecuaciones (2.3) y (3.1) será denominado "*modelo de regresión lineal espacial para respuestas censuradas*", cuya función de verosimilitud es definida a continuación.

3.1.1 Función de verosimilitud

Sea $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \phi, \gamma, \sigma^2)^\top$ el vector de parámetros del modelo. Para determinar la función de verosimilitud de los datos observados y posteriormente los estimadores de máxima verosimilitud, es necesario tratar separadamente los valores observados y los censurados.

En el Capítulo anterior consideramos que si todos los valores del vector de respuestas $\mathbf{y}$ fueran completamente observados, la variable aleatoria $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, donde $\boldsymbol{\Sigma}$ es la matriz de varianzas y covarianzas que fue definida en (2.10). Si no todos los valores de $\mathbf{y}$ son completamente observados, para determinar la función de verosimilitud, debemos tratar por separado los componentes observados y censurados existentes en $\mathbf{y}$, es decir, $\mathbf{y} = (\mathbf{y}_o, \mathbf{y}_c)^\top$, en que $\mathbf{y}_o$ es el vector de valores completamente observados de dimensión n_o y $\mathbf{y}_c$ es el vector de valores censurados o faltantes de dimensión n_c, tal que $n = n_o + n_c$.

Para los datos censurados, defina $\mathbf{C} = C_1 \times \ldots \times C_{n_c}$ como la región de censura. Además, defina $\delta_i = 0$ para todos los elementos en $\mathbf{y}_o$ y $\delta_i = 1$ para todos los elementos en $\mathbf{y}_c$. Luego de un re-ordenamiento de los elementos de $\mathbf{y}$, $\mathbf{X}$ y $\boldsymbol{\Sigma}$, se tiene:

$$\mathbf{y} = \text{vec}(\mathbf{y}_o, \mathbf{y}_m), \quad \boldsymbol{\delta} = \text{vec}(\boldsymbol{\delta}_o, \boldsymbol{\delta}_c), \quad \mathbf{X}^\top = (\mathbf{X}_o^\top, \mathbf{X}_c^\top), \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{oo} & \boldsymbol{\Sigma}_{oc} \\ \boldsymbol{\Sigma}_{co} & \boldsymbol{\Sigma}_{cc} \end{bmatrix},$$

donde $vec(\cdot)$ es el operador que apila columnas dentro de un mismo vector.

Utilizando las propiedades de la distribución normal multivariada, obtenemos que $\mathbf{Y}_o \sim N_{n_o}(\mathbf{X}_o\boldsymbol{\beta}, \boldsymbol{\Sigma}_{oo})$, $\mathbf{Y}_c|\mathbf{y}_o \sim N_{n_c}(\boldsymbol{\mu}^*, \boldsymbol{S}^*)$, tal que $\boldsymbol{\mu}^* = \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\Sigma}_{co}\boldsymbol{\Sigma}_{oo}^{-1}(\mathbf{y}_o - \mathbf{X}_o\boldsymbol{\beta})$ y $\mathbf{S}^* = \boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{co}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{oc}$. Sea $\phi_n(\mathbf{y}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ la función de densidad normal multivariada con media $\boldsymbol{\mu}$ y matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$, la función de verosimilitud de los datos observados está dada por:

$$L(\boldsymbol{\theta}; \mathbf{y}) = \int_{\mathbf{C}} \phi_n(\mathbf{y}; \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}) d\mathbf{y}_c = \phi_{n_o}(\mathbf{y}_o; \mathbf{X}_o\boldsymbol{\beta}, \boldsymbol{\Sigma}_{oo}) P(\mathbf{Y}_c \in \mathbf{C}|\mathbf{y}_o), \tag{3.2}$$

que corresponde al producto de la función de densidad marginal de los componentes completamente observados y la probabilidad condicional (en los valores observados) de los componentes censurados de estar en la región de censura $\mathbf{C}$.

Note que determinar las estimaciones de máxima verosimilitud a partir del logaritmo de la función mostrada en (3.2) puede resultar un desafío debido a la necesidad de realizar cálculos complejos o emplear métodos numéricos en el proceso. Sin embargo, este problema puede ser abordado mediante el uso de algún algoritmo del tipo Expectation-Maximization (EM), el cual fue descrito en el Capítulo 1 (Sección 2.2), de manera similar a lo que se muestra en Ordoñez *et al.* (2018) o Valeriano *et al.* (2021). Aunque esto resolvería el problema de estimación de los parámetros, aún existen dificultades no consideradas por los autores mencionados, como la ineficiencia computacional de los algoritmos propuestos ante la presencia de grandes volúmenes de datos. Por este motivo, a continuación se considerará la aproximación del modelo de regresión lineal espacial, como sugerido por Sahoo *et al.* (2024), y posteriormente se utilizará un algoritmo del tipo EM para la estimación de los parámetros.

3.2 Aproximación del modelo espacial gaussiano

Para mejorar la eficiencia computacional del algoritmo propuesto por Ordoñez *et al.* (2018), consideraremos, en el modelo definido en (2.3), la aproximación del GRF $Z(\cdot)$ mediante $\tilde{Z}(\cdot)$, construido a partir de un GMRF y operando en función de su matriz de precisión, como sugiere Sahoo *et al.* (2024). Esta aproximación presenta la ventaja de evitar el cálculo directo de la matriz de correlación $\mathbf{R}$ del proceso exacto, que es densa. El proceso aproximado se expresa como:

$$\tilde{Y}(\mathbf{s}_i) = \mu(\mathbf{s}_i) + \tilde{Z}(\mathbf{s}_i) + \epsilon(\mathbf{s}_i), \quad \mathbf{s}_i \in \mathcal{S}.$$

El proceso gaussiano con correlación Matérn $Z(\cdot)$, definido según (Lindgren *et al.*, 2011), puede ser obtenido como solución de la siguiente SPDE:

$$(\phi^{-2} - \boldsymbol{\Delta})^{\alpha/2} Z(\mathbf{s}) = 4\pi\phi^{-2}\mathcal{W}(\mathbf{s}), \quad \mathbf{s} \in \mathbb{R}^2,$$

en que $\mathcal{W}(\cdot)$ es un proceso gaussiano de ruido blanco y $\boldsymbol{\Delta} = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2}$ es el operador Laplaciano, además, $\alpha = \nu + 1$. En este trabajo consideramos $\nu = 1$, sugerido por Whittle (1954), como la covarianza elemental en $\mathbb{R}^2$. La solución $Z(\mathbf{s})$ de la SPDE puede ser aproximada a través del método de elementos finitos (Ciarlet, 1978; Brenner & Scott, 2008) mediante una discretización del espacio a través de una triangulación definida en una región limitada de $\mathbb{R}^2$. La triangulación puede ser construida usando la función `inla.mesh.2d` de la librería `INLA` (Rue *et al.*, 2009) disponible en CRAN

(R Core Team, 2023) o a través de la función `fm_mesh_2d_inla` del paquete `fmesher` (Lindgren & Shewchuk, 2024).

Sea $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_N^*\}$ el conjunto de nodos dentro de la triangulación. La solución aproximada de $Z(\mathbf{s})$ por el método de los elementos finitos es dada por

$$\tilde{Z}(\mathbf{s}) = \sum_{j=1}^{N} \zeta_j(\mathbf{s}) Z_j^*,$$

en que $\{\zeta_j(\cdot)\}$ son funciones base, lineales por partes y con soporte compacto, definidas de tal forma que preserve una estructura markoviana sobre la triangulación. Por otra parte $\{Z_j^*\}$ son ponderaciones normalmente distribuidas, definida para cada función base, así entonces, $\mathbf{Z}^* = (Z_1^*, \ldots, Z_N^*)^\top \sim N_N(\mathbf{0}, \mathbf{Q}_\phi^{-1})$, en que $\mathbf{Q}_\phi$ es una matriz de precisión de dimensiones $N \times N$, dada por

$$\mathbf{Q}_\phi = \frac{\phi^2}{4\pi}\left[\frac{1}{\phi^4}\mathbf{D} + \frac{2}{\phi^2}\mathbf{G}_1 + \mathbf{G}_2\right],$$

donde $\mathbf{D}, \mathbf{G}_1$ y $\mathbf{G}_2$ son matrices sparse de dimensiones $N \times N$ y N denota el número de nodos considerados en la triangulación. La matriz $\mathbf{D}$ es diagonal, con el elemento de la posición (j, j) dada por $D_{jj} = \langle \zeta_j(\cdot), 1\rangle$, en que $\langle f, g\rangle = \int f(\mathbf{s})g(\mathbf{s})d\mathbf{s}$ denota el producto interno. De manera similar, el elemento en la posición (i, j) de la matriz $\mathbf{G}_1$ es dado por $G_{1,ij} = \langle \boldsymbol{\nabla}\zeta_i(\cdot), \boldsymbol{\nabla}\zeta_j(\cdot)\rangle$ y $\mathbf{G}_2 = \mathbf{G}_1\mathbf{D}^{-1}\mathbf{G}_1$. Estas matrices pueden ser calculadas eficientemente a partir de la función `inla.mesh.fem` disponible en la librería `INLA`.

Luego, el vector aleatorio $\mathbf{Z}(\mathcal{S})$ definido sobre las localidades $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ puede ser aproximado a través de $\tilde{\mathbf{Z}}(\mathcal{S}) = (\tilde{Z}(\mathbf{s}_1), \ldots, \tilde{Z}(\mathbf{s}_n))^\top$ que es dado por

$$\tilde{\mathbf{Z}}(\mathcal{S}) = \sigma_0 \mathbf{A}\mathbf{Z}^* = \sigma\sqrt{\gamma}\mathbf{A}\mathbf{Z}^*,$$

en que $\mathbf{A}$ es la matriz de proyección de dimensiones $n \times N$, cuyo elemento en la posición (i, j) es $\zeta_j(\mathbf{s}_i)$ para cada localización $\mathbf{s}_i \in \mathcal{S}$. Esta matriz puede ser obtenida a través de la función `inla.spde.make.A` disponible también en `INLA`. Entonces, la matriz de varianzas y covarianzas de $\tilde{\mathbf{Z}}(\mathcal{S})$ es dada por $\boldsymbol{\Sigma}_{\tilde{\mathbf{Z}}} = \sigma_0^2 \mathbf{A}\mathbf{Q}_\phi^{-1}\mathbf{A}^\top$.

Luego, la representación jerárquica del vector $\tilde{\mathbf{Y}}(\mathcal{S}) = (\tilde{Y}(\mathbf{s}_1), \ldots, \tilde{Y}(\mathbf{s}_n))^\top$, considerando que $\tilde{\mathbf{Z}}(\mathcal{S}) = \sigma\sqrt{\gamma}\,\mathbf{Z}^*$, es dada por

$$\begin{aligned} \tilde{\mathbf{Y}}(\mathcal{S})|\tilde{\mathbf{Z}}(\mathcal{S}) &\sim N_n(\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\tilde{\mathbf{Z}}, \sigma^2(1-\gamma)\mathbf{I}_n), \\ \tilde{\mathbf{Z}}(\mathcal{S}) &\sim N_N(\mathbf{0}, \sigma^2\gamma\mathbf{Q}_\phi^{-1}). \end{aligned}$$

Si estuviéramos trabajando con un enfoque bayesiano, la representación jerárquica mostrada previamente sería ideal para resolver el problema de estimación

de parámetros. Sin embargo, en el contexto frecuentista y considerando que algunas observaciones son censuradas o faltantes, la estimación de los parámetros del modelo se llevará a cabo mediante la distribución marginal del proceso aproximado.

$$\tilde{\mathbf{Y}}(\mathcal{S}) \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2[\gamma \mathbf{A}\mathbf{Q}_\phi^{-1}\mathbf{A}^\top + (1-\gamma)\mathbf{I}_n]). \tag{3.3}$$

3.3 Estimación de los parámetros

Para estimar los parámetros del modelo de regresión lineal espacial con respuestas censuradas, utilizaremos un algoritmo del tipo EM. Sea $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma^2, \gamma, \phi)^\top$ el vector de parámetros. La función de verosimilitud de los datos completos del proceso aproximado (3.3), considerando que $\boldsymbol{\Psi} = \gamma \mathbf{A}\mathbf{Q}_\phi^{-1}\mathbf{A}^\top + (1-\gamma)\mathbf{I}_n$, es

$$\ell_c(\boldsymbol{\theta}|\tilde{\mathbf{y}}) \propto -\frac{1}{2}\left[n\ln(\sigma^2) + \ln|\boldsymbol{\Psi}|\right] - \frac{1}{2\sigma^2}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\Psi}^{-1}(\tilde{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}).$$

Dada la estimación actual del vector de parámetros $\widehat{\boldsymbol{\theta}}^{(k)}$, en el paso E del algoritmo EM, calculamos la esperanza condicional del logaritmo de la función de verosimilitud de los datos completos, como sigue

$$Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)}) \propto -\frac{1}{2}\left[n\ln(\sigma^2) + \ln|\boldsymbol{\Psi}|\right] - \frac{1}{2\sigma^2}\widehat{A}^{(k)},$$

tal que $\widehat{A}^{(k)} = \text{tr}(\widehat{\mathbf{y}}^{2(k)}\boldsymbol{\Psi}^{-1}) - 2\widehat{\mathbf{y}}^{(k)\top}\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top\mathbf{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{X}\boldsymbol{\beta}$. Note que este paso implica el cálculo de las esperanzas condicionales

$$\widehat{\mathbf{y}}^{2(k)} = \mathbb{E}[\tilde{\mathbf{Y}}\tilde{\mathbf{Y}}^\top|\mathbf{V}, \boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}^{(k)}] \quad \text{y} \quad \widehat{\mathbf{y}}^{(k)} = \mathbb{E}[\tilde{\mathbf{Y}}|\mathbf{V}, \boldsymbol{\delta}, \widehat{\boldsymbol{\theta}}^{(k)}]. \tag{3.4}$$

Aunque estas esperanzas pueden calcularse explícitamente mediante librerías de R como `tmvtnorm` o `MomTrunc`, este proceso puede ser computacionalmente costoso. Por ello, utilizaremos una variación del algoritmo EM, conocida como algoritmo SAEM, que sustituye el cálculo exacto de las esperanzas por una aproximación estocástica basada en la simulación de un pequeño número de valores provenientes de la distribución condicional, como se detalla a continuación.

Paso E:

1. **Paso E-1 (Simulación):** Simular $\mathbf{y}_c$ a partir de la distribución normal truncada con región de truncamiento $\mathbf{C} = C_1 \times \cdots \times C_{n_c}$, esto es, $\mathbf{Y}_c|\mathbf{V}, \boldsymbol{\delta} \sim TN_{n_c}(\boldsymbol{\mu}^*, \mathbf{S}^*; \mathbf{C})$, con media $\boldsymbol{\mu}^* = \mathbf{X}_c\boldsymbol{\beta} + \boldsymbol{\Sigma}_{co}\boldsymbol{\Sigma}_{oo}^{-1}(\mathbf{y}_o - \mathbf{X}_o\boldsymbol{\beta})$ y matriz de varianzas y covarianzas $\mathbf{S}^* = \boldsymbol{\Sigma}_{cc} - \boldsymbol{\Sigma}_{co}\boldsymbol{\Sigma}_{oo}^{-1}\boldsymbol{\Sigma}_{oc}$. De esta forma se obtiene una nueva observación $\mathbf{y}^{(k,l)} = \left(y_1, \ldots, y_{n_o}, y_{n_o+1}^{(k,l)}, y_{n_o+2}^{(k,l)}, \ldots, y_n^{(k,l)}\right)^\top$, que contiene los datos observados y una muestra de los n_c casos censurados (o faltantes).

2. **Paso E-2 (Aproximación):** Dada la secuencia $\mathbf{y}^{(k,l)}$, para $l = 1, 2, \ldots, M$, en la k-ésima iteración remplazamos las esperanzas condicionales dadas en (3.4) con las siguientes aproximaciones:

$$\begin{aligned}\widehat{\mathbf{y}}^{2(k)} &= \widehat{\mathbf{y}}^{2(k-1)} + \lambda_k \left[\frac{1}{M}\sum_{l=1}^{M}\mathbf{y}^{(k,l)}\mathbf{y}^{(k,l)\top} - \widehat{\mathbf{y}}^{2(k-1)}\right] \\ \widehat{\mathbf{y}}^{(k)} &= \widehat{\mathbf{y}}^{(k-1)} + \lambda_k \left[\frac{1}{M}\sum_{l=1}^{M}\mathbf{y}^{(k,l)} - \widehat{\mathbf{y}}^{(k-1)}\right].\end{aligned}$$

A partir de esto se realiza un paso de Maximización Condicional (CM), maximizando $Q(\boldsymbol{\theta}|\widehat{\boldsymbol{\theta}}^{(k)})$ respecto a $\boldsymbol{\theta}$, para obtener $\widehat{\boldsymbol{\theta}}^{(k+1)}$.

Paso M: Las estimaciones de $\boldsymbol{\theta}$ en la iteración $(k+1)$ están dadas por:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{(k+1)} &= (\mathbf{X}^\top\widehat{\boldsymbol{\Psi}}^{-1(k)}\mathbf{X})^{-1}\mathbf{X}^\top\widehat{\boldsymbol{\Psi}}^{-1(k)}\widehat{\mathbf{y}}^{(k)} \\ \widehat{\sigma}^{2(k+1)} &= \frac{1}{n}\left\{\mathrm{tr}(\widehat{\mathbf{y}}^{2(k)}\widehat{\boldsymbol{\Psi}}^{-1(k)}) - 2\widehat{\mathbf{y}}^{(k)\top}\widehat{\boldsymbol{\Psi}}^{-1(k)}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)} + \widehat{\boldsymbol{\beta}}^{(k+1)\top}\mathbf{X}^\top\widehat{\boldsymbol{\Psi}}^{-1(k)}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)}\right\} \\ \widehat{\boldsymbol{\alpha}}^{(k+1)} &= \underset{\boldsymbol{\alpha}\in\mathbb{R}^+\times[0,1]}{\mathrm{argmax}}\left\{-\frac{1}{2}\ln|\boldsymbol{\Psi}| - \frac{1}{2\widehat{\sigma}^{2(k+1)}}\left(\mathrm{tr}(\widehat{\mathbf{y}}^{2(k)}\boldsymbol{\Psi}^{-1}) - 2\widehat{\mathbf{y}}^{(k)\top}\boldsymbol{\Psi}^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)}\right.\right. \\ &\quad \left.\left. + \widehat{\boldsymbol{\beta}}^{(k+1)\top}\mathbf{X}^\top\boldsymbol{\Psi}^{-1}\mathbf{X}\widehat{\boldsymbol{\beta}}^{(k+1)}\right)\right\}\end{aligned}$$

con $\boldsymbol{\alpha} = (\phi, \gamma)^\top$. Luego, las estimaciones de la varianza del proceso espacial y la varianza del error, pueden ser obtenidas a través de $\widehat{\sigma}_o^2 = \widehat{\gamma}\widehat{\sigma}^2$ y $\hat{\sigma}_\epsilon^2 = (1-\widehat{\gamma})\widehat{\sigma}^2$, respectivamente. La estimaciones de $\boldsymbol{\alpha}$ pueden ser obtenidas a través de la función `roptim` de la librería con el mismo nombre. Los pasos E y M son ejecutados iterativamente hasta que la distancia entre dos evaluaciones sucesivas del logaritmo de la función de verosimilitud de los datos observados sea lo más pequeña posible, es decir, $|\ell(\widehat{\boldsymbol{\theta}}^{(k+1)}) - \ell(\widehat{\boldsymbol{\theta}}^{(k)})| < \epsilon$. Usualmente se considera $\epsilon = 10^{-5}$.

Finalmente, para mejorar la eficiencia en la estimación de los parámetros, se considerará la siguiente identidad para calcular la inversa de $\boldsymbol{\Psi}$:

$$\boldsymbol{\Psi}^{-1} = \frac{1}{1-\gamma}\mathbf{I}_n - \frac{\gamma}{1-\gamma}\mathbf{A}\left((1-\gamma)\mathbf{Q}_\phi + \gamma\mathbf{A}^\top\mathbf{A}\right)^{-1}\mathbf{A}^\top, \tag{3.5}$$

donde, en esta expresión, únicamente es necesario calcular la inversa de una matriz sparse de dimensiones $N \times N$. Asimismo, para el cálculo del determinante de $\boldsymbol{\Psi}$ se empleará la identidad:

$$|\boldsymbol{\Psi}| = \frac{(1-\gamma)^{n-N}}{|\mathbf{Q}_\phi|}\left|(1-\gamma)\mathbf{Q}_\phi + \gamma\mathbf{A}^\top\mathbf{A}\right|, \tag{3.6}$$

que también involucra el cálculo del determinante de una matriz sparse.

Además, se logra una reducción adicional en el costo computacional al evitar el cálculo repetido de la matriz de correlación espacial $\mathbf{R}$ en cada iteración del algoritmo, ya que esta se expresa en función de las matrices $\mathbf{D}$, $\mathbf{G}_1$, $\mathbf{G}_2$ y $\mathbf{A}$, las cuales se calculan una única vez a lo largo de todo el proceso.

3.4 Aproximación del error estándar

Además de determinar las estimaciones de los parámetros del modelo, siempre es de interés aproximar la varianza de estas estimaciones. Para ello, consideraremos la matriz de información de Fisher, la cual constituye una medida adecuada de la cantidad de información que se obtiene de un conjunto de datos. Con ella, es posible calcular, además de una cota inferior para la varianza de los estimadores, una aproximación de la varianza asintótica. En escenarios donde se emplea el algoritmo EM para obtener las estimaciones de máxima verosimilitud, Louis (1982) desarrolló un procedimiento que permite calcular la matriz de información de Fisher en problemas con datos incompletos.

Sea $\mathbf{S}_c(\mathbf{y};\boldsymbol{\theta}) = \dfrac{\partial \ell_c(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}$ el Jacobiano (vector de primeras derivadas) y $\mathbf{B}_c(\mathbf{y};\boldsymbol{\theta}) = -\dfrac{\partial^2 \ell_c(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}$ el negativo de la matriz Hessiana, correspondiente a las segundas derivadas del logaritmo de la función de verosimilitud de los datos completos respecto al vector de parámetros $\boldsymbol{\theta}$. De manera análoga, se define $\mathbf{S}_o(\mathbf{y}_o;\boldsymbol{\theta}) = \mathbb{E}[\mathbf{S}_c(\mathbf{y};\boldsymbol{\theta})|\mathbf{y}_o]$ y $\mathbf{B}_o(\mathbf{y}_o;\boldsymbol{\theta}) = \mathbb{E}[\mathbf{B}_c(\mathbf{y};\boldsymbol{\theta})|\mathbf{y}_o]$, el Jacobiano y el negativo de la matriz Hessiana de los datos observados, respectivamente. Louis (1982) demostró que la matriz de información de Fisher observada $\boldsymbol{\mathcal{I}}_o(\boldsymbol{\theta})$, obedece la siguiente relación:

$$\begin{aligned}\boldsymbol{\mathcal{I}}_o(\boldsymbol{\theta}) &= \mathbf{B}_o(\mathbf{y}_o;\boldsymbol{\theta}) \\ &= \mathbb{E}[\mathbf{B}_c(\mathbf{y};\boldsymbol{\theta})|\mathbf{y}_o] - \mathbb{E}[\mathbf{S}_c(\mathbf{y};\boldsymbol{\theta})\mathbf{S}_c^\top(\mathbf{y};\boldsymbol{\theta})|\mathbf{y}_o] + \mathbf{S}_o(\mathbf{y}_o;\boldsymbol{\theta})\mathbf{S}_o^\top(\mathbf{y}_o;\boldsymbol{\theta}). \quad (3.7)\end{aligned}$$

El primer término en (3.7) corresponde a la esperanza condicional en los datos observados de la matriz de información de los datos completos, mientras que, los términos restantes representan la matriz de información observada asociada con los datos censurados, faltantes o latentes.

Las esperanzas condicionales que conforman la matriz de información observada en (3.7), no siempre son fáciles de calcular y por lo tanto, Delyon *et al.* (1999) propuso aproximarla a cada iteración del algoritmo SAEM siguiendo la siguiente relación:

$$\begin{aligned}\Delta_k &= \Delta_{k-1} + \lambda_k \left[\frac{1}{M}\sum_{l=1}^{M} \frac{\partial \ell_c(\theta|\mathbf{y}^{(l,k)})}{\partial \boldsymbol{\theta}} - \Delta_{k-1}\right], \\ G_k &= G_{k-1} + \lambda_k \left[\frac{1}{M}\sum_{l=1}^{M} \frac{\partial \ell_c^2(\boldsymbol{\theta}|\mathbf{y}^{(l,k)})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} + \left(\frac{\partial \ell_c(\boldsymbol{\theta}|\mathbf{y}^{(l,k)})}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial \ell_c(\boldsymbol{\theta}|\mathbf{y}^{(l,k)})}{\partial \boldsymbol{\theta}}\right)^\top \right. \\ &\quad \left. - G_{k-1}\right] \\ H_k &= -G_k + \Delta_k \Delta_k^\top.\end{aligned}$$

Delyon *et al.* (1999) también puntuó que la inversa de la matriz de información observada $\boldsymbol{\mathcal{I}}_o(\boldsymbol{\theta}) = H_k$ converge a la matriz de varianzas y covarianzas asintótica del estimador de máxima verosimilitud. Las expresiones necesarias para calcular el Jacobiano y la matriz Hessiana son proporcionadas en el Anexo A.

3.5 Predicción

Además de la interpretabilidad, la cuantificación de la incertidumbre y la exploración integral que proporciona la modelización estadística espacial, la predicción del fenómeno de estudio en localizaciones donde la respuesta no fue observada, constituye uno de sus principales intereses. En este contexto, la predicción resulta fundamental para sustentar decisiones, implementar acciones preventivas, implementar estrategias de gestión y, en síntesis, tomar decisiones informadas.

Considerando a $\mathbf{Y}_{obs}$ como el vector aleatorio en las localidades observadas, definido como en la Sección 2.1, el interés radica en predecir $\mathbf{Y}_{pred}$ utilizando la información contenida en $\mathbf{Y}_{obs}$. Sin embargo, si $\mathbf{Y}_{obs}$ contiene observaciones censuradas o faltantes, el primer paso consiste en imputar las componentes censuradas mediante su esperanza condicional. Es decir, definimos $\mathbf{Y}^*_{obs} = \widehat{\mathbf{y}}^{(W)}$, donde $\widehat{\mathbf{y}}^{(W)}$ representa la esperanza condicional de la variable de respuesta, dada la información de los componentes observados. Esta cantidad se aproxima utilizando los valores obtenidos en la última iteración del algoritmo SAEM.

Expresando a $\mathbf{Y}_{pred} = (Y^{(1)}_{pred}, Y^{(2)}_{pred}, \ldots, Y^{(n_{pred})}_{pred})$, siendo un GRF, en las localizaciones $\mathcal{S}^+ = \{\mathbf{s}^+_1, \mathbf{s}^+_2, \ldots, \mathbf{s}^+_{n_{pred}}\}$. El vector aleatorio $\mathbf{Y}^\star = vec(\mathbf{Y}^*_{obs}, \mathbf{Y}_{pred})$, es también un GRF, de esta forma, con $\mathbf{X}^\star = (\mathbf{X}^\top_{obs}, \mathbf{X}^\top_{pred})^\top$ la matriz de diseño de orden $(n_{obs} + n_{pred}) \times p$, entonces $\mathbf{Y}^\star \sim N_{n_{obs}+n_{pred}}(\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star)$, con vector de medias $\boldsymbol{\mu}^\star = \mathbf{X}^\star\boldsymbol{\beta}$ y matriz de varianzas y covarianzas

$$\boldsymbol{\Sigma}^\star = \begin{bmatrix} \boldsymbol{\Sigma}_{obs,obs} & \boldsymbol{\Sigma}_{obs,pred} \\ \boldsymbol{\Sigma}_{pred,obs} & \boldsymbol{\Sigma}_{pred,pred} \end{bmatrix}. \tag{3.8}$$

Luego, $\mathbf{Y}_{pred}|\mathbf{y}^*_{obs} \sim N_{n_{pred}}(\boldsymbol{\mu}_p, \boldsymbol{\Sigma}_p)$, similar que en la Sección 3.1.1, usando propiedades de distribución multivariada,

$$\boldsymbol{\mu}_p = \mathbf{X}_{pred}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{pred,obs}\boldsymbol{\Sigma}^{-1}_{obs,obs}(\mathbf{y}^*_{obs} - \mathbf{X}_{obs}\boldsymbol{\beta}), \tag{3.9}$$

$$\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_{pred,pred} - \boldsymbol{\Sigma}_{pred,obs}\boldsymbol{\Sigma}^{-1}_{obs,obs}\boldsymbol{\Sigma}_{obs,pred}, \tag{3.10}$$

De este modo, el mejor predictor lineal de $\mathbf{Y}_{pred}$, que minimiza el error cuadrático medio (Mean Squared Error, MSE), está dado por la esperanza condicional $\mathbb{E}[\mathbf{Y}_{pred}|\mathbf{y}^*_{obs}] = \boldsymbol{\mu}_p$. Para una revisión más detallada, consulte De Oliveira (2005).

Adicionalmente, la matriz $\boldsymbol{\Sigma}^*$ definida en (3.8) es estimada como

$$\widehat{\boldsymbol{\Sigma}}^{\star} = \widehat{\sigma}^2 \left[\widehat{\gamma} \mathbf{A}^{\star} \mathbf{Q}_{\widehat{\phi}}^{-1} (\mathbf{A}^{\star})^{\top} + (1 - \widehat{\gamma}) \mathbf{I}_{(n_{obs}+n_{pred})} \right], \tag{3.11}$$

en que $\widehat{\sigma}^2$, $\widehat{\gamma}$ y $\widehat{\phi}$ son las estimaciones obtenidas en la última iteración del algoritmo SAEM. Por otro lado, $\mathbf{A}^{\star} = (\mathbf{A}^{\top}, \mathbf{A}_{pred}^{\top})^{\top}$ representa la matriz de proyección, de dimensiones $N \times (n_{obs} + n_{pred})$, del proceso espacial definido en las localidades $\mathcal{S}^{\star} = \{\mathcal{S}, \mathcal{S}^{+}\}$. Aquí, $\mathbf{A}$ es la matriz de proyección usada en el proceso de estimación dentro del algoritmo SAEM y $\mathbf{A}_{pred}$ es la matriz de proyección para las localidades $\mathcal{S}^{+}$, de tal forma que el elemento en la posición (i, j) es $\zeta_j(\mathbf{s}_i^{+})$. La matriz $\mathbf{A}_{pred}$ puede ser obtenida a través de la función `inla.spde.make.A`, especificando las localidades de predicción en el argumento `loc`.

4 Resultados de estudios de simulación y aplicaciones

Este capítulo presenta los resultados de tres estudios de simulación. En el primer estudio, se determina el número óptimo de nodos necesarios para aproximar un GRF mediante un GMRF, el cual es solución de una SPDE. El segundo estudio analiza las propiedades asintóticas de las estimaciones obtenidas utilizando el algoritmo propuesto. Finalmente, el tercer estudio compara los tiempos de cómputo requeridos por el método desarrollado en este trabajo para estimar los parámetros de un modelo de regresión espacial con respuestas censuradas, frente a los tiempos obtenidos mediante las funciones de los paquetes `RcppCensSpatial` y `Censpatial`. Asimismo, se incluyen los resultados de la aplicación de los métodos propuestos en un conjunto de datos real descrito en el Capítulo 2.

4.1 Estudios de Simulación

Para los tres estudios de simulación, se generaron aleatoriamente 100 conjuntos de datos siguiendo el modelo $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma})$. En los estudios de simulación I y III, se utilizó un tamaño de muestra de $n = 300$, mientras que en el segundo estudio se evaluaron tamaños de muestra de $n = 100$, 300 y 800. Las posiciones espaciales (localidades) correspondientes a cada observación se generaron de manera aleatoria dentro de un cuadrado de dimensiones 20×20.

La matriz de diseño $\mathbf{X} = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2)$ se construyó asignando 1 a la primera columna, mientras que las columnas 2 y 3 se generaron a partir de una distribución normal estándar $N(0, 1)$. Los coeficientes de regresión empleados fueron $\boldsymbol{\beta} = (1, 3, -2)^\top$.

Por otro lado, la matriz de varianzas y covarianzas $\boldsymbol{\Sigma}$ se generó utilizando la función de correlación Matérn descrita en el Capítulo 2. Los parámetros utilizados fueron los siguientes: varianza total $\sigma^2 = 5$, proporción de la varianza explicada por el proceso espacial $\gamma = 0.70$, y parámetro de escala $\phi = 3$. Esto implica que la correlación espacial se reduce a menos de 0.05 para distancias mayores a 12 unidades.

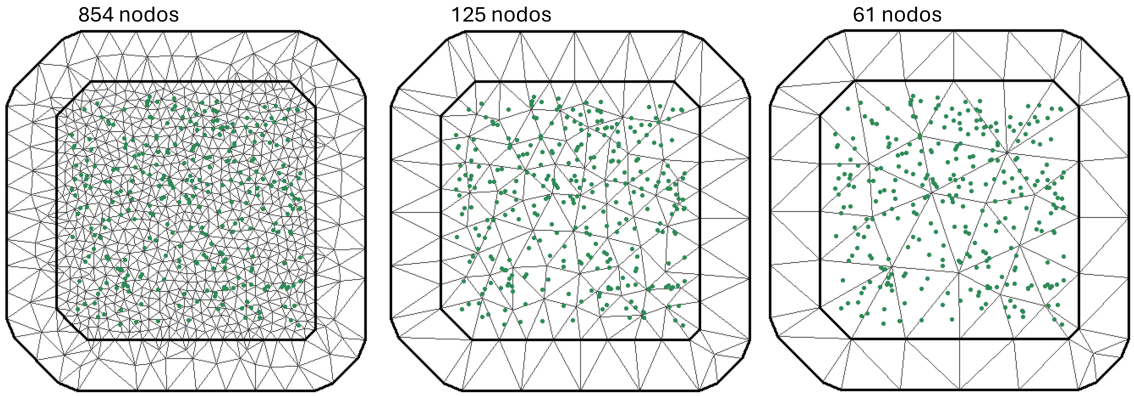
El proceso de censura de las observaciones se llevó a cabo siguiendo la metodología propuesta por Schelin & Sjöstedt-de Luna (2014). Para cada muestra simulada, se calculó el percentil correspondiente al nivel de censura deseado. En el caso de censura por la derecha, los valores que superaban dicho percentil se consideraron

censurados y fueron reemplazados por el límite de censura especificado. De manera similar, para la censura izquierda, se identificó el percentil correspondiente y los valores inferiores a este se sustituyeron por el límite de censura definido.

4.1.1 Simulación I: Selección del número de nodos

Como se describe en la Sección 3.2, una parte importante dentro del proceso de estimación de los parámetros consiste en la aproximación del proceso espacial por un GMRF. En este trabajo, para tal fin, utilizamos la función `inla.mesh.2d`, que realiza una triangulación de Delaunay restringida formada por puntos en la región de estudio, denominados nodos, donde cada nodo está conectado con dos nodos adyacentes mediante una recta, formando una arista. La selección del número de nodos está determinada por diversos factores, como los ángulos interiores y exteriores de cada triángulo, la longitud de las aristas, entre otros parámetros, según lo detallado en Lindgren & Rue (2015).

Figura 2 – Simulación I. Triangulación utilizada en el estudio de simulación considerando diferente número de nodos, los cuales fueron definidos al seleccionar el argumento `iedge` de la función creada para este proyecto igual a 0.05, 0.15 y 0.30, respectivamente.



Por lo tanto, definir el número de nodos es importante ya que estos afectan los tiempos computacionales; un mayor número de nodos implica un mayor número de cálculos. Como se menciona en Krainski *et al.* (2018), una buena aproximación se caracteriza por tener triángulos lo más regulares posible, lo cual puede controlarse estableciendo la longitud de las aristas mediante el argumento `max.edge` de la función `inla.mesh.2d`. En este sentido, proponemos controlar el número de nodos mediante un coeficiente denominado `iedge`, de manera que la longitud máxima de las aristas en cada triángulo interior, esté dada por una proporción (`iedge`) de la distancia máxima entre dos puntos de los datos observados. Figura 2 muestra la triangulación obtenida cuando se consideran 300 coordenadas (puntos verdes) e `iedge` asume los valores 0.05,

0.15 y 0.30, gráficos de izquierda a derecha, respectivamente. Note que entre menor es el valor de `iedge`, mayor es el número de nodos considerados en la triangulación.

Para este estudio de simulación, se generaron 100 muestras de Monte Carlo (MC) siguiendo el procedimiento descrito previamente, considerando un 10 % de observaciones censuradas a la derecha. La estimación de los parámetros se realizó mediante el algoritmo SAEM, presentado en el Capítulo 3, evaluando distintos valores de `iedge`. Esto dio lugar a triangulaciones con 854, 125, 69, 61 y 44 nodos.

En el Cuadro 1 se presentan los resultados obtenidos, donde se reportan la media (MC-AV) y la desviación estándar (MC-SD) de las 100 estimaciones obtenidas para cada parámetro del modelo. Además, se incluye la mediana del tiempo computacional requerido por nuestra metodología en función del número de nodos considerado. Cabe resaltar que las estimaciones de los coeficientes de regresión son satisfactorias en todas las configuraciones de `iedge`. Como era de esperarse, los parámetros relacionados con la estructura de covarianza del proceso (σ^2, ϕ, γ) muestran mayor sensibilidad a la definición de la triangulación. No obstante, se observa un desempeño óptimo, con un balance adecuado entre precisión y tiempo de cómputo, cuando `iedge` $= 0.15$.

Cuadro 1 – Simulación I. Media (MC-AV) y desviación estándar (MC-SD) de las estimaciones de los parámetros del modelo propuesto basado en 100 muestras MC de tamaño $n = 300$ con 10 % de censura a derecha y diferente número de nodos en la aproximación. También se reporta la mediana y desviación estándar del tiempo de ejecución (en minutos).

Nodos (`iedge`)	Medida	β_0 (1)	β_1 (3)	β_2 (-2)	σ^2 (5)	ϕ (3)	γ (0.7)	Tiempo (min)
854	MC-AV	0.967	3.002	-1.988	4.589	2.708	0.658	45.726
(0.05)	MC-SD	0.755	0.087	0.077	0.975	0.767	0.088	8.294
125	MC-AV	0.966	3.003	-1.993	5.076	2.853	0.645	2.389
(0.15)	MC-SD	0.758	0.083	0.076	1.011	0.885	0.087	0.406
69	MC-AV	0.974	3.001	-1.988	6.564	2.730	0.678	2.512
(0.25)	MC-SD	0.802	0.091	0.081	2.193	1.041	0.089	0.676
61	MC-AV	0.994	2.998	-1.984	8.610	2.662	0.719	2.460
(0.30)	MC-SD	0.805	0.090	0.085	4.372	1.371	0.097	0.743
44	MC-AV	0.998	2.998	-1.983	12.953	2.303	0.758	1.800
(0.40)	MC-SD	0.781	0.095	0.090	7.750	1.377	0.131	0.510

4.1.2 Simulación II: Propiedades Asintóticas

En esta sección, se analizan las propiedades asintóticas de las estimaciones obtenidas mediante el método propuesto. Para ello, se generaron 100 muestras de Monte

Carlo (MC) con tamaños de 100, 300 y 800, considerando dos escenarios distintos. En el primer escenario, el 8 % de las observaciones fueron censuradas a la izquierda, mientras que un 2 % adicional fue seleccionado aleatoriamente y tratado como valores faltantes. En el segundo escenario, la proporción de observaciones censuradas a la izquierda se incrementó al 24 %, y el 6 % del total de observaciones fue seleccionado aleatoriamente para ser tratado como valores faltantes.

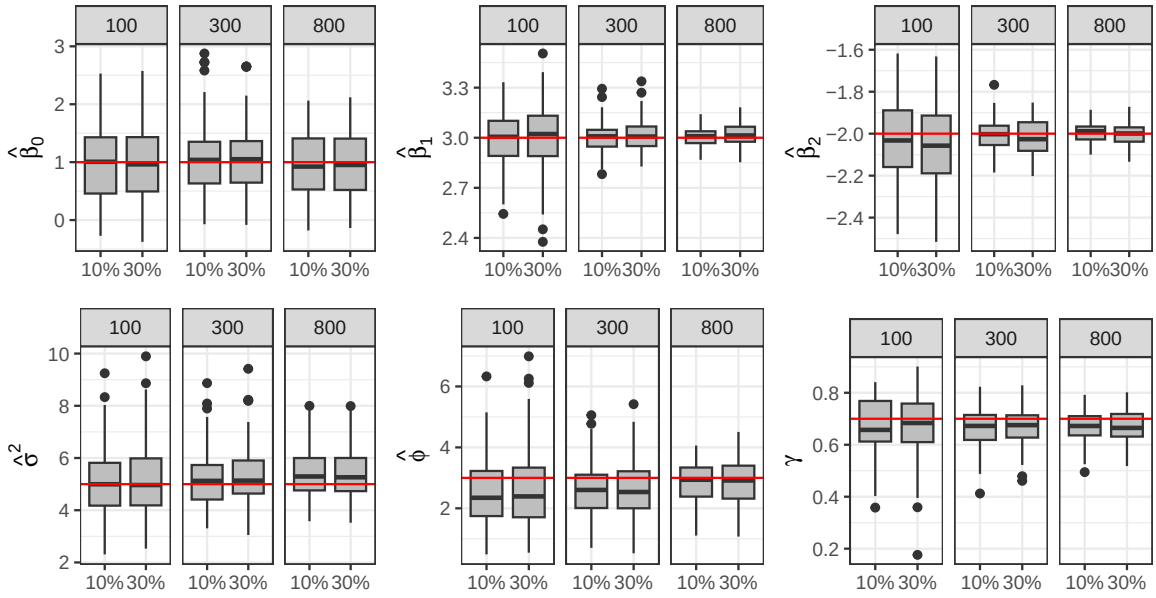
Los resultados obtenidos se presentan en el Cuadro 2, donde MC-AV y MC-SD corresponden a la media y la desviación estándar de las 100 estimaciones, respectivamente. Además, se reporta IM-AV, que representa la media de la aproximación del error estándar de las estimaciones, calculada a partir de la matriz de información observada. Los resultados indican que las estimaciones obtenidas se aproximan a los valores reales del modelo. Asimismo, los errores estándar estimados mediante la matriz de información observada muestran una concordancia notable con aquellos derivados de las muestras de Monte Carlo. Cabe destacar que los errores estándar disminuyen a medida que aumenta el tamaño de la muestra, mientras que tienden a incrementarse con niveles más altos de censura.

Cuadro 2 – Simulación II. Media (MC-AV) y desviación estándar (MC-SD) de las 100 estimaciones obtenidas del ajuste del modelo propuesto considerando muestras de tamaño $n = 100, 300$ y 800 con niveles de censura 10 % y 30 % a la izquierda y missings. IM-AV representa la media de la estimación del error estándar definido en Sección 3.4.

n	Censura	Métrica	β_0 (1)	β_1 (3)	β_2 (-2)	σ^2 (5)	ϕ (3)	γ (0.7)
100	10 %	MC-AV	1.013	2.998	-2.036	5.055	2.593	0.672
		MC-SD	0.662	0.163	0.185	1.270	1.176	0.106
		IM-AV	0.667	0.162	0.168	1.700	1.108	0.120
	30 %	MC-AV	0.992	3.013	-2.047	5.140	2.619	0.673
		MC-SD	0.667	0.212	0.193	1.380	1.289	0.122
		IM-AV	0.677	0.192	0.188	1.745	1.208	0.127
300	10 %	MC-AV	1.056	3.001	-2.006	5.201	2.641	0.663
		MC-SD	0.638	0.085	0.077	1.059	0.892	0.072
		IM-AV	0.645	0.094	0.084	1.124	0.832	0.075
	30 %	MC-AV	1.043	3.012	-2.017	5.281	2.639	0.666
		MC-SD	0.621	0.094	0.085	1.174	0.902	0.070
		IM-AV	0.655	0.111	0.096	1.265	0.869	0.081
800	10 %	MC-AV	0.969	3.006	-1.995	5.433	2.878	0.670
		MC-SD	0.569	0.053	0.045	0.989	0.666	0.063
		IM-AV	0.689	0.052	0.050	1.092	0.774	0.064
	30 %	MC-AV	0.956	3.017	-2.003	5.401	2.887	0.669
		MC-SD	0.559	0.064	0.055	1.001	0.741	0.065
		IM-AV	0.691	0.061	0.056	1.110	0.799	0.066

En la Figura 3 se presentan los boxplots de las 100 estimaciones para cada parámetro del modelo, considerando distintos tamaños de muestra y niveles de censura. La línea roja indica el valor verdadero del parámetro. Se observa que la mediana de las estimaciones de los coeficientes de regresión $(\beta_0, \beta_1, \beta_2)$ se encuentra muy próxima a los valores reales, independientemente del tamaño de la muestra y del nivel de censura. En el caso de ϕ, la mediana tiende a subestimar este parámetro en muestras de tamaño 100 y 300, sin depender del nivel de censura. Además, ϕ es generalmente subestimado en casi todos los escenarios analizados. Cabe destacar que el rango intercuartílico de las estimaciones para la mayoría de los parámetros disminuye a medida que aumenta el tamaño de la muestra, lo que indica una reducción en la variabilidad de las estimaciones. Sin embargo, este rango tiende a aumentar ligeramente con mayores niveles de censura, lo que concuerda con los resultados previamente observados.

Figura 3 – Simulación II. Boxplot de las estimaciones de los parámetros del modelo, considerando 100 muestras de tamaño $n = 100, 300$ y 800 con 10 % y 30 % de censura a la izquierda y valores faltantes. La línea roja representa el verdadero valor del parámetro.



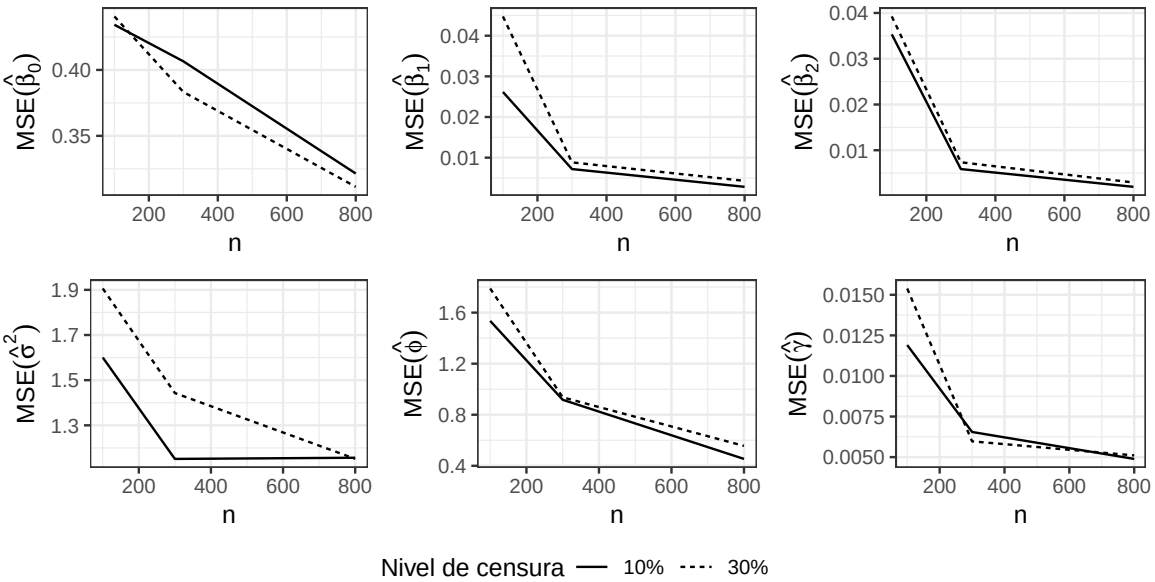
También se evalúa el error cuadrático medio (MSE), definido para un estimador $\widehat{\theta}$ como

$$MSE(\widehat{\theta}) = \frac{1}{100}\sum_{i=1}^{100}(\widehat{\theta}_i - \theta)^2,$$

en que $\widehat{\theta}_i$ es la estimación de θ obtenida en la i-ésima iteración, para $i = 1, \ldots, 100$. En la Figura 4 se muestra el MSE calculado para distintos tamaños de muestra, considerando dos niveles de censura: 10 % (línea sólida) y 30 % (línea punteada). Se observa que,

a medida que aumenta el tamaño de la muestra, el MSE disminuye, lo que refleja propiedades asintóticas deseables para las estimaciones.

Figura 4 – Simulación II. MSE de las estimaciones de los parámetros del modelo, considerando 100 muestras de tamaño $n = 100, 300$ y 800 con 10 % y 30 % de censura a la izquierda y valores faltantes.



4.1.3 Simulación III: Tiempos de ejecución

En esta sección, se comparan los tiempos computacionales del método propuesto en este trabajo, denominado `SAEM-SPDE`, con las alternativas disponibles en R, específicamente las implementadas en las librerías `RcppCensSpatial` y `CensSpatial`.

Cuadro 3 – Simulación III. Media de las estimaciones de los parámetros obtenidos a través de nuestra propuesta (`SAEM-SPDE`) y los paquetes `RcppCensSpatial` y `CensSpatial` basado en 100 muestras MC de tamaño $n = 300$ con 10 % de censura a derecha. También se reporta la mediana y la desviación estándar (en paréntesis) del tiempo de ejecución (en minutos).

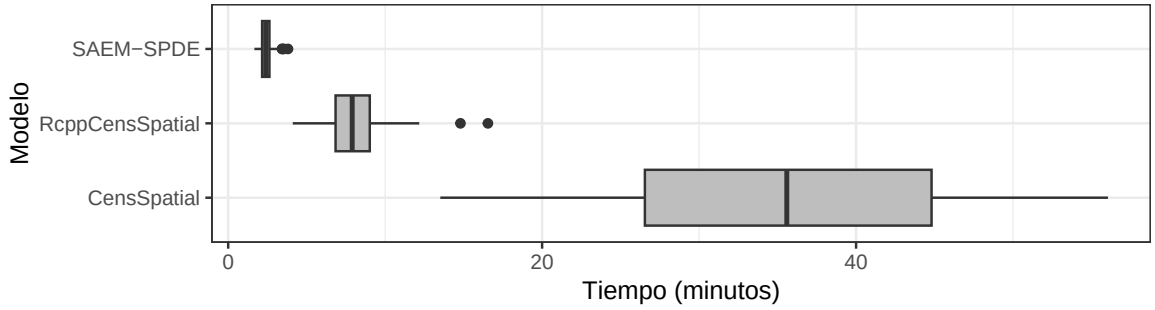
n	Método	β_0 (1)	β_1 (3)	β_2 (-2)	σ^2 (5)	ϕ (3)	γ (0.7)	Tiempo (desviación)
300	`SAEM-SPDE`	0.966	3.003	-1.993	5.076	2.853	0.645	2.389 (0.406)
	`RcppCensSpatial`	0.965	3.002	-1.988	4.674	2.830	0.665	7.900 (2.021)
	`CensSpatial`	0.965	3.002	-1.988	4.689	2.838	0.666	35.587 (10.562)

Para ello, se generaron 100 muestras de Monte Carlo con un tamaño de $n = 300$, considerando un 10 % de observaciones censuradas a la derecha. En el Cuadro 3, se presentan la media de las 100 estimaciones obtenidas por cada método, junto con la mediana de los tiempos de ejecución y sus respectivas desviaciones estándar.

Los resultados muestran que las estimaciones obtenidas son muy similares entre los distintos métodos. Sin embargo, los tiempos computacionales varían significativamente, destacándose que el método propuesto en este trabajo (`SAEM-SPDE`) es el que reporta los menores tiempos de ejecución, lo que resalta su eficiencia en términos de costos computacionales.

Adicionalmente, la Figura 5 presenta un boxplot de los tiempos de ejecución requeridos por cada uno de los métodos evaluados. Se observa que nuestra propuesta, `SAEM-SPDE`, es computacionalmente eficiente en comparación con las alternativas disponibles. En particular, el método implementado completamente en R (`CensSpatial`) mostró los tiempos de ejecución más elevados, evidenciando una mayor demanda computacional en relación con las demás opciones.

Figura 5 – Simulación III. Boxplot de los tiempos de ejecución (en minutos) del modelo propuesto (`SAEM-SPDE`) y las propuestas de los paquetes `RcppCensSpatial` y `CensSpatial` para estimar los parámetros del modelo espacial con respuestas censuras en muestras de tamaño 300 con 10 % de censura a la derecha.

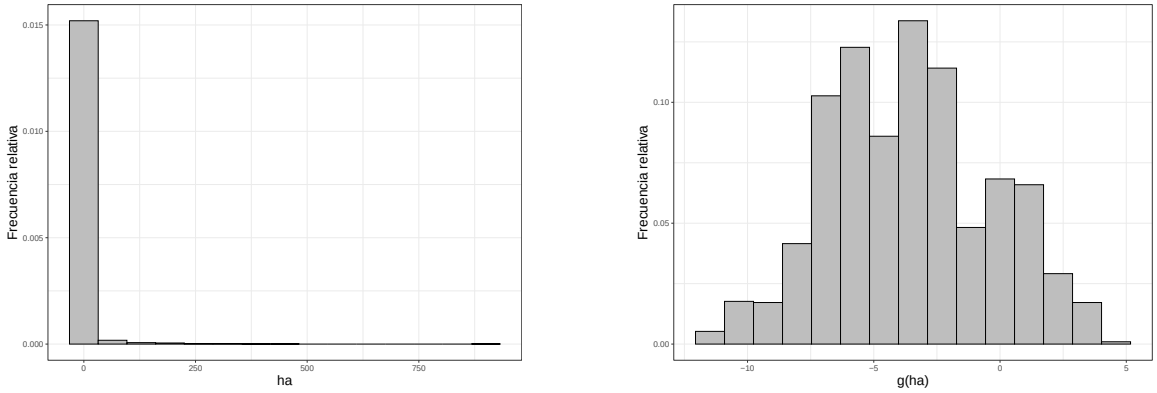


4.2 Aplicación: Incendios forestales

En este apartado se analiza el conjunto de datos sobre la cobertura vegetal afectada por incendios forestales y descrito en la Sección 2.3. Se seleccionaron los incendios de Nivel 1 debido a su alta incidencia en la provincia de Pichincha. Posteriormente, se aplicó la transformación de Box-Cox $g(\text{Ha}) = -10(\text{Ha}^{-0.1} - 1)$ para obtener una distribución simétrica que cumpla con el supuesto de normalidad, permitiendo así realizar una inferencia estadística válida sobre los datos transformados, como se ilustra en la Figura 6.

Dado que el modelo está especificado para operar con distancias euclidianas, se transformaron las coordenadas de longitud y latitud a la proyección cartesiana del sistema de coordenadas geocéntricas EPSG:4978, el cual nos permite interpretar las distancias en kilómetros. En la Figura 7 se muestra la triangulación de Delaunay, usada

Figura 6 – Histograma de los datos crudos (izquierda) e histograma de los datos transformados (derecha).

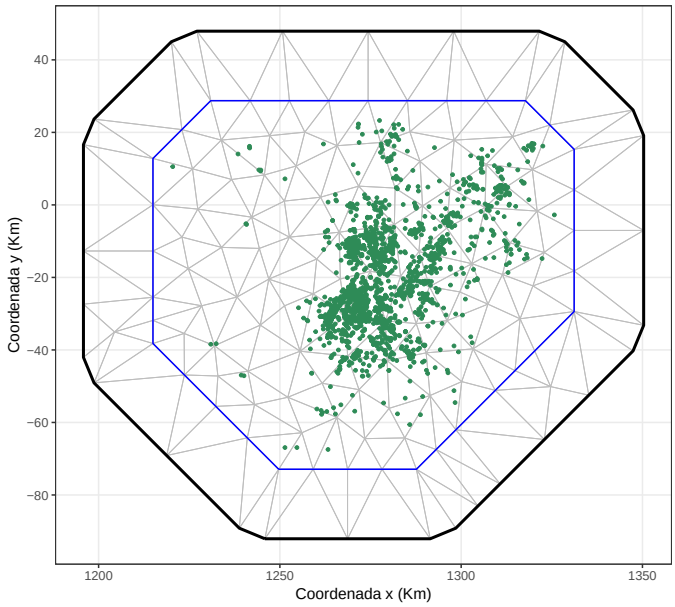


para la estimación de los parámetros del modelo especificado a continuación:

$$g(Ha_i) = \beta_0 + \beta_1 \text{Coord}_x + \beta_2 \text{Coord}_y + \tilde{Z}_i + \epsilon_i, \quad i = 1, 2, \ldots, 2031,$$

en que la variable de respuesta es dada por la transformación del total de hectáreas quemadas $y_i = g(Ha_i)$ y las covariables consideradas son las coordenadas cartesianas.

Figura 7 – Triangulación de Delaunay para la región de Pichincha con un total de 143 nodos, las localidades en estudio se muestran de color verde.



Para obtener las estimaciones del modelo, utilizamos la función de R desarrollada en este trabajo, denominada `SPDEsclm`. Esta función recibe los siguientes argumentos: `Y` es el vector de respuestas, `X` es la matriz de diseño, `ci` es el indicador de censura (valores faltantes), `lcl` y `ucl` corresponden a los límites inferior y superior de censura, y `coords` es la matriz de coordenadas cuyas unidades están en km, para

las 2031 localizaciones consideradas en la provincia de Pichincha. Adicionalmente, los valores iniciales para los parámetros $\boldsymbol{\beta}$ y σ^2 son calculadas dentro del modelo considerando únicamente los datos observados y los límites de censura, descartando las posiciones que poseen valores faltantes. A continuación, se presenta el `output` de la función con los resultados obtenidos.

Output SPDEsclm

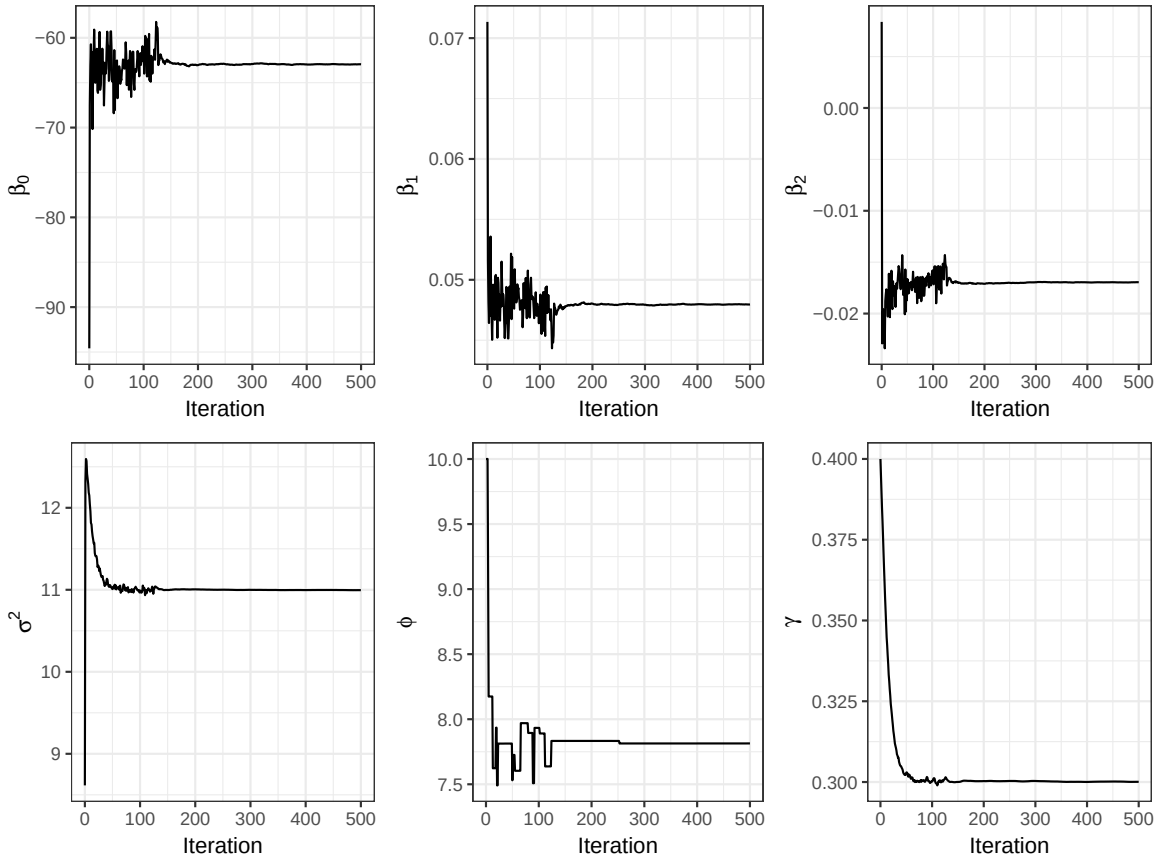
```
------------------------------------------------------------
      Censored Linear Spatial Regression Model using SPDE
------------------------------------------------------------
Call:
SPDEsclm(
        y = Y, x = X, ci = ci, lcl = LI, ucl = LS,
        coords = coords,  phi0 = 10, gamma0 = 0.4,
        lower = c(0.001, 0.001), upper = c(300, 0.999),
        iedge = 0.15, MaxIter = 500, M = 20, pc = 0.25,
        error = 1e-05
)

Estimated parameters:
        beta0  beta1   beta2  sigma2    phi  gamma
     -62.9477 0.0480 -0.0169 10.9955 7.8140 0.3001
s.e.  26.2878 0.0205  0.0198  1.0807 3.6093 0.0684


Model selection criteria:
         Loglik      AIC      BIC
Value -4492.453 8996.907 9030.604

Details:
Number of censored/missing values: 203
Convergence reached?: FALSE
Iterations: 500 / 500
```

Código 1– Output de la función `SPDEsclm`.

En el Código 1 se presenta la salida de la función `SPDEsclm`, que incluye las estimaciones de los parámetros del modelo espacial junto con sus errores estándar (SE). Además, se reportan los criterios de selección: log-verosimilitud (LogLik), criterio de información de Akaike (Akaike Information Criterion, AIC) y criterio de información Bayesiano (Bayesian Information Criterion, BIC).

La Figura 8 muestra la convergencia de los parámetros estimados en cada iteración de la función `SPDEsclm`. Se observa que, a partir de la iteración `200`, la estimación de los parámetros alcanza un estado de estabilidad, lo que indica una adecuada convergencia del método.

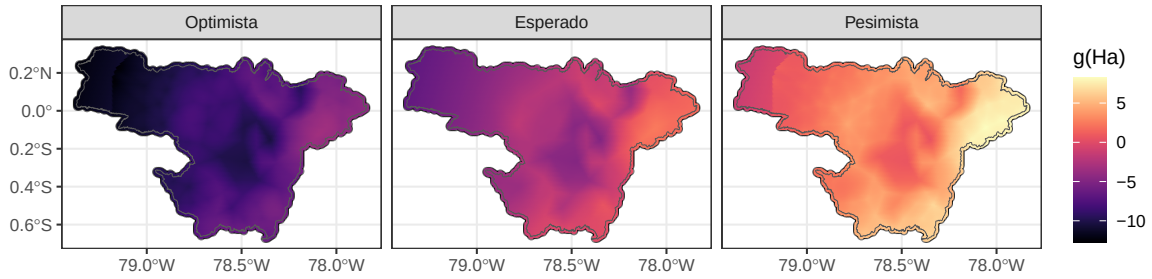
Figura 8 – Convergencia de los parámetros estimados a cada iteración de la función `SPDEsclm`.



La Figura 9 presenta los escenarios de predicción para la provincia de Pichincha, generados siguiendo la metodología descrita en la Sección 3.5. En esta figura se ilustran tres escenarios que reflejan la variabilidad en la predicción del proceso. Escenario optimista, los niveles de afectación por incendios forestales se mantienen en niveles de alerta principalmente en la frontera occidental, mientras que en la frontera oriental se observan niveles considerablemente más bajos. En el escenario pesimista, los niveles de afectación se intensifican en toda la provincia, reflejando un panorama

de mayor riesgo. Sin embargo, la zona central, que corresponde mayoritariamente al Distrito Metropolitano de Quito (DMQ), continúa siendo la región con menor impacto. En el escenario esperado se observa una distribución en la que los menores niveles de afectación se concentran alrededor del DMQ, con una intensificación progresiva conforme se avanza hacia la frontera oeste.

Figura 9 – Predicción de niveles de afectación por incendios forestales. El escenario optimista (izquierda) está representado por el límite inferior del intervalo de predicción, el escenario esperado (centro) es representado por los valores predichos, mientras que el escenario pesimista (derecha) esta representado por el límite superior del intervalo de predicción. Se utilizó un nivel de credibilidad del 95 %.



De los parámetros estimados podemos inferir que las regiones con mayores niveles de afectación se concentran mayoritariamente en la frontera oeste de la provincia de Pichincha, donde hay mayor densidad arbórea dada su conexión con la Amazonía y gran extensión de parques forestales. Además, El valor estimado del parámetro de escala ϕ nos indica que a distancias mayores de 31.25 km, los niveles de afectación tienen una correlación menor a 0.05. Por otra parte de los parámetros asociados a la varianza del proceso (σ^2, γ), se puede identificar que los niveles de afectación presentan una variabilidad total de 10.996 unidades cuadradas. De esta, un 30 % se atribuye a la distribución natural y local de los niveles de afectación, mientras que el 70 % restante se asocia con factores externos vinculados a la aleatoriedad no espacial del evento.

5 Consideraciones Finales

5.1 Producción técnica

En este trabajo, se desarrolló la función `SPDEsclm` aprovechando la integración entre `R/C++` a través de la librería `Rcpp`. Esta implementación permite optimizar las operaciones algebraicas, mejorando la eficiencia computacional del método propuesto. Esta función está disponible en el repositorio de `GitHub`, el cual se puede acceder a través del siguiente enlace: https://github.com/pabzun/SPDEsclm.

5.2 Conclusiones

El presente trabajo abordó los desafíos en los modelos estadísticos geoespaciales, particularmente en la presencia de datos censurados y faltantes. A través de un enfoque innovador basado en ecuaciones diferenciales parciales estocásticas (SPDEs), el modelo propuesto ofrece una solución eficiente para el manejo de grandes volúmenes de datos. En particular, se utiliza la aproximación de un campo aleatorio gaussiano (GRF) mediante un campo aleatorio gaussiano de Markov (GMRF), lo que optimiza el proceso de estimación por máxima verosimilitud y mejora significativamente la eficiencia computacional. Además, se proporciona una implementación de la metodología en el software estadístico R.

Las simulaciones realizadas validan propiedades clave del modelo, como el comportamiento asintótico esperado en los estimadores de máxima verosimilitud, y analizan el impacto del número de nodos en los tiempos de ejecución, ofreciendo criterios concretos para optimizar su desempeño.

La aplicación del modelo a datos reales de incendios forestales en la provincia de Pichincha demuestra su capacidad para capturar patrones espaciales complejos, identificar regiones críticas y generar predicciones confiables. Este enfoque no solo permite una comprensión más profunda del fenómeno analizado, gracias a la capacidad inferencial de los modelos estadísticos, sino que también respalda la formulación de políticas y estrategias fundamentadas en datos.

Finalmente, este trabajo establece un marco metodológico sólido y extensible, que integra innovaciones teóricas con implementaciones prácticas. Su contribución al campo de la estadística espacial no solo responde a las demandas actuales del análisis geoespacial, sino que también abre nuevas perspectivas para abordar escena-

rios en los que los datos no siguen formatos convencionales y donde las limitaciones computacionales representan desafíos cruciales.

5.3 Trabajos futuros

Una extensión natural de este estudio es explorar valores adicionales para el parámetro $\nu > 0$ de la función de correlación de Matérn, como se menciona en Lindgren *et al.* (2011). Esto implica que el proceso gaussiano con una correlación de Matérn de parámetro ν es solución de la SPDE $(\phi^{-2} - \boldsymbol{\Delta})^{\alpha/2} Z(\mathbf{s}) = 4\pi\phi^{-2}\mathcal{W}(\mathbf{s})$ donde $\alpha = \nu + 1$, lo que resulta en la construcción recursiva de la matriz de precisión $\mathbf{Q}_\phi$.

Adicionalmente, las observaciones pueden recopilarse en distintas estaciones o localidades a intervalos regulares, ya sean diarios, mensuales o anuales. En consecuencia, una línea de investigación futura podría ser la ampliación de este estudio para incluir el análisis de datos con respuestas censuradas o faltantes en variables que presentan correlación tanto espacial como temporal. Este enfoque tiene como objetivo optimizar los tiempos de ejecución del modelo propuesto por Valeriano *et al.* (2021).

Bibliografía

Anselin, L. (2013). *Spatial econometrics: methods and models*, volume 4. Springer Science & Business Media. Citado en la página 13.

Banerjee, S., Gelfand, A. E., Finley, A. O. & 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. Citado en la página 14.

Brenner, S. C. & Scott, L. R. (2008). *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer, third edition. ISBN 978-0-387-75934-0. Citado en la página 27.

Brody, S. D., Peck, B. M. & Highfield, W. E. (2004). Examining localized patterns of air quality perception in texas: A spatial and statistical analysis. *Risk Analysis: An International Journal*, **24**(6), 1561–1574. Citado en la página 13.

Ciarlet, P. G. (1978). *The Finite Element Method for Elliptic Problems*, volume 4 of *Studies in Mathematics and Its Applications*. North-Holland, Amsterdam. ISBN 9780444850287. Citado en la página 27.

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons. Citado 2 veces en las páginas 13 y 19.

De Oliveira, V. (2005). Bayesian inference and prediction of gaussian random fields based on censored data. *Journal of Computational and Graphical Statistics*, **14**(1), 95–115. Citado en la página 32.

Delyon, B., Lavielle, M. & Moulines, E. (1999). Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics*, **27**(1), 94–128. Citado 3 veces en las páginas 22, 23 y 31.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**(1), 1–38. Citado en la página 21.

Fridley, B. L. & Dixon, P. (2007). Data augmentation for a bayesian spatial model involving censored observations. *Environmetrics: The official journal of the International Environmetrics Society*, **18**(2), 107–123. Citado en la página 14.

Gaetan, C., Guyon, X. *et al.* (2010). *Spatial statistics and modeling*, volume 90. Springer. Citado en la página 13.

Griffin, T. W. (2010). The spatial analysis of yield data. *Geostatistical applications for precision agriculture*, pages 89–116. Citado en la página 13.

Guttorp, P. & Gneiting, T. (2006). Studies in the history of probability and statistics xlix on the matérn correlation family. *Biometrika*, **93**(4), 989–995. Citado en la página 20.

Hristopulos, D. T. (2020). *Gaussian Random Fields*, pages 245–307. Springer Netherlands, Dordrecht. Citado 2 veces en las páginas 13 y 18.

Juntto, S. & Paatero, P. (1994). Analysis of daily precipitation data by positive matrix factorization. *Environmetrics*, **5**(2), 127–144. Citado en la página 13.

Krainski, E., Gómez-Rubio, V., Bakka, H., Lenzi, A., Castro-Camilo, D., Simpson, D., Lindgren, F. & Rue, H. (2018). *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC. Citado en la página 35.

Kuhn, E. & Lavielle, M. (2005). Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*, **49**(4), 1020–1038. Citado en la página 23.

Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press. Citado en la página 23.

Lawson, A. B. (2006). *Exploratory Approaches, Parametric Estimation and Inference*, chapter 5, pages 67–107. John Wiley & Sons, Ltd. Citado en la página 14.

Lawson, A. B. (2013). *Statistical methods in spatial epidemiology*. John Wiley & Sons. Citado en la página 13.

Lindgren, F. & Rue, H. (2015). Bayesian spatial modelling with r-inla. *Journal of statistical software*, **63**(19). Citado en la página 35.

Lindgren, F. & Shewchuk, J. R. (2024). fmesher: Triangle meshes and related geometry tools. Version: 0.2.0. *The Journal of Open Source Software*. Citado en la página 28.

Lindgren, F., Rue, H. & 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. Citado 3 veces en las páginas 14, 27 y 46.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **44**(2), 226–233. Citado 2 veces en las páginas 16 y 31.

McIlhatton, D., McGreal, W., Taltavul de la Paz, P. & Adair, A. (2016). Impact of crime on spatial analysis of house prices: evidence from a uk city. *International Journal of Housing Markets and Analysis*, **9**(4), 627–647. Citado en la página 13.

Ordoñez, J. A., Bandyopadhyay, D., Lachos, V. H. & Cabral, C. R. (2018). Geostatistical estimation and prediction for censored responses. *Spatial statistics*, **23**, 109–123. Citado 3 veces en las páginas 14, 21 y 27.

R Core Team (2023). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Citado en la página 28.

Rue, H., Martino, S. & Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **71**(2), 319–392. Citado en la página 27.

Sahoo, I., Majumder, S., Hazra, A., Rappold, A. G. & Bandyopadhyay, D. (2024). Computationally scalable bayesian spde modeling for censored spatial responses. *arXiv preprint arXiv:2403.15670*. Citado 2 veces en las páginas 14 y 27.

Schabenberger, O. & Gotway, C. A. (2017). *Statistical Methods for Spatial Data Analysis*. Chapman and Hall/CRC, New York. ISBN 978-1-315-27508-6. Citado en la página 20.

Schelin, L. & Sjöstedt-de Luna, S. (2014). Spatial prediction in the presence of left-censoring. *Computational Statistics & Data Analysis*, **74**, 125–141. Citado en la página 34.

Sun, Y., Li, B. & Genton, M. G. (2012). Geostatistics for large datasets. In *Advances and Challenges in Space-time Modelling of Natural Events*, pages 55–77. Springer Berlin Heidelberg. Citado en la página 14.

Valeriano, K. A., Lachos, V. H., Prates, M. O. & Matos, L. A. (2021). Likelihood-based inference for spatiotemporal data with censored and missing responses. *Environmetrics*, **32**(3), e2663. Citado 4 veces en las páginas 21, 26, 27 y 46.

Wang, K., Abdulah, S., Sun, Y. & Genton, M. G. (2023). Which parameterization of the matérn covariance function? Citado en la página 20.

Wei, G. C. G. & Tanner, M. A. (1990). A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, **85**(411), 699–704. Citado en la página 22.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, **41**(3/4), 434–449. Citado en la página 27.

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*. Academic press. Citado en la página 13.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**(465), 250–261. Citado en la página 21.

Anexos

ANEXO A – Matriz de información observada

A.1 Cálculos necesarios para determinar la matriz de información observada

Para calcular la matriz de información de Fisher de los datos observados, mostrada en (3.7), es necesario determinar el Jacobiano y la matriz Hessiana del logaritmo de la función de verosimilitud de los datos completos. Estos cálculos son mostrados a continuación.

A.1.1 Primeras derivadas de la función de verosimilitud de los datos completos

Sea $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Psi}$, donde $\boldsymbol{\Psi} = \gamma\mathbf{R} + (1-\gamma)\mathbf{I}_n$, con $\mathbf{R} = \mathbf{A}\mathbf{Q}_\phi^{-1}\mathbf{A}^\top$ y $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$. Siendo $\dot{\ell}_c(\boldsymbol{\theta}|\mathbf{y}) = \dfrac{\partial \ell_c(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}$, $\dot{\mathbf{Q}}_\phi^{-1} = \dfrac{\partial \mathbf{Q}_\phi^{-1}}{\partial \phi} = \mathbf{Q}_\phi^{-1}\left(\dfrac{1}{2\pi\phi^3}\mathbf{D} - \dfrac{\phi}{2\pi}\mathbf{G}_2\right)\mathbf{Q}_\phi^{-1}$ y $\dot{\boldsymbol{\Psi}}_\phi = \dfrac{\partial \boldsymbol{\Psi}}{\partial \phi} = \gamma\mathbf{A}\dot{\mathbf{Q}}_\phi^{-1}\mathbf{A}^\top$, entonces los elementos del Jacobiano, $\dot{\ell}_c(\boldsymbol{\theta}|\mathbf{y})$, están dados por:

$$
\begin{aligned}
\dot{\ell}_{\boldsymbol{\beta}} &= \mathbf{X}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\dot{\ell}_{\sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\dot{\ell}_{\phi} &= -\frac{1}{2}tr\left(\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi\right) + \frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\dot{\ell}_{\gamma} &= -\frac{1}{2}tr\left(\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)\right) + \frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}).
\end{aligned}
$$

A.1.2 Segundas derivadas de la función de verosimilitud de los datos completos

Sean $\ddot{\boldsymbol{\Psi}}_\phi = \dfrac{\partial^2\boldsymbol{\Psi}}{\partial\phi^2} = \gamma\mathbf{A}\ddot{\mathbf{Q}}_\phi^{-1}\mathbf{A}^\top$ y $\ddot{\mathbf{Q}}_\phi^{-1} = \dfrac{\partial^2\mathbf{Q}_\phi^{-1}}{\partial\phi^2} = \dot{\mathbf{Q}}_\phi^{-1}\left(\dfrac{1}{2\pi\phi^3}\mathbf{D} - \dfrac{\phi}{2\pi}\mathbf{G}_2\right)\mathbf{Q}_\phi^{-1} - \boldsymbol{Q}_\phi^{-1}\left(\dfrac{3}{2\pi\phi^4}\mathbf{D} + \dfrac{1}{2\pi}\mathbf{G}_2\right)\mathbf{Q}_\phi^{-1} + \mathbf{Q}_\phi^{-1}\left(\dfrac{1}{2\pi\phi^3}\mathbf{D} - \dfrac{\phi}{2\pi}\mathbf{G}_2\right)\dot{\mathbf{Q}}_\phi^{-1}$. Luego, los elementos de la matriz Hessiana son dados por:

$$
\begin{aligned}
\ddot{\ell}_{\boldsymbol{\beta}\boldsymbol{\beta}^\top} &= -\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}\mathbf{X},\\
\ddot{\ell}_{\boldsymbol{\beta}\sigma^2} &= -\frac{1}{\sigma^2}\mathbf{X}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}),
\end{aligned}
$$

$$
\begin{aligned}
\ddot{\ell}_{\beta\phi} &= -\frac{1}{\sigma^2}\mathbf{X}^\top\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{\beta\gamma} &= -\frac{1}{\sigma^2}\mathbf{X}^\top\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{(\sigma^2)^2} &= \frac{n}{2\sigma^4}-\frac{1}{\sigma^6}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{\sigma^2\phi} &= -\frac{1}{2\sigma^4}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{\sigma^2\gamma} &= -\frac{1}{2\sigma^4}(\mathbf{y}-\boldsymbol{\mu})^\top\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{\phi^2} &= \frac{1}{2}tr\left(\mathbf{L}-\boldsymbol{\Psi}^{-1}\ddot{\boldsymbol{\Psi}}_\phi\right)-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top\left(2\mathbf{L}\boldsymbol{\Psi}^{-1}-\boldsymbol{\Psi}^{-1}\ddot{\boldsymbol{\Psi}}_\phi\boldsymbol{\Psi}^{-1}\right)(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{\phi\gamma} &= \frac{1}{2}tr\left(\mathbf{K}-\boldsymbol{\Psi}^{-1}\mathbf{A}\dot{\mathbf{Q}}_\phi^{-1}\mathbf{A}^\top\right)-\frac{1}{2\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top\left(\mathbf{K}\boldsymbol{\Psi}^{-1}-\mathbf{M}+\boldsymbol{\Psi}^{-1}\mathbf{K}^\top\right)(\mathbf{y}-\boldsymbol{\mu}),\\
\ddot{\ell}_{\gamma^2} &= \frac{1}{2}tr\left(\mathbf{J}\right)-\frac{1}{\sigma^2}(\mathbf{y}-\boldsymbol{\mu})^\top\mathbf{J}\boldsymbol{\Psi}^{-1}(\mathbf{y}-\boldsymbol{\mu}),
\end{aligned}
$$

tal que $\mathbf{J}=\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)$, $\mathbf{K}=\boldsymbol{\Psi}^{-1}(\mathbf{R}-\mathbf{I}_n)\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi$, $\mathbf{L}=\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi\boldsymbol{\Psi}^{-1}\dot{\boldsymbol{\Psi}}_\phi$ y $\mathbf{M}=\boldsymbol{\Psi}^{-1}\mathbf{A}\dot{\mathbf{Q}}_\phi^{-1}\mathbf{A}^\top\boldsymbol{\Psi}^{-1}$.