PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Modelo bayesiano espacio-temporal DAGAR para estimar los casos de operaciones sospechosas en la provincia de Lima

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

Joseph Bryan Diaz Quispe

ASESORA:

Zaida Jesús Quiroz Cornejo

Lima, 2025

Informe de Similitud

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Modelo bayesiano espacio-temporal DAGAR para estimar los casos de operaciones sospechosas en la provincia de Lima*, del autor Joseph Bryan Diaz Quispe, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 17 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 22/08/2025.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 22 de agosto de 2025

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

Dedicatoria

Dedico esta tesis, en primer lugar, a mi esposa, por ser mi mayor fuente de motivación para seguir estudiando y aprendiendo, por su apoyo incondicional y su compañía constante a lo largo de todo este proceso.

Asimismo, dedico este trabajo a mis padres, quienes desde mis primeros pasos me inculcaron el valor del esfuerzo, la responsabilidad y el deseo de superación. Gracias por creer en mí, y por brindarme las oportunidades y el respaldo necesarios para mi formación académica.

Agradecimentos

En primer lugar agradezco a Dios por la oportunidad de completar mis estudios de maestría.

En segundo lugar, agradezco a la profesora Zaida Quiroz, mi asesora, por su valioso apoyo, guía e instrucción a lo largo de la elaboración de esta tesis. Su acompañamiento constante, sus conocimientos y compromiso han sido fundamentales para el desarrollo de este trabajo. Le agradezco también por haberme introducido al campo de la estadística espacial, que ha enriquecido profundamente mi formación académica.

Finalmente, agradezco al staff docente de la maestría en estadística por haber contribuido a mi formación, y a mis compañeros por el apoyo que me brindaron.

Resumen

En esta tesis se propone un enfoque bayesiano para la estimación de modelos espacio-temporales de áreas basados en la estructura DAGAR (Directed Acyclic Graph Autoregressive). Se desarrollan dos variantes del modelo: una con componentes espacio-temporales lineales y otra que incorpora una dinámica temporal modelada mediante un paseo aleatorio de primer orden. Ambos modelos incluyen efectos espaciales estructurados y no estructurados, y son estimados utilizando inferencia bayesiana a través del método de aproximaciones de Laplace integradas e iteradas (INLA). Se realizaron estudios de simulación para evaluar el desempeño de los modelos bajo distintos niveles de autocorrelación espacial, analizando la precisión en la estimación de parámetros. Finalmente, se aplicó el modelo a datos reales sobre reportes de operaciones sospechosas (ROS) en el Perú, permitiendo identificar patrones espacio-temporales relevantes para las políticas de prevención y lucha contra el lavado de activos.

Palabras-clave: DAGAR, Datos de área, INLA, Lavado de Activos, modelo espacio-temporal.

Abstract

This thesis proposes a Bayesian approach for the estimation of area-level spatio-temporal models based on the DAGAR (Directed Acyclic Graph Autoregressive) structure. Two model variants are developed: one with linear space-time components, and another that incorporates a temporal dynamic modeled through a first-order random walk. Both models include structured and unstructured spatial effects and are estimated using Bayesian inference through the Integrated Nested Laplace Approximation (INLA) method. Simulation studies were conducted to evaluate model performance under different levels of spatial autocorrelation, focusing on parameter estimation accuracy. Finally, the model was applied to real data on suspicious transaction reports (ROS) in Peru, enabling the identification of relevant spatio-temporal patterns to inform anti-money laundering and prevention policies.

Keywords: Areal data, DAGAR, INLA, Money Laundering, spatio-temporal model.

Índice general

Capítulo 1

Introducción

1.1. Consideraciones preliminares

El lavado de activos es un fenómeno omnipresente y complejo que amenaza la integridad de las instituciones financieras y socava la estabilidad económica y social de los países. En el caso del Perú, donde existen múltiples fuentes de ingresos ilegales, como el narcotráfico, la minería ilegal y la corrupción, el lavado de activos representa un desafío significativo (SBS, 2021). Por lo tanto, llevar a cabo una investigación exhaustiva sobre este tema es esencial para comprender la magnitud del problema, identificar sus dinámicas y así proveer de información a las autoridades para proteger así los intereses económicos y sociales del país.

El análisis de registros de operaciones sospechosas (ROS) proporciona una ventana única para comprender la dinámica y la escala del lavado de activos en el Perú. Estos registros contienen información detallada sobre transacciones financieras que levantan sospechas debido a su naturaleza inusual o sospechosa. Al examinar los ROS, se pueden identificar patrones, tendencias y redes de lavado de activos, lo que permite una comprensión más profunda de cómo se llevan a cabo estas actividades ilícitas en el país.

Se han llevado a cabo diversos estudios utilizando registros de operaciones sospechosas para abordar distintos aspectos del lavado de activos. Los análisis de tendencias, por ejemplo, han permitido identificar cambios a lo largo del tiempo en los patrones de actividades sospechosas, proporcionando información crucial sobre la evolución de las estrategias de lavado de activos y la efectividad de las medidas de prevención (SBS, 2023). Por otro lado, los estudios sectoriales se han enfocado en sectores económicos específicos, como el minero, pesquero o maderero, para comprender cómo estas industrias son utilizadas para lavar dinero y para desarrollar estrategias de prevención adaptadas a sus características particulares (SBS, 2018). Además, las evaluaciones de riesgos basadas en ROS han permitido identificar

áreas de mayor vulnerabilidad y concentración de actividades sospechosas, proporcionando una base sólida para la asignación de recursos y la implementación de medidas de control más efectivas (SBS, 2023).

En particular en esta tesis se van a estudiar los registros de operaciones sospechosas en la provincia de Lima, datos que incluyen información geográfica detallada, como el distrito donde se realizó la operación. Debido a las características de los casos de registros de operaciones a nivel distrital, es razonable pensar que el número de casos de ROS es similar en distritos vecinos. Esta información permitiría identificar patrones espaciales y entender la distribución espacial de las actividades sospechosas. Al analizar la cantidad de ROS a nivel distrital y analizar estos datos a lo largo de varios años, se pueden identificar áreas de mayor concentración de actividades sospechosas, así como cambios en los patrones de lavado de activos a nivel regional. Esto proporcionaría información valiosa para la formulación de políticas y la asignación de recursos, permitiendo una respuesta más efectiva y focalizada en la prevención y combate del lavado de activos en diferentes áreas geográficas de la provincia en estudio.

En lo que respecta a estudios estadísticos con datos relacionados al lavado de activos, Ferwerda et al. (2020) emplean un modelo de gravedad en su estudio para estimar los flujos de lavado de dinero a nivel mundial, revelando las preferencias de los lavadores de dinero y proporcionando estimaciones que distinguen entre los desafíos de políticas relacionados con el lavado de ganancias de crimen doméstico, inversión internacional de dinero sucio y flujos de dinero a través de un país. Xia et al. (2022) proponen un modelo híbrido de predicción espacio-temporal de lavado de dinero que combina redes neuronales convolucionales y memoria a corto y largo plazo, destacando la importancia de capturar las dependencias temporales y espaciales complejas en los datos de lavado de activos. Por otro lado, Yulia et al. (2021) utilizan una regresión logística multinomial geoponderada para modelar la variación espacial de los delitos de lavado de dinero en Indonesia, considerando la influencia de la ubicación geográfica en los factores causales de estos delitos. Además, Singh and Best (2019) exploran el uso de técnicas de visualización de datos para identificar actividades sospechosas de lavado de dinero, resaltando la importancia de las soluciones tecnológicas mejoradas en la lucha contra el lavado de activos.

Con respecto a modelos espaciales de área, según Gelfand et al. (2010) los modelos espaciales son particularmente útiles en el contexto de datos de área, donde la autocorrelación espacial juega un papel importante en la distribución de los datos. Estos modelos son capaces de capturar la dependencia espacial entre áreas vecinas al introducir términos que incorporan

la autocorrelación espacial en el modelo. Los modelos de datos de áreas más conocidos son el modelo autoregresión condicional (CAR) y el modelo de autoregresión simultánea (SAR), (Cressie, 1993; Banerjee et al., 2004).

Otro enfoque para modelos espaciales es el modelo dirigido acíclico autoregresivo (DAGAR) propuesto por Datta et al. (2019), donde en lugar de considerar el dominio geográfico como un grafo no dirigido, el modelo DAGAR utiliza un grafo acíclico dirigido (DAG) derivado del grafo no dirigido original. Definen distribuciones condicionales para los efectos aleatorios espaciales utilizando secuencia de árboles locales creados a partir del DAG, cuya distribución es normal y tiene una de covarianza similar a la de un proceso autorregresivo. Una característica importante del modelo DAGAR es que garantiza que la matriz de covarianza resultante sea definida positiva, lo que permite generar o modelar directamente datos multivariados gaussianos con una estructura de dependencia derivada de un grafo. Además, el parámetro ρ de la matriz de precisión en el modelo DAGAR tiene una mejor interpretación como un parámetro de autocorrelación espacial, lo que resuelve una importante dificultad que tienen los modelos autorregresivos condicionales (CAR) y simultáneos (SAR).

Con respecto a modelos espacio-temporales para datos de áreas, Blangiardo et al. (2013) incorporan en un modelo espacial un efecto aleatorio temporal para capturar la evolución espacial a lo largo del tiempo. En este modelo cada área está influenciada tanto por las observaciones pasadas en la misma ubicación como por las observaciones pasadas en áreas espaciales y temporales vecinas. Esto implica que la estructura del modelo incorpora términos de efectos aleatorios tanto espaciales como temporales, lo que permite capturar patrones complejos de dependencia en los datos espacio-temporales.

En cuanto a la inferencia, el enfoque de Aproximación de Laplace Anidada e Integrada (INLA) se destaca sobre el método de Cadenas de Markov y Monte Carlo (MCMC) en inferencia bayesiana para modelos espaciales y espacio-temporales debido a su eficiencia computacional, amplia aplicabilidad a una variedad de modelos gaussianos latentes y facilidad de uso a través del R. El método INLA proporciona una alternativa computacionalmente eficiente que permite realizar análisis bayesianos en modelos complejos o con grandes conjuntos de datos de manera más rápida (Rue et al., 2009).

En este contexto, esta tesis tiene como objetivo implementar el modelo espacio-temporal DAGAR para datos de áreas a través del enfoque de inferencia bayesiana usando INLA, y aplicar el modelo a datos de registros de operaciones sospechosas (ROS) en la provincia de Lima a nivel distrital. La elección del modelo DAGAR se fundamenta en su capacidad para capturar de manera eficiente la estructura espacio-temporal de los datos, proporcionando es-

timaciones más precisas y robustas del parámetro de autocorrelación espacial. Esto permitirá una mejor comprensión y análisis de los patrones y dinámicas del lavado de activos. La implementación del modelo se evaluará mediante estudios de simulación intensiva en diferentes escenarios. Finalmente, se aplicará el modelo a los datos reales del ROS y se comparará su desempeño con otros modelos existentes para validar su eficacia y potencial como herramienta para la detección y prevención de actividades sospechosas.

1.2. Objetivos

El objetivo general de la tesis es implementar el modelo espacio-temporal DAGAR, cuya inferencia bayesiana se realizará usando INLA, para finalmente aplicarlo a la base de datos real del ROS en la provincia de Lima a nivel distrital.

De manera específica:

- Proponer el modelo espacio-temporal DAGAR mediante el enfoque INLA.
- Implementar la estimación de los parámetros de los modelos propuestos usando INLA.
- Realizar estudios de simulación acerca del modelo DAGAR considerando computación intensiva sobre diferentes escenarios.
- Aplicar el modelo al conjunto de datos sobre los ROS y comparar el ajuste con otros modelos competencia.

1.3. Organización del trabajo

En el capítulo 2 se presentan conceptos preliminares sobre los modelos geoestadísticos, como: autocorrelación espacial, datos geoestadísticos, procesos espaciales, modelos CAR y SAR que permitirán comprender y aplicar adecuadamente el modelo propuesto. En el capítulo 3 se aborda el modelo propuesto, es decir, el modelo espacio-temporal DAGAR aplicado a datos de conteo del tipo de área y el método de estimación INLA. En el capítulo 4, se presenta un estudio de simulación con los diferentes escenarios para la generación de datos, con el fin de evaluar la efectividad del modelo. En el capítulo 5, se presentará la aplicación de los modelos propuestos usando datos reales. Finalmente, en el capítulo 6 se presentarán las conclusiones y posibles trabajos futuros.

Capítulo 2

Conceptos

2.1. Introducción a datos de áreas

En esta sección se abordan conceptos preliminares clave relacionados con los modelos de datos de área. Los datos de área se obtienen al dividir un dominio fijo en un número limitado de áreas, que pueden ser regulares (como una grilla) o irregulares (como departamentos o provincias). Sean $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_n)^\top$ las variables aleatorias para cada área $1, 2, \ldots, n$. Una característica importante de estas unidades de área es la presencia de autocorrelación espacial, lo cual indica que las áreas vecinas tienden a tener valores de medición similares. Para tener en cuenta esta autocorrelación espacial en el modelo, se representa la estructura de vecindad entre las áreas mediante un grafo, que a su vez se utiliza para crear la matriz de vecindad.

Consideremos que $\mathcal{G} = \{V, E\}$ es un grafo no dirigido con n nodos en V, donde cada nodo corresponde a un área, y con aristas E entre las áreas vecinas. Por ejemplo, la Figura 2.1 muestra un grafo no dirigido de 5 nodos $\boldsymbol{V} = \{v_1, v_2, v_3, v_4, v_5\}$ y aristas $\boldsymbol{E} = \{(v_1, v_4), (v_4, v_2), (v_2, v_5), (v_4, v_5), (v_5, v_3)\}$ las cuales representan las vecindades entre los nodos.

Otro concepto clave es la matriz de vecindad $\boldsymbol{W}$, esta es una matriz de pesos w_{ij} que representa la estructura de vecindad. Los pesos w_{ij} están asociados a las áreas i y j. La relación entre estas áreas vecinas puede representarse mediante la matriz de vecindad:

$$\boldsymbol{W} = \begin{pmatrix} w_{1,1} & \cdots & w_{1,n} \\ \vdots & \ddots & \vdots \\ w_{n,1} & \cdots & w_{n,n} \end{pmatrix}.$$

De esta forma $w_{ij} \neq 0$ si y solo si i y j son áreas vecinas, y $w_{ij} = 0$ si no son áreas vecinas

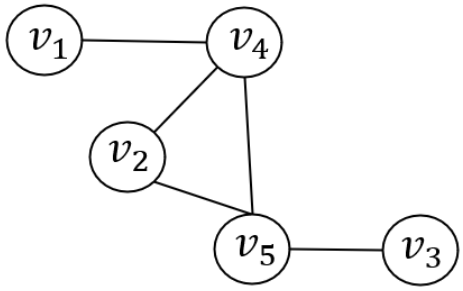

Figura 2.1: Grafo no dirigido de 5 nodos.

(Blangiardo et al., 2013). La matriz w_{ij} se puede interpretar como pesos, donde el peso será mayor cuanto más cercanas estén las áreas. Continuando con el ejemplo de la Figura 2.1, la matriz de vecindad asociada al grafo $\mathcal{G}$ es:

$$\boldsymbol{W} = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix},$$

donde $w_{ij} = 1$ si i y j comparten límites, y $w_{ij} = 0$ de lo contrario.

Para definir un modelo espacial para datos de área, se puede incorporar la autocorrelación espacial mediante efectos aleatorios espaciales u_i para cada área $i = 1, \ldots, n$. Dentro de los modelos espaciales de área, los más populares son los modelos intrínsecamente autoregresivos (ICAR), los modelos autoregresivos condicionales (CAR) y los modelos autoregresivos simultaneos (SAR). A continuación se definen brevemente estos modelos y el modelo dirigido acíclico autoregresivo (DAGAR) que ciertas definiciones de estos modelos incorporando grafos acíclicos.

2.2. Modelos para datos de áreas

2.2.1. Modelo intrínsecamente autoregresivo (IAR)

El modelo IAR fue introducido por Besag (1974). Consideremos que una región geográfica está dividida en áreas $i = 1, \ldots, n$. El vector de efectos aleatorios espaciales para estas n áreas se define como $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^\top$. En el modelo IAR, se asume que el efecto espacial aleatorio estructurado u_i de una área i, condicionado al resto de los efectos espaciales aleatorios $\boldsymbol{u}_{-i} = (u_1, \ldots, u_{i-1}, u_{i+1}, \ldots, u_n)^\top$, tiene una distribución normal:

$$u_i \mid \boldsymbol{u}_{-i} \sim N\left(\sum_{j=1}^{n} b_{ij}u_j, \sigma_i^2\right), \tag{2.1}$$

donde b_{ij} son constantes, con $b_{ii} = 0$ para todo $i = j$ y varianza σ_i^2.

Las distribuciones condicionales completas en la ecuación (2.1) definen una distribución a priori $\pi(\boldsymbol{u})$ si $b_{ij}\sigma_i^2 = b_{ji}\sigma_j^2$ para todo $i, j = 1, \ldots, n$, en cuyo caso se tiene que $b_{ij} = w_{ij}/w_{i+}$ y $\sigma_i^2 = \sigma^2/w_{i+}$ siendo w_{i+} el número de vecinos del área i. Si $\sigma^2 = \frac{1}{\tau}$, donde τ es un parámetro de precisión (inversa de la varianza), luego el modelo IAR asume que la distribución a priori de $\boldsymbol{u}$ tiene la siguiente forma:

$$\pi(\boldsymbol{u}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{u}^\top[\tau(\boldsymbol{D_W} - \boldsymbol{W})]\boldsymbol{u}\right\} \tag{2.2}$$

donde $\boldsymbol{D_W} = diag(w_{i+})$.

2.2.2. Modelo condicional autoregresivo (CAR)

La distribución a priori en la ecuación (2.2) define una distribución conjunta normal multivariada para $\boldsymbol{u}$ si:

$$\pi(\boldsymbol{u}) \propto \exp\left\{-\frac{1}{2}\boldsymbol{u}^\top[\tau(\boldsymbol{D_W} - \rho\boldsymbol{W})]\boldsymbol{u}\right\} \tag{2.3}$$

donde ρ es un parámetro espacial conocido como parámetro de autocorrelación espacial. Entonces de la ecuación (2.3), se tiene que $\boldsymbol{u}$ sigue una distribución normal multivariada, esto es:

$$\boldsymbol{u} \sim N\left(0, (\boldsymbol{D_W} - \rho\boldsymbol{W})^{-1}/\tau\right),$$

y la matriz de precisión es $\boldsymbol{Q}_{CAR} = \tau(\boldsymbol{D_W} - \rho\boldsymbol{W})$.

Luego los efectos aleatorios espaciales u_i pueden ser incorporados en un modelo espacial. Básicamente, si $E(Y_i) = \mu_i$, entonces:

$$g(\mu_i) = z_i^\top\boldsymbol{\beta} + u_i,$$

donde $g(\cdot)$ representa una función de enlace apropiada, z_i forma un vector de covariables, y $\boldsymbol{\beta}$ representa un vector de coeficientes de regresión.

2.2.3. Modelos autoregresivo simultaneo (SAR)

El modelo autorregresivo simultáneo (SAR) fue introducido por Whittle (1954). El modelo SAR modela de manera simultanea los efectos espaciales aleatorios u_i. Se considera que el efecto espacial aleatorio de un área i, depende de los efectos aleatorios de sus vecinos, es decir,

$$u_i = \rho \sum_{j \neq i} b_{ij} u_j + \epsilon_i, \text{ for } i = 1, 2, \ldots, n, \tag{2.4}$$

asumiendo que $\epsilon_i \overset{\text{ind}}{\sim} N(0, 1/\tau_i)$ representan errores no estructurados, es decir, errores que no siguen un patrón espacial, además, se supone que los errores ϵ_i son independientes de los efectos espaciales u_i, b_{ij} son constantes conocidas. A partir de la ecuación (2.4), se deduce lo siguiente:

$$\begin{aligned} u_1 &= \rho \sum_{j \neq 1} b_{1j} u_j + \epsilon_1 \\ u_2 &= \rho \sum_{j \neq 2} b_{2j} u_j + \epsilon_2 \\ &\vdots \\ u_n &= \rho \sum_{j \neq n} b_{nj} u_j + \epsilon_n, \end{aligned}$$

luego $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^T$ sigue una distribución normal, esto es,

$$\boldsymbol{u} \sim N\left(0, (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{F} \left((\boldsymbol{I} - \boldsymbol{B})^{-1}\right)^T\right),$$

donde $\boldsymbol{F} = \text{diag}(1/\tau_1, \ldots, 1/\tau_n) = \boldsymbol{I_n}/\tau$ y la matriz $\boldsymbol{B} = \rho \boldsymbol{W}$ depende de un parámetro de autocorrelación espacial ρ, y $(\boldsymbol{I} - \boldsymbol{B})$ es una matriz positiva definida si ρ está acotada en el intervalo $\left(\lambda_{(1)}^{-1}, \lambda_{(n)}^{-1}\right)$ donde $\left(\lambda_{(1)} \leq \lambda_{(2)} \leq \cdots \leq \lambda_{(n)}\right)$ son los autovalores ordenados de $\boldsymbol{W}$, con $\lambda_{(1)} < 0 < \lambda_{(n)}$. Para restringir $|\rho| < 1$, la matriz de vecindad puede ser reescalada dividiéndola por $\lambda_{(n)}$, es decir, $\widetilde{\boldsymbol{W}} = \boldsymbol{W}/\lambda_{(n)}$ según Haining (2003). Por lo general, ρ se interpretar de manera similar a un parámetro de autocorrelación espacial. Finalmente, la distribución SAR de $\boldsymbol{u}$ está dada por:

$$\boldsymbol{u} \sim N\left(0, \frac{1}{\tau}(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}})^{-1} \left((\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}})^{-1}\right)^\top\right), \tag{2.5}$$

luego la matriz de precisión $\boldsymbol{Q}_{SAR} = \tau(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}})(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}})^\top$.

Los efectos aleatorios espaciales u_i se pueden incorporar en el modelo, de forma similar

que en el modelo CAR. Si $E(Y_i) = \mu_i$ entonces,

$$g(\mu_i) = z_i^\top \boldsymbol{\beta} + u_i$$

En este contexto, $g(\cdot)$ representa una función de enlace apropiada, z_i s un vector que contiene las covariables relevantes, y $\boldsymbol{\beta}$ es un vector de coeficientes de regresión.

En cuanto a la interpretación del parámetro ρ, presenta problemas similares a los modelos CAR, pues tienden a sobreestimar la autocorrelación espacial.

2.2.4. Modelo dirigido acíclico autoregresivo (DAGAR)

Sea $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)^\top$ un vector de dimensión $n \times 1$ que contiene los efectos aleatorios espaciales para cada área numerada de $1, 2, \ldots, n$. La función de densidad conjunta de $\boldsymbol{u}$ puede ser definida considerando las distribuciones condicionales de los u_i:

$$\pi(\boldsymbol{u}) = \pi(u_1)\,\pi(u_2 \mid u_1)\,\pi(u_3 \mid u_1, u_2)\ldots\pi(u_k \mid u_1, u_2, \ldots, u_{k-1}).$$

De esta forma cada efecto aleatorio espacial u_i está influenciado por los efectos aleatorios de las áreas vecinas más próximas en el "pasado". Este enfoque generaliza la definición del modelo SAR, pues cada efecto espacial aleatorio u_i puede ser definido por:

$$\begin{aligned} u_1 &= \epsilon_1; \\ u_2 &= b_{21}u_1 + \epsilon_2; \\ &\vdots \\ u_n &= b_{n1}u_1 + \cdots + b_{n,n-1}u_{n-1} + \epsilon_n, \end{aligned}$$

donde $\epsilon_i \overset{\text{ind}}{\sim} N(0, 1/\tau_i)$ son los errores independientes de $\boldsymbol{u}$. Generalizando este resultado, cualquier efecto aleatorio u_i se define de la siguiente manera:

$$u_i = \sum_{\substack{j<i \\ j\sim i}} b_{ij}u_j + \epsilon_i; \quad \text{para } i = 1, 2, \ldots, n, \tag{2.6}$$

donde $j \sim i$ representa las áreas vecinas de i, b_{ij} son coeficientes y $\epsilon_i \overset{\text{ind}}{\sim} N(0, 1/\tau_i)$.

Para definir el modelo DAGAR, primero se necesita definir el conjunto de vecinos del área i como $N(i) = \{j < i, j \sim i\}$, lo que significa que el área j es vecino del área i. Además, se fija $b_{ij} = 0$ para todos los $j \notin N(i)$. Para determinar qué áreas son vecinas y definir $b_{ij} \neq 0$, es decir, cuáles son las áreas vecinas (pasadas y más cercanas), se utiliza un enfoque

basado en grafos acíclicos (DAG). Sea un grafo acíclico (DAG) $\mathcal{G} = (\boldsymbol{V}, \boldsymbol{E})$ donde los nodos V representan las áreas y las aristas E conectan las áreas vecinas. Llamamos a $\mathcal{G}_i$ a un subgrafo de $\mathcal{G}$ que incluye el nodo i y sus vecinos. Datta et al. (2019) propusieron utilizar árboles de expansión locales de subgrafos pequeños de $\mathcal{G}$ para construir densidades condicionales de menor dimensión, y así definir $b_{ij} \neq 0$. Específicamente, proponen utilizar un árbol de expansión incrustado T_i de $\mathcal{G}_i$ para establecer el conjunto final de vecinos $N(i)$ y la función de densidad condicional $u_i \mid \boldsymbol{u}_{N(i)}$ que se asume una distribución normal con media cero y matriz de covarianza definida positiva. A través del árbol de expansión T_i y la matriz de covarianza de un proceso autoregresivo de orden uno (AR(1)), se define la covarianza entre el efecto aleatorio u_i y los efectos aleatorios de sus vecinos $\boldsymbol{u}_{N(i)}$, de la siguiente manera:

$$\operatorname{cov}\left(u_i, \boldsymbol{u}_{N(i)}\right) = \begin{pmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho^2 & \cdots & \rho^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho^2 & \rho^2 & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & v_i^T \\ v_i & \sum_i \end{pmatrix} \tag{2.7}$$

donde el parámetro de autocorrelación espacial $0 \leq \rho \leq 1$, la matriz con elementos $\rho^{d_{ij}}$ es definida positiva, donde d_{ij} representa la longitud del camino más corto en $\mathcal{G}$ entre los nodos i y j; y v_i es el vector de covarianzas entre u_i y $\boldsymbol{u}_{N(i)}$, mientras que $\sum_i$ es la matriz de covarianza de $\boldsymbol{u}_{N(i)}$. Luego, a partir de la distribución condicional normal de $u_i \mid \boldsymbol{u}_{N(i)}$ con media cero y matriz de covarianza definida en la ecuación (2.7), y la definición del modelo DAGAR en la ecuación(2.6), el valor esperado y varianza condicional de $u_i \mid \boldsymbol{u}_{N(i)}$ son iguales a $E\left(u_i \mid \boldsymbol{u}_{N(i)}\right) = \sum_{j \in N(i)} b_{ij} u_j$, y $\tau_i = 1/\operatorname{Var}\left(u_i \mid \boldsymbol{u}_{N(i)}\right)$, obteniéndose que:

$$b_{ij} = \frac{\rho}{1 + (n_{<i})\,\rho^2}; \quad i = 2, \ldots, n; \quad j \in N(i) \tag{2.8}$$

$$\tau_i = \frac{1 + (n_{<i} - 1)\,\rho^2}{1 - \rho^2}, \quad i = 1, \ldots, n, \tag{2.9}$$

donde $n_{<i}$ representa la cardinalidad de $N(i)$, es decir, el número de vecinos de un área i.

De forma matricial, la ecuación (2.6) se puede escribir como $\boldsymbol{u} = \boldsymbol{B}\boldsymbol{u} + \boldsymbol{\epsilon}$, donde $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{F})$, $\boldsymbol{F} = \operatorname{diag}(\tau_1, \tau_2, \ldots, \tau_n)$ y $\boldsymbol{B} = b_{ij}$ es una matriz estrictamente triangular inferior, y las definiciones de b_{ij} y τ_i están dadas en las ecuaciones (2.8) y (2.9), respectivamente. Por lo tanto, $\boldsymbol{u} \sim N\left(0, (\boldsymbol{I} - \boldsymbol{B})^\top \boldsymbol{F}(\boldsymbol{I} - \boldsymbol{B})\right)$. Es crucial notar que la matriz de precisión del modelo DAGAR, denotada por $\boldsymbol{Q}(\rho) = \left[(\boldsymbol{I} - \boldsymbol{B})^T \boldsymbol{F}(\boldsymbol{I} - \boldsymbol{B})\right]^{-1}$ (inversa de la matriz de covarianza) es dispersa, esto significa que tiene muchos ceros, lo que la hace adecuada para conjuntos de datos masivos que involucran áreas geográficas extensas.

La definición completa del modelo DAGAR usado como distribución de los efectos aleatorios asume que: $\boldsymbol{u} \sim N\left(0, \boldsymbol{Q}^{-1}(\rho)/\tau_u\right)$, donde $\boldsymbol{Q}(\rho)$ es la matriz de precisión, y $1/\tau_u$ es la varianza marginal.

Como no existe un orden natural entre las áreas, Datta et al. (2019) asumen que $\boldsymbol{u} \sim N\left(0, \boldsymbol{Q}^{\star -1}(\rho)/\tau_u\right)$, donde $\boldsymbol{Q}^{\star}(\rho)$ es definida como el promedio de las matrices de precisión DAGAR bajo todas las posibles permutaciones π, es decir,

$$\boldsymbol{Q}^{\star}(\rho) = \frac{1}{k!} \sum_{\pi} \boldsymbol{P}_{\pi}^{\top} \boldsymbol{L}_{\pi}^{\top} \boldsymbol{F}_{\pi} \boldsymbol{L}_{\pi} \boldsymbol{P}_{\pi}, \tag{2.10}$$

donde $\boldsymbol{P}_{\pi}$ es la matriz de permutación matriz correspondiente a π, es decir, $\boldsymbol{P}_{\pi} \left(u_1, \ldots, u_n\right)^{\top} = \left(u_{\pi(1)}, \ldots, u_{\pi(k)}\right)^{\top}$, como para un orden π, se tiene que $\boldsymbol{Q}_{\pi}(\rho) = \boldsymbol{L}_{\pi}^{T} \boldsymbol{F}_{\pi} \boldsymbol{L}_{\pi}$, entonces $\boldsymbol{F}_{\pi} = \text{diag}\left(\tau_{\pi(1)}, \tau_{\pi(2)}, \ldots, \tau_{\pi(k)}\right)$, $\boldsymbol{L}_{\pi} = \boldsymbol{I} - \boldsymbol{B}_{\pi}$ y $\boldsymbol{B}_{\pi} = \{b_{ij}\}$ para el orden π determinado. Cabe resaltar que $\boldsymbol{Q}^{\star}(\rho)$ también es una matriz dispersa.

2.3. Inferencia bayesiana

La inferencia bayesiana constituye un enfoque estadístico basado en el uso de la probabilidad para describir la incertidumbre sobre parámetros desconocidos. Bajo este enfoque, el parámetro de interés θ se trata como variable aleatoria cuya incertidumbre se modela mediante una distribución de probabilidad. El procedimiento inferencial se fundamenta en el teorema de Bayes, tal que:

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\pi(y)} \propto \pi(y|\theta)\pi(\theta), \tag{2.11}$$

donde $\pi(\theta)$ es la función de probabilidad o densidad (fdp) de la distribución a priori, y representa la información previa sobre los parámetros, $\pi(y|\theta)$ es la función de verosimilitud asociada a la información de los datos de la muestra aleatoria, y $\pi(\theta|y)$ es la distribución a posteriori que sintetiza la información previa y la evidencia empírica (Bernardo and Smith, 2000). Así la ecuación (2.11) establece la relación entre la distribución a priori, la función de verosimilitud y la distribución a posteriori.

Este marco probabilístico ofrece varias ventajas respecto al enfoque frecuentista. En particular, permite cuantificar la incertidumbre de manera más completa al proporcionar distribuciones posteriores de los parámetros en lugar de solo estimadores puntuales, así como construir intervalos de credibilidad con una interpretación directa en términos de probabilidad. Además, facilita la especificación de modelos jerárquicos y la incorporación de conocimiento

experto a través de las distribuciones a priori, características esenciales en contextos de alta complejidad estructural.

En términos computacionales, el cálculo exacto de las distribuciones posteriores es en la mayoría de los casos intratable, ya que requiere evaluar integrales de alta dimensión. Para superar esta limitación se han desarrollado métodos aproximados, entre los que destacan los algoritmos de Monte Carlo basados en cadenas de Markov (MCMC), ampliamente utilizados por su flexibilidad (Robert and Casella, 2004), y el enfoque de Aproximación de Laplace Anidada Integrada (INLA), que resulta especialmente eficiente en modelos latentes gaussianos con aplicaciones espaciales y espacio-temporales (Rue et al., 2009).

Capítulo 3

Modelo espacio-temporal DAGAR

En este capítulo se define el modelo espacio temporal usando efectos aleatorios cuya distribución se define a través del modelo DAGAR. Asimismo se detalla cómo se realiza la inferencia bayesiana para el modelo propuesto usando INLA.

3.1. Definición de modelos espacio-temporales DAGAR

Sea y_{it} variable aleatoria que representa el número de casos en el área i y tiempo t, para $i = 1, \ldots, n; \quad t = 1, \ldots, T$. Se asume que y_{it} sigue una distribución de Poisson, tal que $y_{it} \sim \text{Poisson}(\lambda_{it})$, donde la media λ_{it} es el número de casos esperado en cada área i y tiempo t. Además, λ_{it} está definida en términos de una tasa ρ_{it} y el número esperado de recuentos E_{it}, de tal forma que $\lambda_{it} = \rho_{it} E_{it}$.

Se asume que el predictor lineal η_{it} es asociado a ρ_{it} a través de una función de enlace logarítmica y se puede expresar como:

$$\eta_{it} = \log(\rho_{it}) = \alpha + f_{si} + f_{ti}, \tag{3.1}$$

donde α es un intercepto, f_{si} es una estructura espacial y f_{ti} es una estructura temporal.

3.1.1. Estructura espacial DAGAR

Respecto a la estructura espacial se asume que:

$$f_{si} = v_i + \nu_i,$$

donde v_i en un efecto aleatorio espacial estructurado y ν_i es un efecto aleatorio no estructurado.

Se asume que v_i es definido mediante un modelo autoregresivo acíclico dirigido (DAGAR). Entonces el modelo espacio temporal DAGAR que se propone en esta tesis asume que:

$$\boldsymbol{v} \sim N\left(0, \boldsymbol{Q}^{-1}(\rho)/\tau_v\right),$$

donde $1/\tau_v$ es la varianza marginal, y

$$\boldsymbol{Q}(\rho) = \frac{1}{k!}\sum_{\pi} \boldsymbol{P}_\pi^\top \boldsymbol{L}_\pi^\top \boldsymbol{F}_\pi \boldsymbol{L}_\pi \boldsymbol{P}_\pi,$$

donde $\boldsymbol{P}_\pi$ es la matriz de permutación matriz correspondiente a π, es decir, $\boldsymbol{P}_\pi \left(u_1, \ldots, u_n\right)^\top = \left(u_{\pi(1)}, \ldots, u_{\pi(k)}\right)^\top$, ρ es el parámetro de autocorrelación espacial, $\boldsymbol{F}_\pi = \text{diag}\left(\tau_{\pi(1)}, \tau_{\pi(2)}, \ldots, \tau_{\pi(k)}\right)$, $\boldsymbol{L}_\pi = \boldsymbol{I} - \boldsymbol{B}_\pi$ y $\boldsymbol{B}_\pi = \{b_{ij}\}$, tal que

$$b_{ij} = \frac{\rho}{1 + (n_{\pi_{(i)}} - 1)\rho^2}; \quad \tau_i = \frac{1 + (n_{\pi(i)} - 1)\rho^2}{1 - \rho^2},$$

donde $n_{\pi_{(i)}}$ representa la cardinalidad de $N(i)$, es decir, el número de vecinos de un área i, para $N(i) = \{j < i, j \sim i\}$.

Por otro lado, el parámetro ν_i modela la variabilidad restante no estructurada, la cual se modela usando una distribución normal, específicamente suponiendo independencia, se asume que $\nu_i \overset{ind}{\sim} \text{Normal}\left(0, \sigma_\nu^2 = 1/\tau_\nu\right)$.

Dependiendo de la estructura temporal en la ecuación (3.1) se definen dos modelos espacio-temporales como se detalla en la siguiente subsección.

3.1.2. Estructura temporal

Modelo espacio temporal DAGAR 1 (MDAGAR1)

Un modelo espacio temporal clásica fue introducido por Bernardinelli et al. (1995), y asume que el predictor lineal se puede definir como:

$$\begin{aligned}\eta_{it} = \log\left(\rho_{it}\right) &= \alpha + f_{si} + (\beta + \delta_i) \times t, \\ &= \alpha + v_i + \nu_i + (\beta + \delta_i) \times t,\end{aligned}$$

donde además de la estructura espacial, se incluye una tendencia lineal principal, β, que captura el efecto global del tiempo, y una tendencia diferencial, δ_i, que describe cómo interactúan el tiempo y el espacio. Para asegurar la identificabilidad del modelo, se impone una restricción de suma cero en $\boldsymbol{\delta}$ y $\boldsymbol{\nu}$. Además se asume que $\delta_i \overset{ind}{\sim} \text{Normal}\left(0, \tau_\delta\right)$.

Modelo espacio temporal DAGAR 2 (MDAGAR2)

Según lo señalado por Knorr-Held (2020) se propone una formulación dinámica no paramétrica para el predictor lineal:

$$\begin{aligned}\eta_{it} = \log(\rho_{it}) &= \alpha + f_{si} + \gamma_t + \phi_t, \\ &= \alpha + v_i + \nu_i + + \gamma_t + \phi_t,\end{aligned}$$

donde los parámetros α, v_i and ν_i tienen la mismas propiedades que en el Modelo 1. El parámetro γ_t representa un efecto temporal estructurado, que se modela dinámicamente mediante la estructura del tiempo previo y posterior, es decir un paseo aleatorio de orden uno:

$$\gamma_t \mid \gamma_{-t} \sim \mathrm{N}(\gamma_{t+1}, \tau_\gamma); \qquad \text{si} \quad t = 1$$

$$\gamma_t \mid \gamma_{-t} \sim \mathrm{N}\left(\frac{\gamma_{t-1} + \gamma_{t+1}}{2}, \frac{\tau_\gamma}{2}\right); \qquad \text{si} \quad t = 2, \ldots, T-1$$

$$\gamma_t \mid \gamma_{-t} \sim \mathrm{N}(\gamma_{t-1}, \tau_\gamma); \qquad \text{si} \quad t = T.$$

Bajo esta definición se tiene que $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T) \sim N(0, \boldsymbol{Q}_\gamma)$ donde,

$$\boldsymbol{Q}_\gamma = \tau_\gamma \begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \vdots & \vdots & \vdots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix}.$$

Respecto del parámetro ϕ_t este sigue una distribución a priori gausiana, $\phi_t \overset{ind}{\sim} \mathrm{N}(0, \tau_\phi)$.

3.2. Inferencia bayesiana usando INLA

El método de estimación INLA (Integrated Nested Laplace Approximation) propuesto por Rue et al. (2009), es una técnica que permite realizar inferencia bayesiana aproximada para modelos latentes gaussianos (LGM). Este enfoque se basa en una combinación de aproximaciones de Laplace anidadas de forma iterativa que facilitan la estimación eficiente de las distribuciones posteriores, superando las limitaciones computacionales de los métodos tradicionales como métodos de Monte Carlo vía cadenas de Markov (MCMC).

Un modelo Gaussiano Latente (LGM) es un modelo jerárquico compuesto por tres niveles, donde en el segundo nivel se incluyen las distribuciones a priori con distribución normal. En particular el modelo espacio temporal DAGAR propuesto es un LGM como se detalla a continuación:

- En el primer nivel, asumiendo que la y_{it} condicionadas al campo gaussiano latente $\mathbf{x}$ e hiperparámetros $\boldsymbol{\theta}$ son independientes, entonces la función de verosimilitud se define como:

$$\pi(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\theta}) = \prod_{t=1}^{T} \prod_{i=1}^{n} \pi\left(y_{i,t} \mid \mathbf{x}, \boldsymbol{\theta}\right).$$

- En el segundo nivel, se define el campo gausiano latente $\boldsymbol{x}$, de forma general para los modelos MDAGAR1 y MDAGAR2, puede estar compuesto por:

$$\boldsymbol{x} = (\alpha, \boldsymbol{v}, \nu_1, \ldots, \nu_n, \delta_1, \ldots, \delta_n, \boldsymbol{\gamma}, \phi_1, \ldots, \phi_T),$$

  el cual sigue una ditribución a priori multivariada gausiana:

$$\boldsymbol{x} \mid \boldsymbol{\theta} \sim N\left(0, \boldsymbol{Q}^{-1}(\boldsymbol{\theta})\right),$$

  donde:

$$\begin{aligned}
&\alpha \sim N\left(0, \tau_\alpha\right) \\
&\boldsymbol{v} \sim N\left(0, \boldsymbol{Q}^{-1}(\rho)/\tau_v\right), \\
&\nu_i \overset{ind}{\sim} N\left(0, \tau_\nu\right) \\
&\delta_i \overset{ind}{\sim} \text{Normal}\left(0, \tau_\delta\right) \\
&\boldsymbol{\gamma} \sim N(0, \boldsymbol{Q}_\gamma) \\
&\phi_t \overset{ind}{\sim} N\left(0, \tau_\phi\right).
\end{aligned}$$

- En el tercer nivel se asigna una distribución al resto de parámetros llamados hiperparámetros $\boldsymbol{\theta}$, cuya distribución no sea normal. Para los modelos MDAGAR1 y MDAGAR2, puede estar compuesto por:

$$\boldsymbol{\theta} = (\tau_v, \rho, \tau_\nu, \tau_\delta, \tau_\gamma, \tau_\phi),$$

  donde se asume que $\rho \sim U(0,1)$ y los demás hiperparámetros que son básicamente parámetros de precisión siguen una distribución gamma no informativa.

De esta forma, la función de densidad conjunta a posteriori de $\boldsymbol{x}$ y $\boldsymbol{\theta}$ es:

$$\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta}).$$

INLA calcula las distribuciones marginales posteriores para cada elemento del campo gaussiano latente $\boldsymbol{x}$, esto es

$$\pi\left(x_i \mid \boldsymbol{y}\right) = \int \pi\left(x_i, \boldsymbol{\theta} \mid \boldsymbol{y}\right) d\boldsymbol{\theta} = \int \pi\left(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}\right) \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}.$$

Además, la distribución marginal a posteriori del vector de hiperparámetros se define de la siguiente manera

$$\pi\left(\theta_m \mid \boldsymbol{y}\right) = \int \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) d\boldsymbol{\theta}_{-m},$$

donde $\boldsymbol{\theta}_{-m} = (\theta_1, \ldots, \theta_{m-1}, \theta_{m+1}, \ldots, \theta_k)$.

El método INLA aprovecha los supuestos del modelo para obtener una aproximación numérica de las distribuciones a posteriori de interés. En primer lugar, se tiene que:

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \boldsymbol{y}) &= \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \\ &= \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}, \boldsymbol{\theta})}{\pi(\boldsymbol{y})} \frac{1}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \\ &= \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{y})} \frac{1}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})} \\ &\propto \frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x} \mid, \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}. \end{aligned} \tag{3.2}$$

La ecuación (3.2) se aproxima utilizando el método de aproximación gaussiana en el denominador, y este procedimiento es equivalente a la aproximación de Laplace para $\pi(\boldsymbol{\theta} \mid \boldsymbol{y})$:

$$\widetilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) \approx \left.\frac{\pi(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}, \theta)\pi(\boldsymbol{\theta})}{\pi_G(\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y})}\right|_{x=x^*(\boldsymbol{\theta})} = \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}),$$

donde $\tilde{\pi_G}$ representa una aproximación gaussiana de la condicional completa de $\boldsymbol{x}$, y $x^*(\theta)$ es la moda de la distribución completa condicional de $\boldsymbol{x} \mid \boldsymbol{\theta}, \boldsymbol{y}$.

Siguiendo la misma idea se obtiene la aproximación de Laplace de $\pi\left(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}\right)$:

$$\begin{aligned} \pi\left(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}\right) &= \frac{\pi\left(x_i, \boldsymbol{x_{-i}} \mid \boldsymbol{\theta}, \boldsymbol{y}\right)}{\pi\left(x_i \mid x_i, \boldsymbol{\theta}, \boldsymbol{y}\right)}, \\ &= \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\pi(\boldsymbol{\theta} \mid \boldsymbol{y})} \frac{1}{\pi\left(\boldsymbol{x_{-i}} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y}\right)}, \\ &\propto \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\pi\left(\boldsymbol{x_{-i}} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y}\right)}, \end{aligned}$$

donde $\boldsymbol{x} = (x_i, \boldsymbol{x_{-i}})$, y $\boldsymbol{x_{-i}}$ se define como el vector $\boldsymbol{x}$ sin el i-ésimo elemento. Luego, el denominador $\pi(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})$ se aproxima mediante la aproximación gausiana:

$$\widetilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y}) \approx \left. \frac{\pi(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{y})}{\widetilde{\pi_G}(\boldsymbol{x_{-i}} \mid x_i, \boldsymbol{\theta}, \boldsymbol{y})} \right|_{x_{-i}=x^*_{(x_{-i},\boldsymbol{\theta})}},$$

donde $x^*_{(x_{-i},\theta)}$ es la moda de la distribución condicional completa de x_{-i} la cuál se obtiene mediante expansiones de Taylor.

Las distribuciones marginales a posteriori para los hiperparámetros $\tilde{\pi}(\boldsymbol{\theta_m} \mid \mathbf{y})$ se obtienen a partir de $\tilde{\pi}(\boldsymbol{\theta} \mid \mathbf{y})$ mediante integración numérica. Del mismo modo, se aproxima $\pi(x_i \mid \boldsymbol{\theta}, \mathbf{y})$

$$\tilde{\pi}(x_i \mid \boldsymbol{y}) = \int \tilde{\pi}(x_i \mid \boldsymbol{\theta}, \boldsymbol{y})\, \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\theta.$$

$$\tilde{\pi}(\theta_m \mid \boldsymbol{y}) = \int \tilde{\pi}(\boldsymbol{\theta} \mid \boldsymbol{y}) d\theta_{-m}.$$

Para más detalles sobre el INLA ver Rue et al. (2009).

Capítulo 4

Estudio de Simulación

En este capítulo se presenta un estudio de simulación con el objetivo de evaluar la correcta estimación de los parámetros de los modelos MDAGAR1 y MDAGAR2, descritos en la Sección 3.1 a través de la implementación a realizar en el INLA.

Para la simulación de los datos, se consideraron $n = 43$ áreas geográficas correspondientes a los distritos de la provincia de Lima. Se construyó un grafo que define una matriz de vecindad ($\boldsymbol{W}$) basada en estas áreas. En la Figura 4.2 se muestra que los cuadrados amarillos representan distritos no vecinos (con valores de $w_{ij} = 0$), mientras que los cuadrados rojos indican distritos vecinos (con valores de $w_{ij} = 1$). Esta matriz de vecindad evidencia que cada distrito solo está conectada con un conjunto limitado de distritos vecinos.

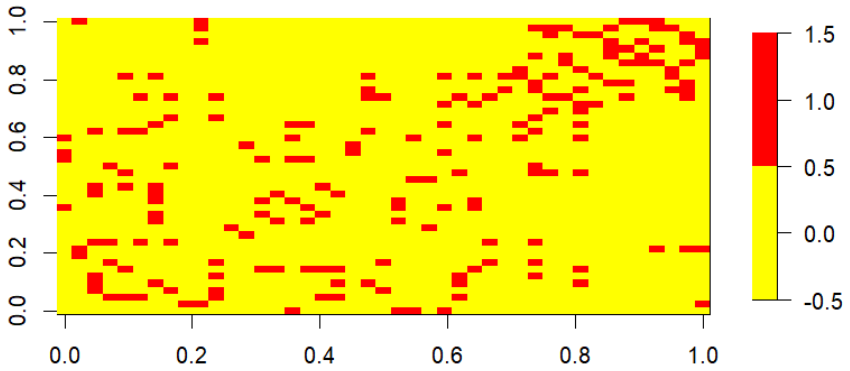

Figura 4.1: Matriz de vecindad de los distritos de Lima.

El proceso de simulación abarca tanto datos espaciales como temporales, generados para las $n = 43$ áreas a lo largo de $T = 11$ periodos de tiempo. Se utilizaron dos componentes espaciales (uno estructurado y otro no estructurado), además de una componente temporal local.

El procedimiento para simular los efectos aleatorios espaciales se describe a continuación:

- **Componente espacial estructurado**: Los efectos espaciales estructurados se modelan según lo señalado en la sección 3.1.1, para diferentes valores de $\rho = 0.2, 0.4, 0.6, 0.8$. La matriz de precisión $\boldsymbol{Q}(\rho)$ se calcula con un hiperparámetro de precisión $\tau_v = 6$. Los efectos espaciales v_i se generan mediante la descomposición de Cholesky de $\boldsymbol{Q}^{-1}(\rho)$, donde $v_i = \boldsymbol{z} \times \boldsymbol{L}$, siendo $\boldsymbol{z}$ un vector de variables normales estandarizadas, y $\boldsymbol{L}$ la matriz resultante de la descomposición.

- **Componente espacial no estructurado**: Se introducen efectos espaciales no estructurados, generados a partir de una distribución normal independiente con media cero y una precisión τ_ν. Estos efectos ν_i se suman a los efectos estructurados para formar el componente espacial total $\zeta_i = v_i + \nu_i$.

Los efectos temporales se pueden simular de acuerdo a los modelos MDAGAR1 y MDAGAR2, este procedimiento se describe en las siguientes secciones, respectivamente.

4.1. Modelo MDAGAR1

- En este modelo se asumió una precisión $\tau_\nu = 150$.

- **Componente temporal local**: Para capturar variaciones temporales a nivel local, se genera un efecto aleatorio $\delta_i \sim \mathcal{N}(0, \sigma_\delta^2)$, para $i = 1, \ldots, n$, donde el hiperparámetro de precisión $\tau_\delta = 1/\sigma_\delta^2 = 1000$.

- **Predictor lineal**: El predictor lineal del modelo se define como:

$$\eta_{it} = \alpha + \zeta_i + (\beta + \delta_i) \times t,$$

  donde $\alpha = -0.45$ es el intercepto, $\beta = 1$ es el parámetro global temporal, los tiempos $t = 1, \ldots, T$ y δ_i sirve para modelar las fluctuaciones temporales locales.

- **Generación de datos**: Finalmente, la variable de respuesta y_{it} se genera siguiendo una distribución de Poisson con media $\mu_{it} = E_i \times \rho_{it}$, donde E_i es un factor de exposición y $\rho_{it} = \exp(\eta_{it})$ es la tasa de ocurrencia en cada área i y en cada tiempo t.

Para evaluar la capacidad del Modelo MDAGAR1 de recuperar los valores reales de los parámetros utilizados en la simulación, se realizaron 30 réplicas, para cada uno de los diferentes escenarios, es decir para los diferentes valores del parámetro de autocorrelación

espacial $\rho = 0.2, 0.4, 0.6, 0.8$. En cada réplica, se generaron datos espacio-temporales bajo el modelo MDAGAR1, con el resto de valores de los parámetros ya descritos.

A manera de ejemplo la Figura 4.2 presenta para una réplica la evolución espacial de los valores simulados de y_{it} en los $n = 43$ distritos de Lima en $T = 11$ tiempos, obtenidos mediante el Modelo MDAGAR1. Se observan patrones diferenciados en los distritos, destacando regiones con mayor intensidad conforme avanza el periodo analizado.

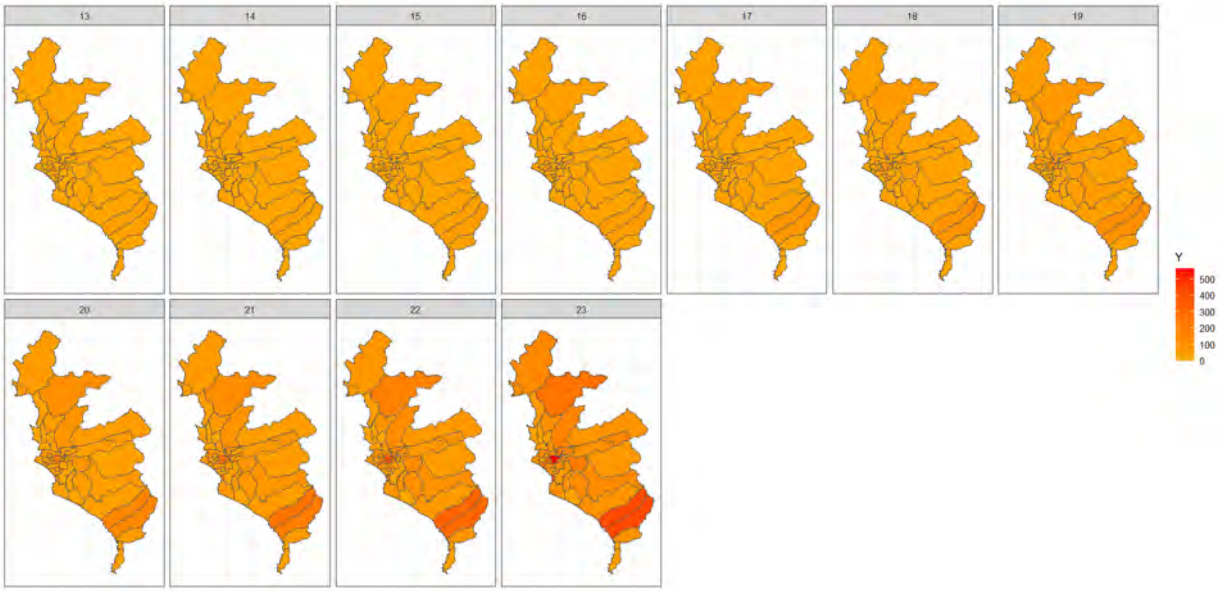

Figura 4.2: Mapas de valores y_{it} simulados para los distritos de Lima.

Posteriormente se estimaron los parámetros mediante inferencia bayesiana usando INLA. La inferencia bayesiana para el modelo MDAGAR1 fue implementada a través del método INLA. A partir de los resultados de cada ajuste, se obtuvieron las medias a posteriori de los parámetros de interés. Las Figuras 4.3 y 4.4 presentan diagramas de caja que muestran la distribución empírica de las medias a posteriori de los parámetros de autocorrelación espacial ρ y precisión espacial τ_v, respectivamente, según las estimaciones a lo largo de las 30 réplicas por escenario. En general, los diagramas de cajas muestran que las medianas de las estimaciones del parámetro ρ se aproximan consistentemente al valor verdadero simulado para los escenarios $\rho = 0.2, 0.4, 0.6$, y que las cajas contienen el verdadero valor simulado, lo que indica que el modelo es capaz de recuperar adecuadamente distintos grados de nivel de autocorrelación espacial. Cuando el valor de ρ aumenta, aunque disminuye la dispersión de las estimaciones, sugiriendo estabilidad y precisión en la inferencia de este parámetro, sin embargo también se tiende a subestimar el parámetro ligeramente como se observa cuando $\rho = 0.8$. De manera similar, las estimaciones del parámetro de precisión espacial τ_v se en-

cuentran, en su mayoría, dentro del rango intercuartílico esperado cuando ρ no es muy alto, lo que evidencia una buena capacidad del modelo para capturar la variabilidad espacial estructurada cuando la autocorrelación espacial es pequeña o moderada. Por otro lado, cuando el valor de ρ aumenta, se tiende a sobreeestimar ligeramente este parámetro.

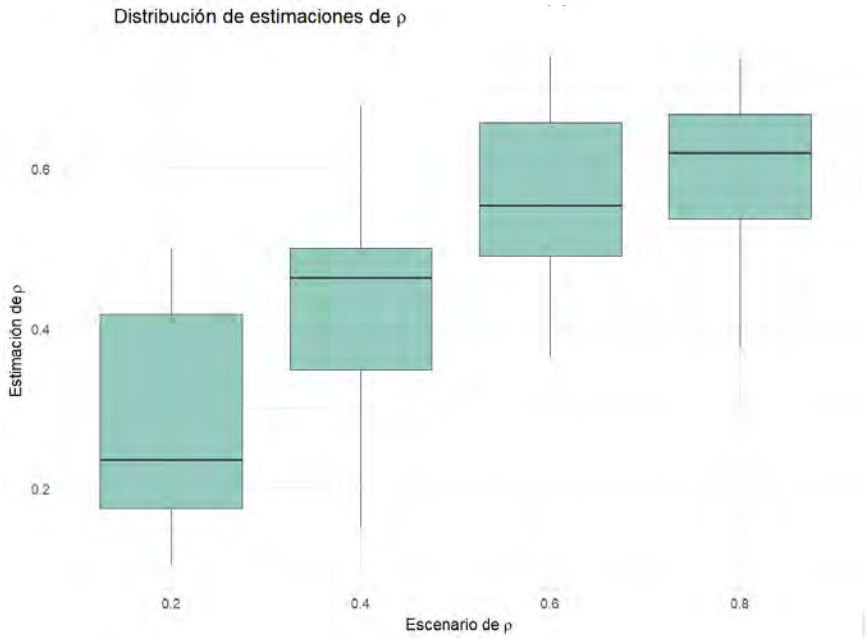

Figura 4.3: Diagramas de caja de estimaciones medias a posteriori de ρ en 30 réplicas por escenario (Modelo MDAGAR1).

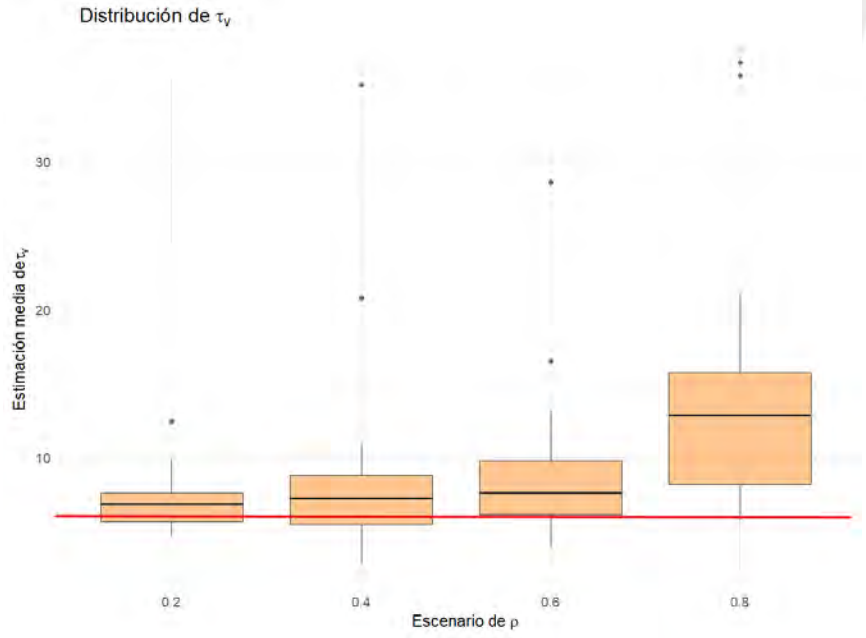

Figura 4.4: Diagramas de caja de estimaciones medias a posteriori de τ_{DAGAR} en 30 réplicas por escenario (Modelo MDAGAR1). La línea roja indica el valor real del parámetro τ_v utilizado en la simulación.

A continuación se presentan resultados adicionales para una de las réplicas. La Tabla 4.1 resume las estimaciones obtenidas para el modelo MDAGAR1. Se incluyen tanto los valores originales como las estimaciones de la media a posteriori, desviación estándar a posteriori, y el intervalos de credibilidad (IC al 95 %) de los parámetros, el IC al 95 % es definido por los cuantiles a posteriori al 2.5 % y 97.5 %. Respecto a la estimación de los parámetros, el parámetro β es el que se estima con mayor precisión y menor sesgo en todos los escenarios, pues la media a posteriori es muy cercana al valor original (1.0) y los intervalos de credibilidad al 95 %(IC) contienen dicho valor, con desviaciones estándar menores a 0.01. Lo mismo ocurre con el intercepto, aunque con un poco más de variabilidad, especialmente en los escenarios extremos $\rho = 0.2, 0.8$.

Respecto a los parámetros de autocorrelación y precisión espacial:

- Según este resultado, el parámetro ρ es estimado mejor cuando el $\rho = 0.2$, pues el IC contiene este valor. El parámetro ρ tiende a subestimarse en el resto de los escenarios, aunque las medias a posteriori siguen una tendencia creciente coherente con los valores simulados.

- El parámetro τ_v presenta mejor desempeño cuando ρ es moderado a alto. En los escenarios 3 y 4, las medias están más próximas al valor original (6.0) y el intervalo de credibilidad contiene dicho valor. En cambio, con $\rho = 0.2$ hay mayor sesgo y mayor dispersión relativa.

En cuanto a los efectos aleatorios no estructurados y temporales:

- El parámetro τ_ν muestra alta variabilidad en todos los escenarios, aunque las medias se acercan más al valor original a partir de $\rho = 0.4$. Aun así, los intervalos de credibilidad son amplios, reflejando incertidumbre considerable.

- El parámetro τ_δ es el parámetro más difícil de estimar. En el escenario $\rho = 0.2$ la media se sobreestima significativamente, mientras que en $\rho = 0.4$ se subestima fuertemente. Solo en $\rho = 0.6y0.8$ las estimaciones se acercan más al valor verdadero, aunque siguen mostrando alta varianza. Esto indica que este parámetro es sensible al nivel de autocorrelación espacial y que requiere mayor cuidado para una estimación más precisa.

Cuadro 4.1: Resultados de las estimaciones a posteriori según el modelo MDAGAR1 para distintos valores de ρ.

	Original	Media	Desv.est.	$Q_{2.5}$	$Q_{97.5}$
Escenario 1 ($\rho = 0.2$)					
α	-0.45	-0.616	0.115	-0.842	-0.382
β	1.00	0.984	0.009	0.966	1.002
τ_ν	150.00	110.680	135.787	5.949	468.486
τ_δ	1000.00	11438.890	39179.370	391.413	73166.342
τ_v	6.00	5.344	1.737	1.737	5.344
ρ	0.20	0.373	0.139	0.139	0.373
Escenario 2 ($\rho = 0.4$)					
α	-0.45	-0.305	0.094	-0.493	-0.120
β	1.00	0.993	0.010	0.974	1.013
τ_ν	150.00	163.752	410.197	2.147	993.922
τ_δ	1000.00	91.766	26.515	50.197	153.655
τ_v	6.00	8.114	3.094	3.094	8.114
ρ	0.40	0.341	0.172	0.172	0.341
Escenario 3 ($\rho = 0.6$)					
α	-0.45	-0.412	0.082	-0.575	-0.248
β	1.00	1.008	0.008	0.991	1.024
τ_ν	150.00	138.841	149.239	13.637	532.566
τ_δ	1000.00	1091.742	867.336	267.369	3405.437
τ_v	6.00	11.722	4.590	4.590	11.722
ρ	0.60	0.462	0.150	0.150	0.462
Escenario 4 ($\rho = 0.8$)					
α	-0.45	-0.369	0.117	-0.603	-0.125
β	1.00	0.994	0.009	0.977	1.011
τ_ν	150.00	136.016	168.365	14.864	569.334
τ_δ	1000.00	426.200	207.482	162.615	957.216
τ_v	6.00	9.117	5.309	5.309	9.117
ρ	0.80	0.680	0.157	0.157	0.680

En la Figura 4.5 y 4.6, se presentan las distribuciones marginales de ρ y τ_v, respectivamente, para los cuatro escenarios considerados: $\rho = 0.2$, $\rho = 0.4$, $\rho = 0.6$ y $\rho = 0.8$. Cada panel muestra la distribución marginal correspondiente, una la línea roja que indica el valor original del parámetro y las lineas azules que indican el intervalo de credibilidad al 95 %. Estos resultados permiten observar cómo varían las estimaciones del parámetro bajo diferentes supuestos de autocorrelación espacial.

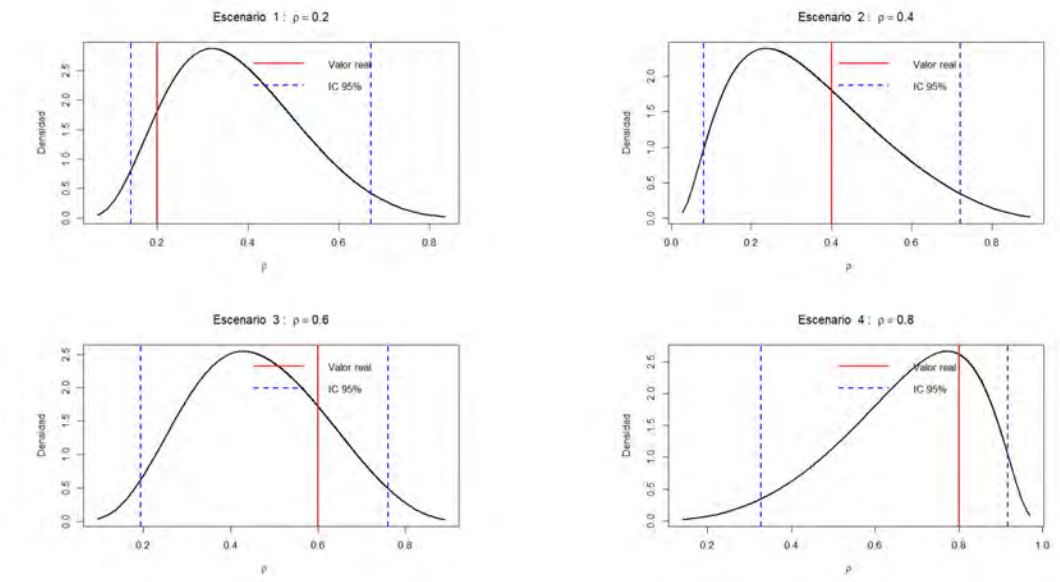

Figura 4.5: Distribución marginal de ρ bajo diferentes valores. La línea roja que indica el valor original del parámetro ρ y las lineas azules que indican el intervalo de credibilidad al 95 %.

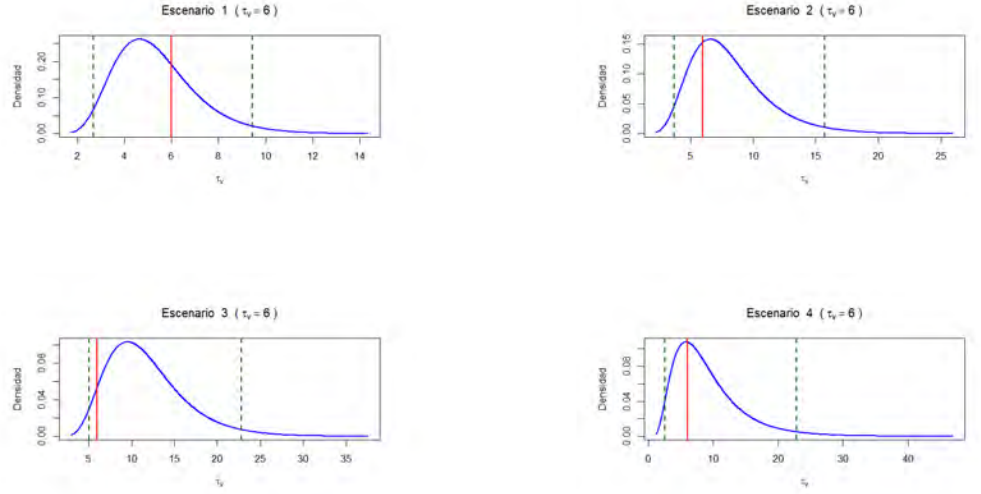

Figura 4.6: Distribuciones marginales de τ_v para los diferentes valores de ρ. La línea roja que indica el valor original del parámetro τ_v y las lineas azules que indican el intervalo de credibilidad al 95 %.

En la Figura 4.7, se presentan las gráficas que comparan los valores simulados y las estimaciones de la tasa ρ_{it} para cada distrito en el tiempo t en cada uno de los escenarios. Cada panel muestra la relación entre los valores simulados de ρ_{it} y las estimaciones a posteriori obtenidas $\hat{\rho}_{it} = E(\rho_{it}|\boldsymbol{y})$. La línea azul que representa la relación ideal, donde ambos valores coinciden perfectamente. Así se observa que las estimaciones a posteriori de ρ_{it} se acercan

mucho a los valores originales simulados, mostrando una buena bondad de ajuste del modelo MDAGAR1.

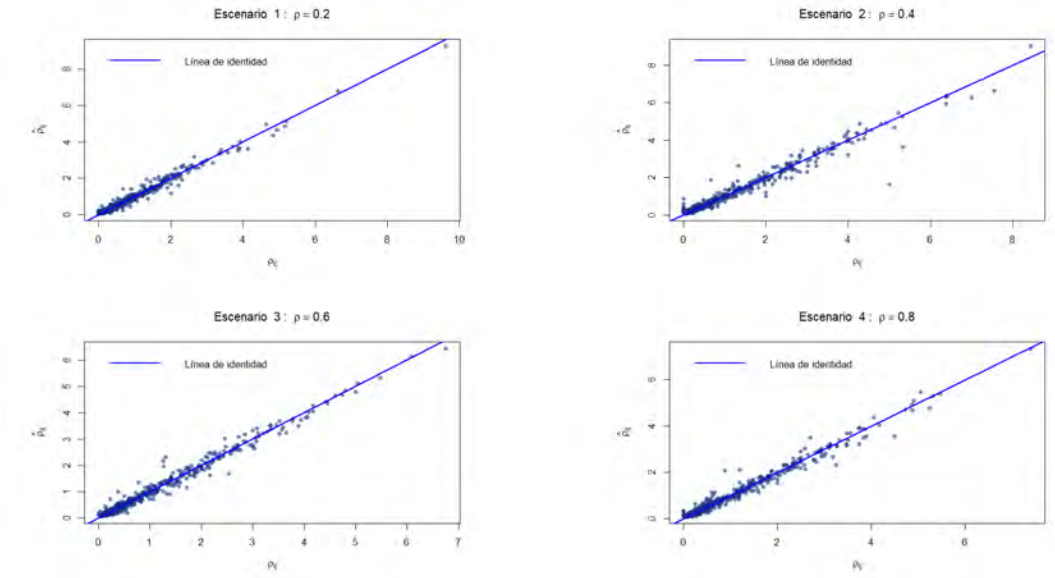

Figura 4.7: Diagramas de dispersión del valor de ρ_{it} versus la estimación a posteriori de ρ_{it} para diferentes valores de ρ.

4.2. Modelo MDAGAR2

El Modelo DAGAR 2 se enfoca en una estructura espacial más compleja, incorporando un modelo de paseo aleatorio (random walk) de primer orden para capturar la dinámica espacial.

- En este modelo se asumió una precisión $\tau_\nu = 45$.
- **Componente temporal local no estructurado**: Se simulan los efectos temporales no estructurados $\phi_t \overset{ind}{\sim} \mathrm{N}(0, \tau_\phi)$ con un hiperparámetro de precisión $\tau_\phi = 5$.
- **Componente temporal local estructurado**: Se simulan los efectos temporales estructurados como un proceso de paseo aleatorio de primer orden $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T) \sim N(0, \boldsymbol{Q}_\gamma)$ donde, $\boldsymbol{Q}_\gamma$ depende del hiperparámetro de precisión $\tau_\gamma = 10$. Este efecto aleatorio permite capturar la dinámica temporal de manera más efectiva.
- **Predictor lineal**: El predictor lineal se define como:

$$\eta_{it} = \alpha + v_i + \nu_i + \gamma_t + \phi_t, \tag{4.1}$$

donde $\alpha = -1$ es el intercepto global.

- **Generación de datos**: Finalmente, la variable de respuesta y_{it} se genera siguiendo una distribución de Poisson con media $\mu_{it} = E_i \times \rho_{it}$, donde E_i es un factor de exposición y $\rho_{it} = \exp(\eta_{it})$ es la tasa de ocurrencia en cada área i y en cada tiempo t.

Para evaluar la capacidad del Modelo MDAGAR2 de recuperar los valores reales de los parámetros utilizados en la simulación, se realizaron 30 réplicas, para cada uno de los diferentes escenarios, es decir para los diferentes valores del parámetro de autocorrelación espacial $\rho = 0.2, 0.4, 0.6, 0.8$. En cada réplica, se generaron datos espacio-temporales bajo el modelo MDAGAR2, con el resto de valores de los parámetros ya descritos.

A manera de ejemplo la Figura 4.8 presenta para una réplica la evolución espacial de los valores simulados de y_{it} en los $n = 43$ distritos de Lima en $T = 11$ tiempos, obtenidos mediante el Modelo MDAGAR2.

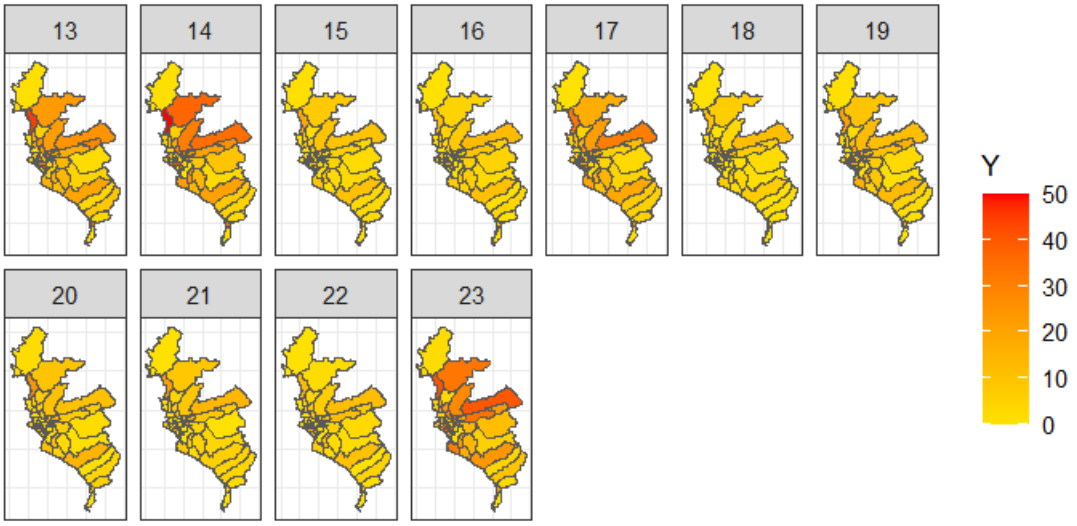

Figura 4.8: Mapas de valores simulados para los distritos de Lima

Posteriormente se estimaron los parámetros mediante inferencia bayesiana usando INLA. La inferencia bayesiana para el modelo MDAGAR2 fue implementada a través del método INLA. A partir de los resultados de cada ajuste, se obtuvieron las medias a posteriori de los parámetros de interés. Las Figuras 4.9 y 4.10 presentan diagramas de caja que muestran la

distribución empírica de las medias a posteriori de los parámetros de autocorrelación espacial ρ y precisión espacial τ_v, respectivamente, según las estimaciones a lo largo de las 30 réplicas por escenario. En la Figura 4.9 se observa que la mediana de las estimaciones del parámetro ρ es muy similar a los valores verdaderos en los escenarios $\rho = 0.2$ y $\rho = 0.4$, lo que indica una buena recuperación del parámetro en condiciones de baja y moderada autocorrelación espacial. En los escenarios con mayor autocorrelación ($\rho = 0.6$ y 0.8), se evidencia una subestimación, siendo mayor la subestimación cuando ρ es mayor, aunque las estimaciones mantienen la tendencia creciente esperada.

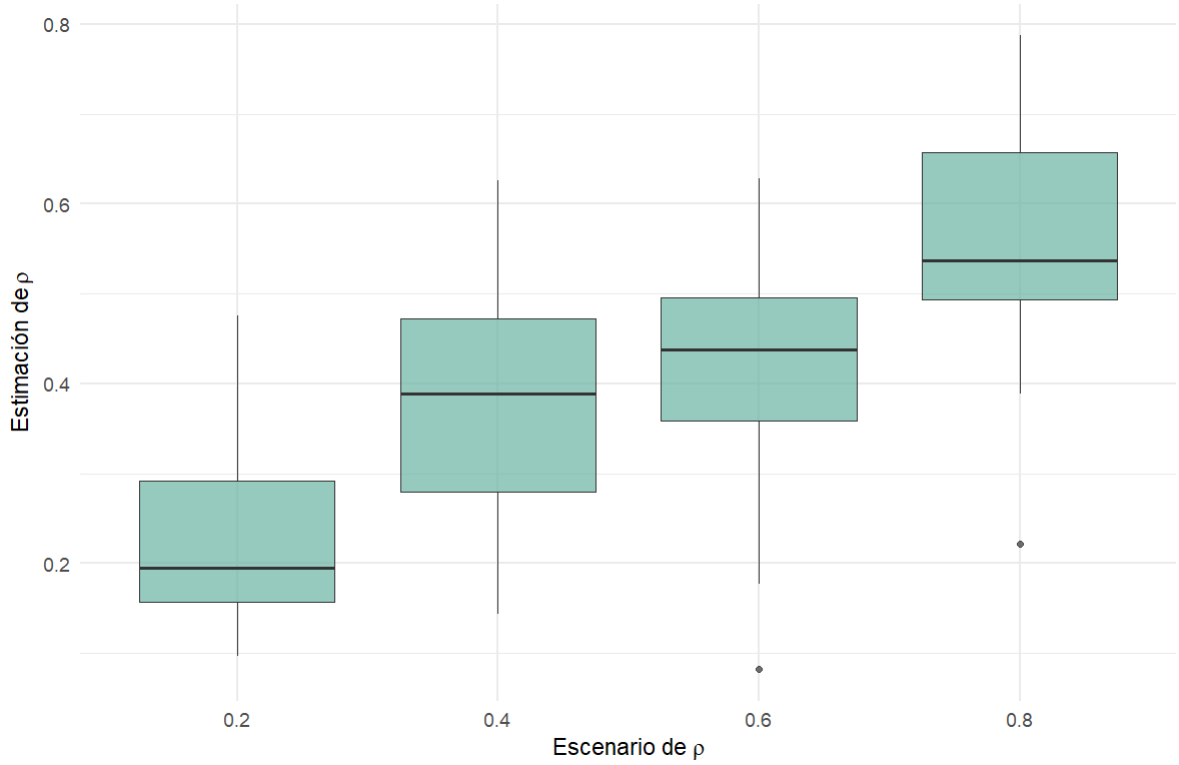

Figura 4.9: Distribución de estimaciones medias posteriores de ρ en 30 réplicas por escenario (Modelo MDAGAR2).

Por otro lado, la Figura 4.10 muestra que las estimaciones del parámetro de precisión espacial τ_v son consistentes en los escenarios $\rho = 0.2$, 0.4 y 0.6, a pesar del incremento en la variabilidad conforme aumenta el nivel de autocorrelación. En el escenario $\rho = 0.8$, el valor verdadero se encuentra ligeramente fuera del rango intercuartílico, lo que sugiere una leve pérdida de precisión en contextos de alta dependencia espacial.

A continuación se presentan resultados adicionales para una de las réplicas. El cuadro 4.2 muestra tanto los valores originales de los parámetros como las estimaciones a posteriori de la media, desviación estándar y los intervalos de credibilidad al 95 %.

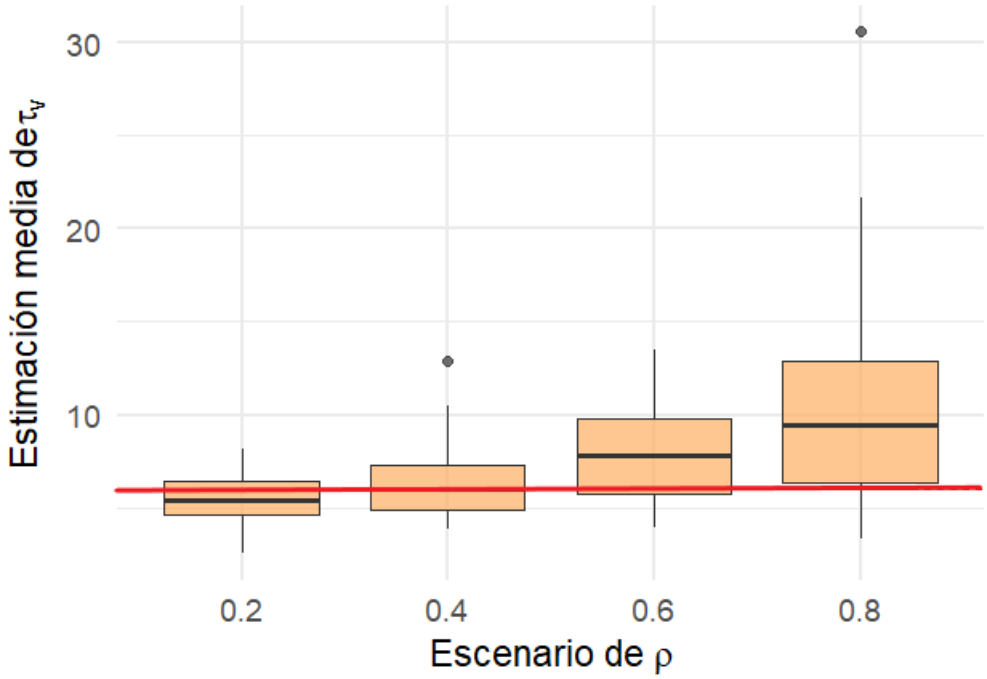

Figura 4.10: Diagramas de caja de estimaciones medias de τ_v (máximo 10,000) en 30 réplicas por escenario (Modelo MDAGAR2). La línea roja discontinua que indica el valor real del parámetro τ_{DAGAR} utilizado en la simulación.

En general, el modelo MDAGAR2 estima con mayor precisión los parámetros en escenarios de baja a moderada autocorrelación espacial ($\rho = 0.2$ y 0.4). Los parámetros más difíciles de estimar son las precisiones τ_γ y τ del efecto aleatorio DAGAR, especialmente cuando ρ es alto, mientras que τ_ϕ y el intercepto se estiman de forma más robusta en todos los escenarios.

En particular, el intercepto α se estima con baja variabilidad y tendencia a la subestimación en todos los escenarios, pero manteniendo una distancia constante respecto del valor original. Para el parámetro ρ, se observa que todos los intervalos de credibilidad contienen el valor verdadero de ρ. En particular, se observa una buena estimación en el escenario $\rho = 0.4$, donde la media estimada (0.420) es muy cercana al valor original y el intervalo de credibilidad contiene claramente el valor verdadero. En el escenario $\rho = 0.2$, hay una ligera subestimación, mientras que en los escenarios $\rho = 0.6$ y 0.8 la subestimación es más notoria, especialmente en $\rho = 0.6$, donde la media es 0.296. Esto indica que el modelo estima mejor la autocorrelación espacial en niveles bajos a moderados, y presenta mayor dificultad cuando ρ es alto. Respecto al parámetro τ_v, el IC contiene el verdadero valor de τ_v cuando $\rho = 0.2, 0.4, 0.6$. En particular, se obtienen resultados razonables en los escenarios $\rho = 0.2$ y 0.4, con medias cercanas al valor original y bajos niveles de variabilidad relativa. Sin embargo, en el escenario $\rho = 0.6$ la media es significativamente mayor y la desviación estándar es alta, lo que sugiere sobreestimación. Este comportamiento se agrava en el escenario $\rho = 0.8$, donde se observa

Cuadro 4.2: Resultados de las estimaciones a posteriori según el modelo MDAGAR2 para distintos valores de ρ.

	Original	Media	Desv.est.	$Q_{2.5}$	$Q_{97.5}$
Escenario 1 ($\rho = 0.2$)					
α	-1	-1.771	0.080	-1.938	-1.617
τ_ν	45	97.848	134.513	3.149	448.621
τ_γ	10	197.894	1448.865	1.146	1339.481
τ_ϕ	5	4.543	2.924	1.079	12.085
τ_v	6	7.901	2.404	4.279	13.647
ρ	0.2	0.144	0.101	0.020	0.400
Escenario 2 ($\rho = 0.4$)					
α	-1	-1.684	0.114	-1.917	-1.462
τ_ν	45	125.876	131.880	11.084	475.655
τ_γ	10	78.822	278.327	1.661	509.184
τ_ϕ	5	7.049	4.752	1.586	19.398
τ_v	6	6.046	2.265	2.742	11.532
ρ	0.4	0.420	0.150	0.158	0.727
Escenario 3 ($\rho = 0.6$)					
α	-1	-1.815	0.066	-1.949	-1.684
τ_ν	45	183.255	4640.059	2.950	1103.737
τ_γ	10	118.741	693.661	1.181	809.959
τ_ϕ	5	4.482	3.019	1.006	12.321
τ_v	6	16.461	8.393	5.840	38.050
ρ	0.6	0.296	0.161	0.060	0.662
Escenario 4 ($\rho = 0.8$)					
α	-1	-1.840	0.021	-1.882	-1.798
τ_ν	45	21.947	6.278	12.057	36.558
τ_γ	10	114.776	655.366	1.136	782.782
τ_ϕ	5	4.897	3.405	1.029	13.756
τ_v	6	21793	23153	1400	84786
ρ	0.8	0.500	0.261	0.059	0.940

una media extremadamente elevada, indicando que la estimación de este parámetro se ve fuertemente afectada en contextos de alta autocorrelación espacial. Respecto a los parámetros asociados a los efectos aleatorios temporales, el parámetro τ_ϕ (componente estructurado) se estima consistentemente bien en todos los escenarios, con medias cercanas a los valores originales y variabilidad moderada, con respecto a τ_γ, todos los IC contienen el valor original, aunque su estimación presenta gran variabilidad y valores medios muy por encima del valor real, especialmente en los escenarios con mayor autocorrelación. Finalmente, en cuanto a τ_ν asociado a la variabilidad no estructurada a nivel de áreas, la variabilidad en la estimación es alta en la mayoría de escenarios, especialmente en $\rho = 0.6$, con una desviación estándar superior a 4600. En contraste, en $\rho = 0.8$ se observa una estimación más precisa (media 21.947), pero alejada del valor original (45), lo que evidencia cierta inestabilidad general en

la estimación de este parámetro.

En la Figura 4.11, se presentan las distribuciones marginales del hiperparámetro de precisión τ_v para diferentes valores del parámetro de autocorrelación espacial ρ =0.2, 0.4, 0.6, 0.8. Cada panel muestra la función de densidad marginal,la línea roja indica el valor original del hiperparámetro. Estos resultados permiten observar cómo varía la estimación de τ_v según los diferentes supuestos de dependencia espacial.

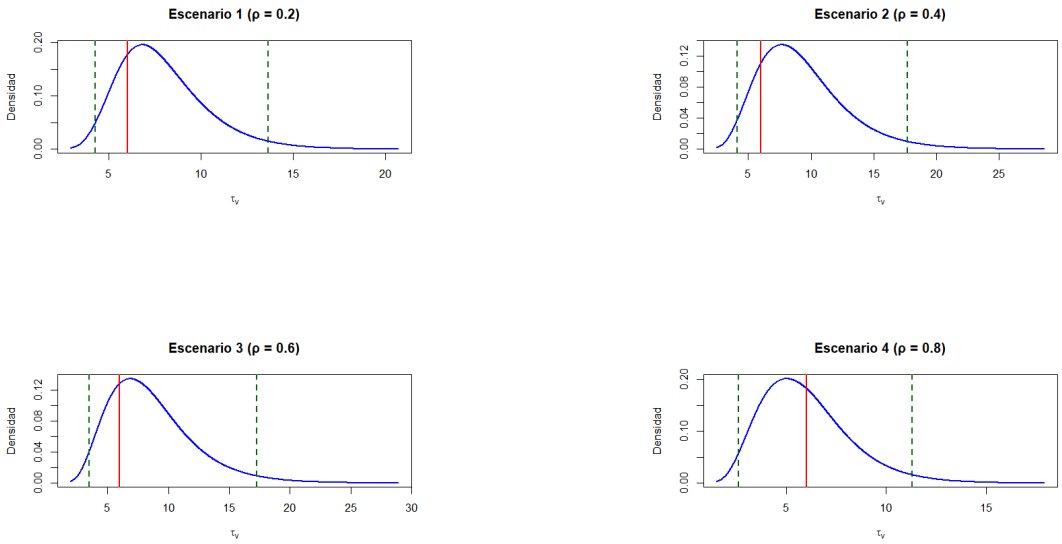

Figura 4.11: Distribuciones marginales del hiperparámetro de precisión τ_v para diferentes valores de ρ.

En la Figura 4.12, se presentan las distribuciones marginales del parámetro de correlación espacial ρ para los cuatro valores simulados: 0.2, 0.4, 0.6 y 0.8. Cada panel ilustra la densidad de la distribución marginal correspondiente, con una línea roja que indica el valor original del parámetro ρ. Esta visualización permite observar cómo varían las estimaciones de ρ bajo diferentes condiciones de dependencia espacial, proporcionando información valiosa sobre la sensibilidad de las estimaciones a los cambios en el parámetro ρ.

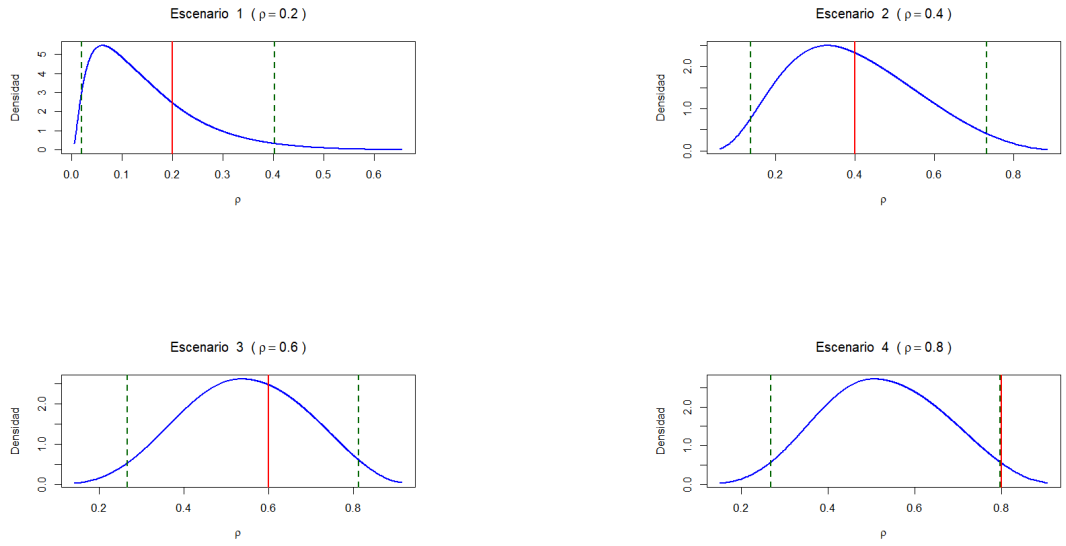

Figura 4.12: Distribuciones marginales de ρ para diferentes valores del parámetro de correlación espacial. Cada gráfico muestra la densidad de la distribución marginal correspondiente, con una línea roja que indica el valor original del parámetro ρ.

En la Figura 4.13, se presentan los gráficos comparativos de los valores observados de ρ_{it} y las estimaciones ajustadas $\hat{\rho}_{it} = E(\rho_{it}|\boldsymbol{y})$ para el modelo MDAGAR2 bajo diferentes valores del parámetro de autocorrelación espacial ρ= 0.2, 0.4, 0.6, 0.8. Cada panel muestra la relación entre los valores observados y las estimaciones de la tasa de ocurrencia, la línea azul indica el ajuste perfecto. Se observa que el ajuste entre los valores observados y las estimaciones de ρ_{it} mejora conforme aumenta el nivel de autocorrelación espacial. En el escenario 2 ($\rho = 0.4$) se presenta la menor dispersión de puntos alrededor de la línea de ajuste perfecto, lo que sugiere una mejor precisión en las estimaciones para este caso.

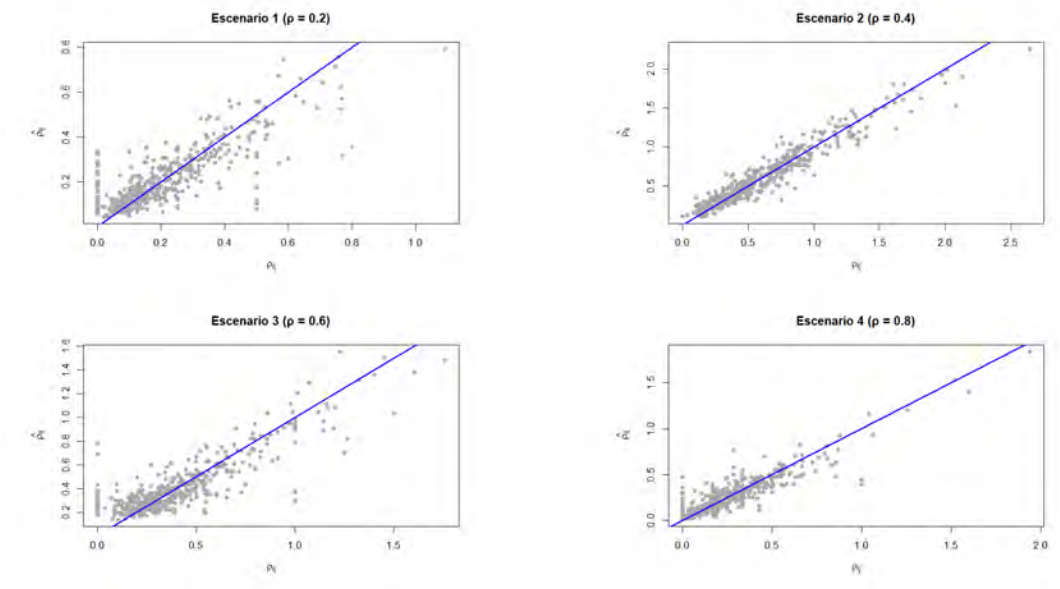

Figura 4.13: Diagramas de dispersión del valor de ρ_{it} versus la estimación a posteriori de ρ_{it} para diferentes valores de ρ.

Capítulo 5

Aplicación

El presente capítulo tiene como objetivo aplicar los modelos MDAGAR1 y MDAGAR2 a un conjunto de datos reales, específicamente para estimar el número de Reportes de Operaciones Sospechosas (ROS) registrados en los distritos de la provincia de Lima. Los ROS son indicadores clave en la detección y prevención del lavado de activos, siendo de gran importancia para identificar patrones y áreas con mayor incidencia.

Los datos utilizados corresponden al número de ROS reportados en $n = 43$ distritos de Lima durante el periodo 2013 - 2023. Este análisis estadístico tiene como propósito explorar la distribución espacial de los ROS y evaluar la influencia de factores regionales en su aparición. La Figura 5.1 muestra la distribución espacial de los Reportes de Operaciones Sospechosas (ROS) en las distritos de Lima, durante los años 2013 a 2023.

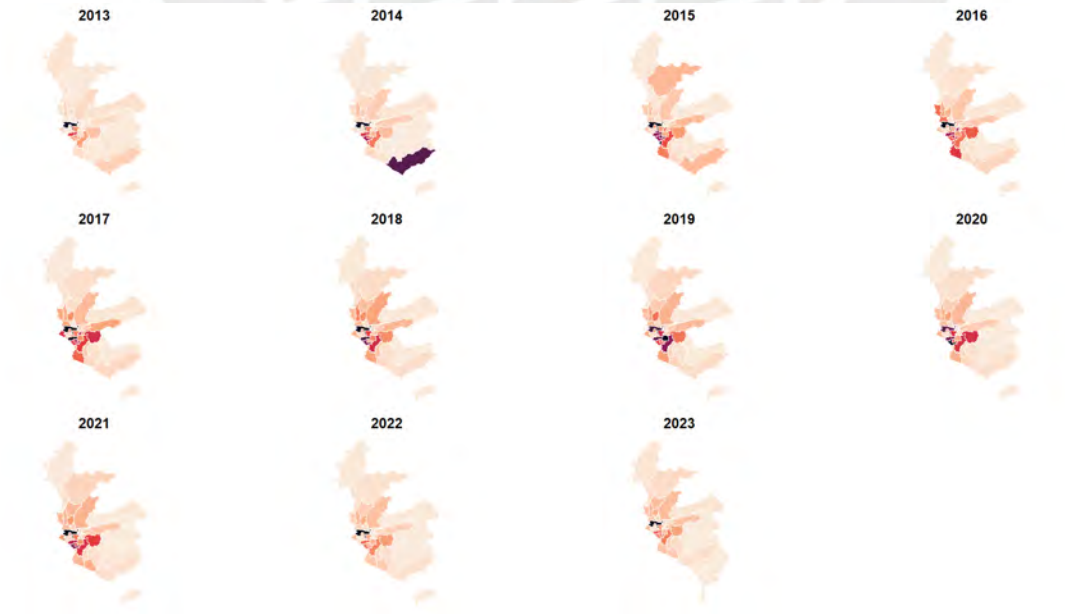

Figura 5.1: Datos reales de ROS por distrito (2013-2023)

Se observa una heterogeneidad espacial marcada, donde ciertos distritos destacan con tasas altas en años específicos, lo que podría reflejar diferencias en los patrones de actividad financiera, la eficacia de los sistemas de detección o posibles focos de operaciones sospechosas. Estas variaciones también pueden estar influenciadas por cambios normativos o económicos en el periodo analizado.

En general, se asume que y_{it} representa el número de Reportes de Operaciones Sospechosas (ROS) en el i-ésima distrito durante el t-ésimo período de tiempo, donde $i = 1, \ldots, 43$ y $t = 1, \ldots, 11$. Consideramos que $y_{it} \sim \text{Poisson}(\mu_{it} = E_{it}\rho_{it})$, donde μ_{it} representa la media de ROS en el distrito i y el período t, E_{it} se define como el factor de exposición, esto es, el número de operaciones de alto riesgo, para el distrito i en el período t, y ρ_{it} se define como la tasa de incidencia de ROS en el distrito i durante el período t.

Se ajustaron los modelos MDAGAR1 y MDAGAR2 usando INLA.

Adicionalmente, se ajustaron modelos espacio temporales con las estructuras temporales de los modelos MDAGAR1 y MDAGAR2, pero en vez de usar el efecto espacial estructurado DAGAR, se ajustó un efecto espacial estructurado condicional autoregresivo (CAR), a estos modelos los llamamos MCAR1 y MCAR2. Luego se comparó el ajuste de los modelos mediante el criterio del Watanabe-Akaike Information Criterion (WAIC) y devianza (DIC). Los resultados se muestran en el cuadro 5.1.

Cuadro 5.1: Comparación de los criterios WAIC y DIC entre los modelos ajustados.

Modelo	**WAIC**	**DIC**
MDAGAR1	36,992.39	-564.34
MDAGAR2	2,187.29	2,185.78
MCAR1	37,025.15	-564.88

El criterio WAIC se utiliza para comparar modelos en términos de ajuste y complejidad, donde valores menores indican un mejor balance entre ambos aspectos. En este estudio, el modelo MDAGAR1 presenta un WAIC de 36,992.39, que resulta menos favorable que el del modelo MDAGAR2 (2,187.29). No obstante, al considerar el modelo MCAR1 con un WAIC de 37,025.15, se observa que el modelo MDAGAR1 obtiene un valor levemente inferior, lo que sugiere un ajuste comparable o ligeramente superior. En términos del DIC, el modelo MDAGAR1 muestra un valor más bajo (-564.34) frente al modelo MDAGAR2 (2185.78) y el modelo MCAR (-564.88), reafirmando su mejor desempeño relativo al balance entre ajuste y penalización por complejidad en su contexto de aplicación.

El cuadro 5.2 muestra un resumen con las estimaciones a posteriori obtenidas mediante inferencia bayesiana con INLA, específicamente presenta los valores estimados a posteriori

de la media, desviación estándar y los intervalos de credibilidad al 95 % (Q025, Q975) de los parámetros de los modelos ajustados.

Cuadro 5.2: Estimaciones a posteriori de los parámetros de los modelos ajustados.

	Parámetro	**Media**	**Desv. Est.**	$Q_{2.5}$	$Q_{97.5}$
MDAGAR1					
	Intercepto	-5.377	0.699	-6.714	-3.883
	β	0.324	0.024	0.277	0.372
	τ_ν	110.763	121.033	7.519	432.132
	τ_δ	6.645	1.759	3.782	10.656
	ρ_{DAGAR}	0.602	0.160	0.602	0.160
	τ_{DAGAR}	0.200	0.108	0.200	0.108
MDAGAR2					
	Intercepto	-1.858	0.077	-2.014	-1.706
	τ_ν	128.037	163.077	8.062	551.955
	τ_γ	127.731	755.581	1.110	872.051
	τ_ϕ	4.932	3.375	1.068	13.701
	τ_{DAGAR}	13.908	6.146	5.557	29.296
	ρ_{DAGAR}	0.401	0.165	0.124	0.743
MCAR1					
	Intercepto	-6.024	0.064	-6.155	-5.904
	β	0.334	0.025	0.286	0.385
	τ_ν	0.358	0.100	0.202	0.592
	τ_δ	6.470	1.799	3.602	10.625

Según los resultados del modelo MDAGAR1, el intercepto estimado es $\alpha = -5.377$, y $\beta = 0.324$, mostrando un efecto positivo del tiempo. Para los hiperparámetros, $\tau_\nu = 110.763$ indica alta variabilidad en los efectos espaciales no estructurados, mientras que $\tau_\delta = 6.645$ refleja mayor estabilidad en los efectos temporales. Según el modelo MDAGAR2, el valor de τ_{dagar} (13.908) indica una precisión considerable en los efectos estructurados, mientras que el valor de ρ_{dagar} (0.401) refleja una autocorrelación espacial positiva media entre áreas vecinas Los intervalos de credibilidad de todos los parámetros se ajustan dentro de rangos razonables, lo que valida la robustez del modelo.

La Figura 5.2 presenta las distribuciones a posteriori de los parámetros de precisión espacial τ_v y de autocorrelación espacial ρ del modelo MDAGAR1, estimados mediante el enfoque bayesiano. Estas distribuciones reflejan la incertidumbre asociada a dichos parámetros y permiten evaluar el grado de estructura espacial capturada por el modelo.

La Figura 5.3 presenta las distribuciones a posteriori de los parámetros de precisión espacial τ_v y de autocorrelación espacial ρ del modelo MDAGAR2, estimados mediante el enfoque bayesiano.

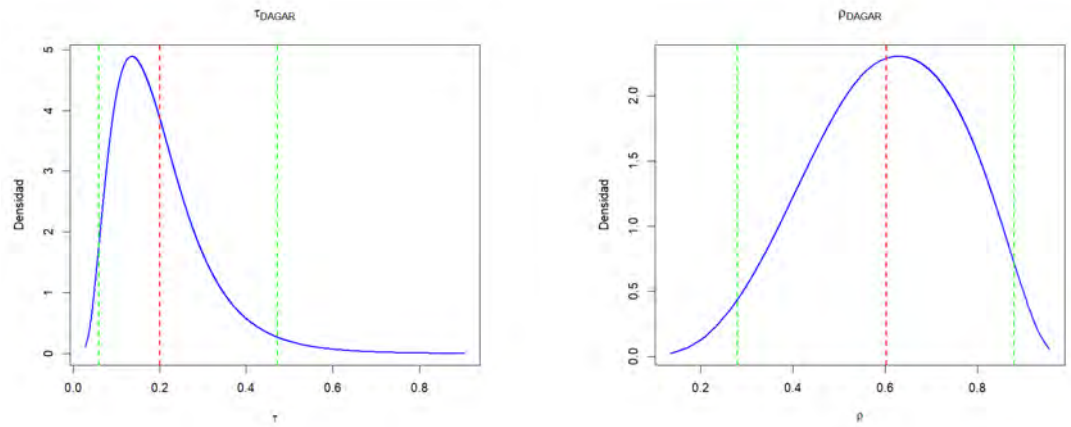

Figura 5.2: Distribución marginal a posteriori de los parámetros τ_v y ρ según el modelo MDAGAR1.

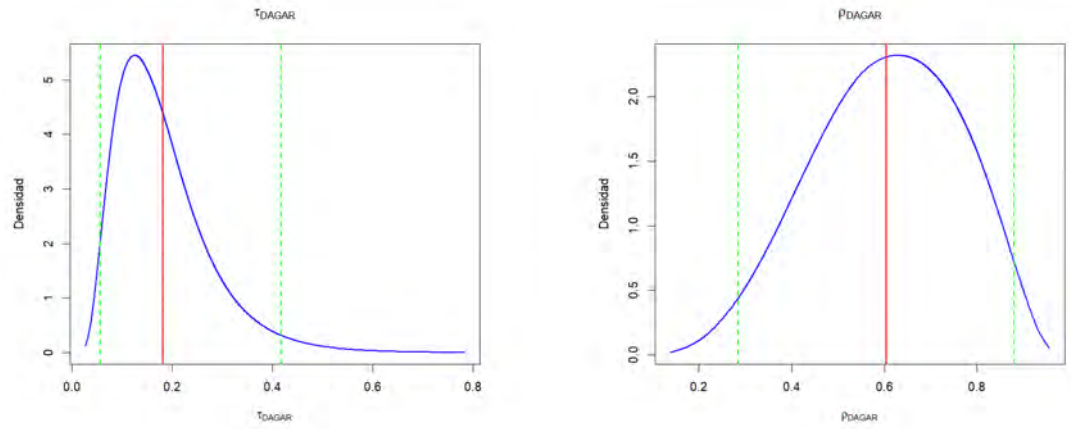

Figura 5.3: Distribución a posteriori de los parámetros τ_v y ρ a través del modelo MDAGAR2.

La Figura 5.4 presenta la comparación entre las tasas de incidencia observadas ($\frac{Y}{E}$) y las estimadas ($\hat{\rho}_{ij}$) desagregada por año a través del modelo MDAGAR1. Cada panel muestra la relación para un año específico entre 2013 y 2023. La línea azul representando la relación perfecta entre ambas. La mayoría de los puntos se agrupan cerca de la línea horizontal, indicando que el modelo estimó correctamente las tasas en la mayoría de las áreas, especialmente en aquellas con tasas bajas.

La Figura 5.5 presenta la comparación entre las tasas de incidencia observadas ($\frac{Y}{E}$) y las estimadas ($\hat{\rho}_{ij}$) desagregada por año según el modelo MDAGAR2. Cada panel muestra la

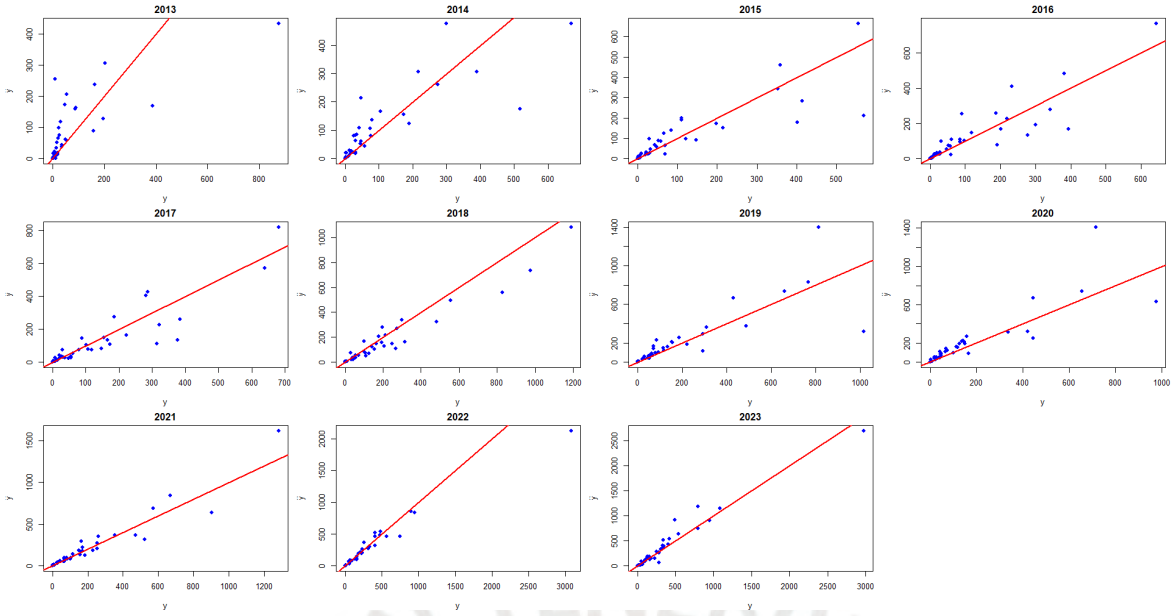

Figura 5.4: Valores estimados por el modelo MDAGAR1 del número de ROS por distrito vs ROS real para los años 2013-2023.

relación para un año específico entre 2013 y 2023, con los puntos azules representando las observaciones y la línea roja la tendencia ajustada.

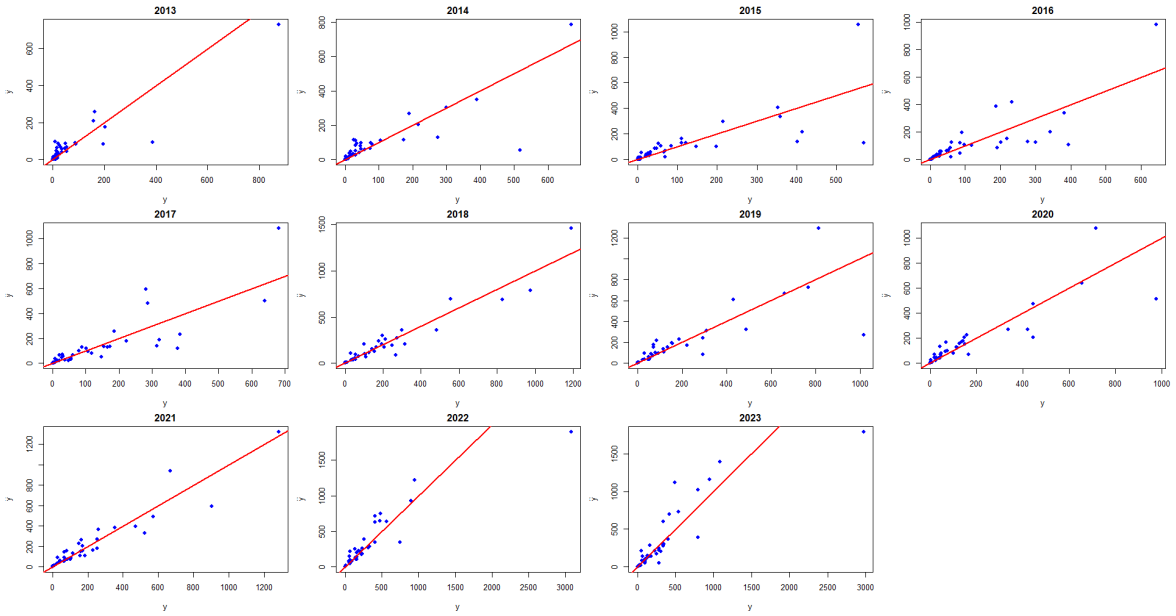

Figura 5.5: Valores estimados por el modelo MDAGAR2 del número de ROS por distrito vs ROS real para los años 2013-2023.

La Figura 5.6 presenta los mapas de las medias a posteriori de los efectos espaciales y espacio-temporales de los ROS para cada uno de los años, desde 2013 hasta 2023 según el modelo MDAGAR1. Estos mapas visualizan la distribución geoespacial y la evolución temporal de los riesgos estimados en los diferentes distritos de Lima.

Figura 5.6: Incidencia de ROS estimada a través del modelo MDAGAR1.

Finalmente, la Figura 5.7 muestra los mapas de las medias a posteriori de los efectos espaciales y espacio-temporales de los ROS para cada uno de los años, desde 2013 hasta 2023 bajo el modelo MDAGAR2.

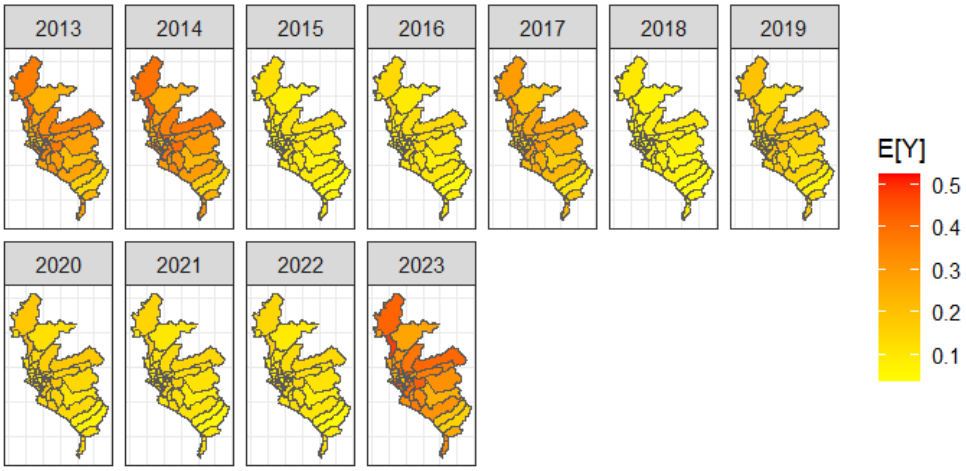

Figura 5.7: Incidencia de ROS estimada a través del modelo MDAGAR2.

Capítulo 6

Conclusiones

El objetivo principal de esta investigación fue estimar la incidencia y el patrón espacial de las Operaciones Sospechosas (ROS) en los distritos de Lima, incorporando la estructura espacial y temporal de los datos mediante modelos DAGAR (Directed Acyclic Graph Autoregressive) y empleando el enfoque bayesiano con el método de integración aproximada anidada de Laplace (INLA). Se evaluaron dos modelos: uno con una estructura temporal local y otro con una dinámica temporal de paseo aleatorio, demostrando la capacidad de estos enfoques para capturar tanto la heterogeneidad espacial como las tendencias temporales, con tiempos de cómputo notablemente bajos en comparación con métodos clásicos.

En cuanto a los resultados, se observó una correlación positiva entre las tasas de ROS estimadas y reales, evidenciando la capacidad de los modelos para ajustar los datos observados. Además, se identificaron patrones espaciales significativos en las tasas de ROS, mostrando una moderada autocorrelación espacial entre distritos vecinos. El análisis también permitió identificar factores explicativos significativos para la incidencia de ROS, lo que aporta información relevante para la toma de decisiones en la prevención del lavado de activos.

Este estudio podría ampliarse con la inclusión de más covariables y factores que ayuden a explicar mejor la incidencia de ROS. Asimismo, el modelo planteado puede extenderse para analizar datos espacio-temporales con mayor granularidad, como registros mensuales o anuales, aprovechando la eficiencia computacional de INLA para trabajar con conjuntos de datos más grandes y complejos. La metodología presentada también puede adaptarse a otras áreas de estudio que requieran análisis espacio-temporales robustos.

Bibliografía

Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*, Chapman and Hall/CRC.

Bernardinelli, L., Clayton, D., Pasutto, C., Montomoli, C., Ghislandi, M. and Songini, M. (1995). Bayesian analysis of space-time variation in disease risk, *Stat Med* **14**(2433-43).

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*, Wiley, Chichester.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems, *Journal of the Royal Statistical Society: Series B (Methodological)* **36**(2): 192–225.

Blangiardo, M., Cameletti, M., Baio, G. and Rue, H. (2013). Spatial and spatio-temporal models with r-inla, *Spatial and spatio-temporal epidemiology* **4**: 33–49.

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley Classics Library.

Datta, A., Banerjee, S., Hodges, J. S. and Gao, L. (2019). Spatial disease mapping using directed acyclic graph auto-regressive (DAGAR) models, *Bayesian Analysis* **14**(4): 1221–1244.
URL: *https://doi.org/10.1214/19-BA1177*

Ferwerda, J., Van Saase, A. and Getzner, M. (2020). Estimating money laundering flows with a gravity model-based simulation, *Scientific Reports* **10**(18552).

Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010). *Handbook of Spatial Statistics*, Chapman and Hall/CRC.

Haining, R. (2003). *Spatial Data Analysis: Theory and Practice*, Cambridge University Press.

Knorr-Held, L. (2020). Bayesian modelling of inseparable space-time variation in disease risk, *Statistics in medicine* **19**(17-18).

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, Springer.

Rue, H., Martino, S. and 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.

SBS (2018). Resumen de las evaluaciones sectoriales de exposición a los riesgos de Lavado de Activos y Financiamiento del Terrorismo de los sectores minero, pesquero y maderero del Perú – UIF-Perú con apoyo de la Cooperación Alemana (GIZ) – 2018, *Technical report*, Superintendencia de Banca, Seguros y AFP (SBS).
URL: *https://www.sbs.gob.pe/prevencion-de-lavado-activos/Estudios-Tecnicos/Estudios-de-Analisis-de-Riesgos*

SBS (2021). EVALUACIÓN NACIONAL DE RIESGOS DE LAVADO DE ACTIVOS, *Technical report*, Superintendencia de Banca, Seguros y AFP (SBS).
URL: *https://www.sbs.gob.pe/prevencion-de-lavado-activos/Estudios-Tecnicos/Estudios-de-Analisis-de-Riesgos*

SBS (2023). Información Estadística Georreferenciada UIF-Perú, *Technical report*, Superintendencia de Banca, Seguros y AFP (SBS).
URL: *https://www.sbs.gob.pe/prevencion-de-lavado-activos/Estudios-Tecnicos/Estudios-Estrategicos*

Singh, K. and Best, P. (2019). Anti-money laundering: Using data visualization to identify suspicious activity, *International Journal of Accounting Information Systems* **34**(100418).

Whittle, P. (1954). On stationary processes in the plane, *Biometrika* **41**(3-4): 434–449.
URL: *https://doi.org/10.1093/biomet/41.3-4.434*

Xia, P., Zhiwei, N., Xiao, H., Zhu, X. and Peng, P. (2022). A novel spatiotemporal prediction approach based on graph convolution neural networks and long short-term memory for money laundering fraud, *Arabian Journal for Science and Engineering* **47**: 1921–1937.

Yulia, A., Astutikand, S. and Sa'adah, U. (2021). Modeling spatial variation of money laundering crime in indonesia using geographically weighted multinomial logistic regression, *IOP Conference Series: Materials Science and Engineering* **1115**(012065).