**PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ**

**ESCUELA DE POSGRADO**


# Modelo espacial para estudiar la distribución del monto del gasto devengado de la inversión pública a nivel provincial en el Perú

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

**Kendy Brigitte Ocola Aguero**

Asesora:

**Zaida Jesús Quiroz Cornejo**

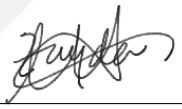
Lima, 2024

**Informe de Similitud**

Yo Zaida Jesús Quiroz Cornejo docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesora de la tesis titulada *Modelo espacial para estudiar la distribución del monto del gasto devengado de la inversión pública a nivel provincial en el Perú*, de la autora Kendy Brigitte Ocola Aguero, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 20 %. Así lo consigna el reporte de similitud emitido por el software Turnitin el 12/08/2024.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lima, 12 de agosto de 2024


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

# Dedicatoria

Esta tesis está dedicada a mis padres, Vilma y Percy, cuyo amor, paciencia y dedicación me han permitido alcanzar hoy otro sueño más.

A mis hermanos, gracias por estar conmigo durante este proceso, brindándome consejos y palabras de aliento que me han hecho mejorar como persona y que de una manera u otra han estado presentes en todos mis sueños y metas.

# Agradecimentos

Quiero expresar mi sincera gratitud a la Pontificia Universidad Católica del Perú y a mis profesores, especialmente a la Dra. Zaida Quiróz, cuya enseñanza y conocimientos valiosos me han permitido evolucionar como profesional día tras día. Agradezco profundamente a cada uno de ustedes por su paciencia, dedicación y apoyo.

# Resumen

Esta tesis presenta una exploración detallada y técnica de los modelos espaciales autorregresivos condicionales (CAR) y autorregresivos simultáneos (SAR) para analizar los datos de inversión pública del año 2022, específicamente para estudiar la distribución espacial del monto del gasto devengado de inversión pública en Perú. A través de una combinación de análisis teóricos y simulaciones, la investigación establece metodologías para evaluar cómo variables como la corrupción, los niveles de inversión del gobierno local, cartera priorizada y avance físico de la inversión, influencian el gasto devengado en diferentes provincias. Este estudio contribuye significativamente al entendimiento de la distribución espacial del gasto público y los factores que lo afectan, utilizando técnicas estadísticas avanzadas para mejorar la precisión y eficacia de las estimaciones de los modelos utilizados. Los resultados del análisis ofrecen perspectivas críticas sobre la gestión y asignación de fondos públicos, proporcionando una herramienta valiosa para los planificadores y responsables de la formulación de políticas públicas.

**Palabras-clave:** datos de áreas, estadística espacial, inversión pública, modelo CAR, modelo SAR, monto devengado.

# Abstract

This thesis presents a detailed and technical exploration of the Conditional Autoregressive (CAR) and Simultaneous Autoregressive (SAR) spatial models to analyze the data on public investment from the year 2022, specifically to study the spatial distribution of accrued expenditure in public investment in Peru. Through a combination of theoretical analysis and simulations, the research establishes methodologies to evaluate how variables such as corruption, levels of local government investment, prioritized portfolio, and physical progress of the investment influence the accrued expenditure in different provinces. This study significantly contributes to the understanding of the spatial distribution of public spending and the factors that affect it, using advanced statistical techniques to improve the accuracy and effectiveness of the estimates of the models used. The results of the analysis provide critical perspectives on the management and allocation of public funds, offering a valuable tool for planners and policymakers in public policy formulation.

**Keywords:** accrued amount, areal data, CAR model, public investment, SAR model, spatial statistics.

# Indice general

# Capítulo 1

# Introducción

## 1.1. Consideraciones preliminares

Según el Boletín anual de ejecución de la inversión pública 2022 elaborado por el Ministerio de Economía y Finanzas (MEF) el año 2022, se consideró como un año importante para el cierre de brechas (Pichihua Serna, 2022). En efecto el conjunto el MEF y las Unidades Ejecutoras de Inversiones del Sistema Nacional de Programación Multianual de Inversiones (SNPMGI) se ejecutó S/ 46,592 millones en los tres niveles de gobierno. Esto refleja un impacto significativo para el país si consideramos que dicha cifra representa un 20 % más que el 2021 (último registro). Asimismo, se llegó a ejecutar el 72 % del presupuesto total asignado para inversiones en dicho año, siendo el monto más alto desde el año 2018. Este logro se debe en gran parte a una mayor asignación de recursos presupuestales, así como a la estrategia de acompañamiento intensivo y continuo liderada por el MEF durante los últimos años destinada a fortalecer los equipos técnicos, principalmente en las regiones, para contribuir a mejorar la gestión de la inversión pública en el país.

Las entidades del SNPMGI de los tres niveles de gobierno tienen la tarea de identificar las principales necesidades y carencias en infraestructura y servicios básicos que tiene la población y de disponer el presupuesto necesario para la ejecución de inversiones públicas, con el objetivo de contribuir al cierre de brechas. El Gobierno Nacional ha ejecutado en el 2022 S/ 16,841 millones, siendo los sectores con mayor porcentaje de avance en la ejecución: Desarrollo e Inclusión Social (99.7 %), Junta Nacional de Justicia (98.4 %), Presidencia del Consejo de Ministros (98.2 %), Defensoría del Pueblo (98 %) y, Transportes y Comunicaciones (96 %). Los gobiernos regionales han ejecutado de manera acumulada S/ 9,247 millones, siendo los de mayor ejecución: el Gobierno Regional de Apurímac (90.8 %), Gobierno Regional de Tacna (87.8 %), Gobierno Regional de Loreto (87 %), Gobierno Regional de Junín (84.3 %),

y el Gobierno Regional de Cusco (82.3 %). Los gobiernos locales de manera acumulada han ejecutado S/ 20,504 millones, siendo los departamentos con mayor ejecución: Loreto (75.8 %), Apurímac (73.4 %), la Provincia Constitucional del Callao (72.9 %), Tacna (72.7 %), y Cusco (71.5 %)

En la literatura existen estudios que analizaron la importancia de la inversión pública en la deisigualdad económica (Costa-i Fon y Rodriguez-Oreggia, 2005), en el crecimiento económico regional (Rodríguez-Pose et al., 2012), asi como trabajos de investigación de los principales factores que explicarían el comportamiento en el monto ejecutado en las inversiones públicas en el Perú, Jimenez et al. (2020) y Lastra (2017).

Costa-i Fon y Rodriguez-Oreggia (2005) investiga la contribución de la inversión pública a la reducción de las desigualdades regionales en México. Examina el impacto de la inversión pública según la posición de cada región en la distribución condicional del ingreso regional utilizando la regresión por cuantiles como técnica empírica. Los resultados confirman la hipótesis de que las desigualdades regionales sí pueden atribuirse a la distribución regional de la inversión pública; el patrón observado muestra que la inversión pública ayudó principalmente a reducir las desigualdades regionales entre las regiones más ricas.

Rodríguez-Pose et al. (2012) estimaron el impacto que tiene la inversión pública en el crecimiento económico regional en Grecia. Utilizaron una base de datos de gasto público por región del año 1978 al año 2007. Ellos usaron un modelo que captura el impacto de la inversión pública en las prefecturas griegas y los efectos indirectos relacionados con las regiones vecinas.

Según el estudio realizado por Lastra (2017), indican haber identificado y cuantificado qué factores están asociados a la ejecución de la inversión pública de los gobiernos locales en el Perú, en particular por categoría de municipalidad y con especial énfasis en la infraestructura básica, es decir, en salud, saneamiento, educación, agropecuario, transporte y electrificación. Para ello usaron una data panel a nivel de 1834 municipalidades para entre los años 2008 y 2014.

Jimenez et al. (2020) estimaron un modelo de datos de panel dinámico para 1796 gobiernos locales del Perú entre los años 2010 y 2018 en el análisis del comportamiento de la inversión pública. Encontraron que los factores más relevantes para impulsar la inversión pública local citan las siguientes: “la disponibilidad de fuentes de financiamiento, en especial las asociadas a ingresos de recursos naturales no renovables; las variables asociadas a la capacidad de cada gobierno local tanto para planificar, presupuestar como ejecutar la inversión pública; y finalmente los efectos del ciclo presupuestario político, sobretodo durante el año siguiente

a las elecciones locales". Además, se preocupan por analizar la diferencia entre gobiernos locales con autoridades reelectas y no reelectas, así como agrupando a los gobiernos locales según su tamaño económico.

En este contexto, en esta tesis se propone estudiar cómo ha evolucionado el promedio del monto devengado de ejecución de las inversiones públicas en las provincias del Perú. Cabe mencionar que el gasto devengado hace mención al reconocimiento de una obligación de pago derivado del gasto comprometido previamente registrado [1]. Para ello, es importante entender primero las características de la variable de interés. Además un modelo estadístico apropiado tomará en cuenta qué otras variables explican la variación del monto promedio de ejecución en las regiones.

Formalmente, se asume que la variable dependiente $Y_i$ representa el promedio de monto devengado en inversiones públicas en la provincia i-ésima para $i = 1, \ldots, n$, donde $n$ es el número de provincias en Perú. Primero, se plantea modelar $Y_i$ tomando en cuenta que el monto de ejecución promedio en una provincia puede depender esencialmente de los montos de ejecución promedio de las provincias vecinas, es decir puede haber evidencia de autocorrelación espacial. En este sentido se propone ajustar estos datos usando los modelos espaciales para datos de áreas más aplicados son el modelo condicional autoregresivo (CAR) y el modelo autoregresivo simultáneo (SAR).

Debido a la distribución normal multivariada de la variable dependiente es viable usar inferencia clásica para estimar los parámetros de los modelos ajustandos usando estimación por máxima verosimilitud. Aunque otros métodos de estimación para modelos de datos de áreas también son viables tales como los métodos de inferencia bayesiana (Banerjee et al., 2014; Rue y Held, 2005).

## 1.2. Objetivos de la tesis

El objetivo general de la tesis es realizar una aplicación del modelos espaciales para analizar datos de las inversiones públicas del Sistema Nacional de Programación Multianual y Gestión de Inversiones a nivel provincial en el Perú. De manera específica:

- Revisar la literatura acerca de los diferentes propuestas de modelos CAR y SAR.
- Proponer, estudiar propiedades, e implementar la estimación de los modelos CAR y SAR con la variable dependiente gaussiana desde la perspectiva clásica.

[1] Se formaliza a través de la conformidad del área correspondiente en la entidad pública o Unidad Ejecutora que corresponda respecto de la recepción satisfactoria de los bienes y la prestación de los servicios solicitados y se registra sobre la base de la respectiva documentación sustentatoria.

- Implementar la estimación de los parámetros a través de inferencia clásica.
- Realizar estudios de simulación acerca del modelo CAR y SAR.
- Aplicar el modelo a un conjunto de datos reales, específicamente para estimar el monto devengado en inversión pública a nivel provincial en el Perú.

## 1.3. Organización del trabajo

En el capítulo 2 se presentan los conceptos que nos permitirá estudiar el modelo CAR así como también el modelo SAR. En el capítulo 3 se aborda la inferencia clásica de ambos modelos. El capítulo 4 muestra los resultados de las simulaciones realizadas. Finalmente, en el capítulo 5 se presentan los resultados de la aplicación de los modelos espaciales a los datos de inversión pública en el Perú para el año 2022.

# Capítulo 2

# Conceptos y modelos

Los modelos estadísticos para analizar datos espaciales que se recolectan en una red (posiblemente irregular) son llamados modelos para datos de área. Estos modelos espaciales para datos de áreas se utilizan en muchos campos, incluyendo la cartografía de las tasas de infecciones, (Elliott y Wartenberg, 2004), agricultura (Besag y Higdon, 1999), econometría (LeSage y Thomas-Agnan, 2015), ecología (Arslan y Akyürek, 2018), entre otras.

Varios autores han propuesto modelos para datos de áreas, mostrando las características de las correlaciones dada una estructura espacial (Besag, 1986; Wall, 2004; Assunção y Krainski, 2009). En este capítulo se presenta una revisión de modelos para datos de áreas.

## 2.1. Introducción a datos de áreas

Según Banerjee et al. (2014) los datos pueden considerarse como una realización (parcial) de un fenómeno aleatorio (es decir, un proceso estocástico) $\{Y(s) : s \in D\}$, en el caso de datos de áreas el conjunto de índices $D$ es una colección contable de áreas en las que se observan datos. La colección $D$ de dichas áreas se denomina lattice, que luego se complementa con información del vecindario. Matemáticamente, los centroides de las áreas se convierten en vértices, que están conectados por aristas (esta es la estructura de vecindad). El formalismo se basa en la teoría de grafos y se brinda una discusión al respecto.

### 2.1.1. Lattices

El lattice de sitios que indexa $Y(.)$ se obtiene (ya sea implícita o explícitamente) con información del vecindario sobre las áreas. Aquí "lattice"se refiere a una colección contable de áreas (espaciales), las cuales pueden ser espacialmente regulares o irregulares. La Figura 2.1 muestra un mapa de los 67 condados de Pensilvania. La numeración de los condados está

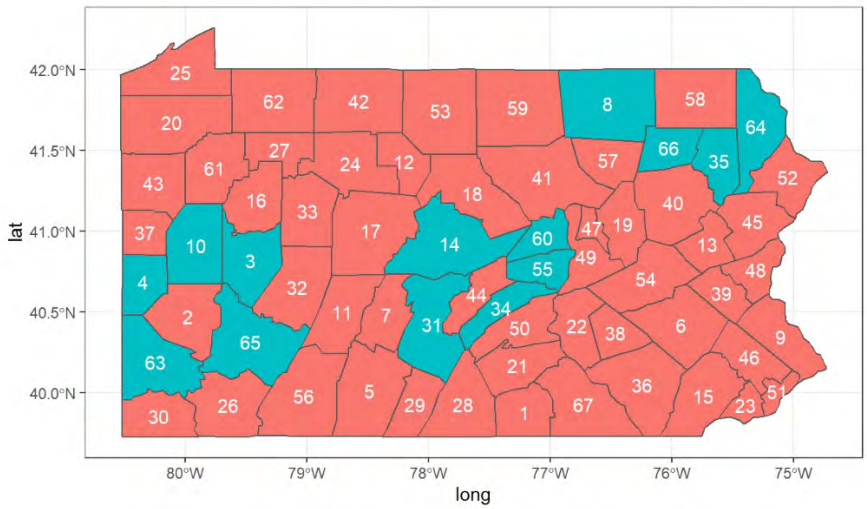



Figura 2.1: Mapa que muestra los estados de Pensilvania, la numeracion es de orden alfabético.

en orden alfabético, el condado 1 es Adams y el condado 67 es Sussex. A continuación se brinda una definición formal del lattice.

**Definición 2.1.1 (Lattice)** *Sea una malla del tipo de (lattice o grid)* $D \in \mathbb{R}^n$ *que contiene una colección contable de áreas espaciales* $D = \{s_i : i = 1, \ldots, n\}$ *, donde la variable respuesta se observa en cada área*

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

Cabe mencionar que los lattice pueden ser regulares o irregulares. Asimismo, la estructura de las áreas vecinas es importante. Por ello una herramienta que puede ser útil en la exploración inicial de datos de unidades de áreas y para construir la matriz de vecindad.

### 2.1.2. Matriz de vecindad

Antes de definir la matriz de vecindad definimos la matriz de distancias.

**Definición 2.1.2 (Matriz de distancias $D$)** *Nos permite representar la noción de proximidad entre las distintas observaciones en una región determinada. En la que cada dato alojado en una determinada celda* $(i, j)$ *de la matriz corresponderá con la distancia entre esa determinada área* $i$*, respecto a un área* $j$*. Esta matriz de distancias es diagonal y simétrica, de la forma:*

$$\boldsymbol{D} = \begin{pmatrix} 0 & d_{12} & \cdots & d_{1j} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2j} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{i1} & d_{i2} & \cdots & 0 & \cdots & d_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nj} & \cdots & 0 \end{pmatrix}.$$

A partir de la matriz de distancias $\boldsymbol{D}$, se puede especificar información de vecindad (basada por ejemplo en la distancia euclidiana).

Dadas las medidas $Y_1, Y_2, \ldots, Y_n$ asociadas con las regiones o áreas $1, 2, \ldots, n$, especificamos una matriz de vecindad $\boldsymbol{W}$ con entradas $\omega_{ij}$ referente a las áreas $i$ y $j$, tal que:

$$\boldsymbol{W} = \begin{pmatrix} 0 & \omega_{12} & \cdots & \omega_{1j} & \cdots & \omega_{1n} \\ \omega_{21} & 0 & \cdots & \omega_{2j} & \cdots & \omega_{2n} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{i1} & \omega_{i2} & \cdots & 0 & \cdots & \omega_{in} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \omega_{n1} & \omega_{n2} & \cdots & \omega_{nj} & \cdots & 0 \end{pmatrix},$$

donde asumimos la diagonal principal $\omega_{ii} = 0$. Algunas posibilidades para $\omega_{ij}$ son: i) $\omega_{ij}$ es la inversa de la distancia entre las unidades; ii) también se podría asumir que $\omega_{ij} = 1$ si la distancia entre las unidades es menor a una distancia fija (0 caso contrario); iii) O de forma más simple $\omega_{ij} = 1$ para las $m$ áreas vecinas con mayor proximidad (0 caso contrario). Los elementos de $\boldsymbol{W}$ pueden ser vistos como pesos; siendo el peso es mayor cuando $i$ y $j$ son areas próximas. Así $\boldsymbol{W}$ brinda un mecanismo para introducir la estructura espacial en el modelado.

Así existen múltiples formas de transformar una matriz de distancias en una matriz de proximidad o de "pesos espaciales". A continuación se muestran algunos ejemplos en detalle.

**Ejemplo 1:** (Transformación basada en relaciones de conectividad)

En el caso ii) puede definirse $\omega_{ij}$ como:

$$\omega_{ij} = \begin{cases} 1 & \text{si } d_{ij} \leq dist \quad \forall\, i, j \in \{1, \ldots, n\},\ i \neq j, \\ 0 & \text{caso contrario.} \end{cases}$$

Una variante en la relación de conectividad donde la distancia varía para cada una de las observaciones tal que se ajusta con un número determinado de $k$ adyacentes (criterio de las

adyacencias mas cercanas), luego,

$$\omega_{ij} = \begin{cases} 1 & si\ d_{ij} \leq dist_{i(k)} \quad \forall\, i,j\ \in \{1,\ldots,n\}\,,\ i \neq j, \\ 0 & \text{caso contrario.} \end{cases}$$

**Ejemplo 2**: (Transformación basada en contiguidades)

En el caso iii) puede definirse $\omega_{ij}$ como:

$$\omega_{ij} = \begin{cases} 1 & \text{si i y j son regiones adyacentes} \quad \forall\, i,j\ \in \{1,\ldots,n\}\,,\ i \neq j, \\ 0 & \text{caso contrario.} \end{cases}$$

Por ejemplo, el cuadro 2.1 define las regiones A, B, C y D, sus regiones vecinas.

| Región | Adyacencia |
|---|---|
| A | B,C |
| B | A,C,D |
| C | A,B |
| D | B |

Cuadro 2.1: Regiones adyacentes

A partir de esta información se obtiene la matriz de vecindad o de pesos espaciales dada por:

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

**Ejemplo 3**: (Matriz de pesos estandarizados)

Se puede estandarizar la matriz de pesos del caso anterior, tal que la suma de las filas de la matriz sea igual a uno. En el ejemplo estandarizando la matriz de pesos inicial, se obtiene la estructura de la nueva matriz de vecindad dada por:

$$\widetilde{W} = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & 0 & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

### 2.1.3. Medidas de asociación espacial

Existen dos estadísticas que son las más usadas para medir la fuerza de asociación espacial entre unidades de áreas, el índice de Moran y C de Geary.

- **Indice de Moran**: Sean las medidas $Y_1, Y_2, \ldots, Y_n$ asociadas con las unidades de área $1, 2, \ldots, n$. La estadística $I$ de Moran tiene la siguiente forma:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij} \left(Y_i - \overline{Y}\right) \left(Y_j - \overline{Y}\right)}{\left(\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij}\right) \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2},$$

  donde $\bar{Y}$ es el valor de la media muestral y $w_{ij}$ son los pesos de la matriz de vecindad. En general tenemos que $I \in [-1, 1]$, donde $I = -1$ sugiere una distribución espacial perfectamente dispersa o inversa, $I = 0$ significa patrón espacial aleatorio e $I = 1$ significa autocorrelación espacial perfecta.

  En términos prácticos, un valor de $I$ cercano a 1 significa si un área tiene un valor alto para una variable particular, sus áreas vecinas también tendrán valores altos, y lo mismo aplica para valores bajos.

- **C de Geary**: Sean las medidas $Y_1, Y_2, \ldots, Y_n$ asociadas con las unidades de área $1, 2, \ldots, n$. La estadística $C$ de Geary tiene la siguiente forma:

$$C = \frac{(n-1) \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij} \left(Y_i - Y_j\right)^2}{2 \left(\sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{ij}\right) \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2},$$

  donde $\bar{Y}$ es el valor de la media muestral y $w_{ij}$ son los pesos de la matriz de vecindad. La estadística $C$ toma cualquier valor positivo, $C > 0$. Valores bajos de C (es decir, entre 0 y 1) indican la presencia de asociación espacial.

### 2.1.4. Independencia Condicional

Para comprender el modelo condicional autoregresivo se revisa brevemente el concepto de la independencia condicional implícita en una cadena de Markov, o más generalmente en el caso espacial, un campo aleatorio de Markov (Rue y Held, 2005). La independencia condicional se discute en el contexto de modelos autorregresivos de series de tiempo.

Ejemplo: Proceso Autoregresivo de orde 1 - AR(1).

Sea el modelo AR(1) definido por:

$$Y_t = \rho Y_{t-1} + \epsilon_t \quad \epsilon_t \overset{iid}{\sim} N(0, 1), \quad |\rho| < 1.$$

Este modelo puede ser expresado en la forma condicional por:

$$Y_t|Y_1,\ldots,Y_{t-1} \sim N\left(\rho Y_{t-1},1\right) \quad t=2,\ldots,n,$$

asumiendo que la distribución marginal de $Y_1$ es normal con media cero y varianza $1/\left(1-\rho^2\right)$.

Luego la fdp conjunta de $\boldsymbol{Y}=(Y_1,\ldots,Y_n)'$ puede ser calculada a partir de las distribuciones condicionales de $Y_t$,

$$f_{\boldsymbol{Y}}\left(\boldsymbol{y}\right)=f\left(y_1\right)f\left(y_2|y_1\right)\ldots f\left(y_n|y_{n-1}\right)$$

$$f_{\boldsymbol{Y}}\left(\boldsymbol{y}\right)=\frac{1}{\left(2\pi\right)^{n/2}}|\boldsymbol{\Sigma}^{-1}|^{1/2}\exp\left(-\frac{1}{2}y'\boldsymbol{\Sigma}^{-1}y\right),$$

donde $\boldsymbol{\Sigma}^{-1}$ es una matriz tridiagonal, de la forma:

$$\boldsymbol{\Sigma}^{-1}=\begin{pmatrix} 1 & -\rho & \cdots & & & \cdots & \\ -\rho & 1+\rho^2 & -\rho & & & \cdots & \\ \vdots & \vdots & \ddots & \ddots & \vdots & \ddots & \vdots \\ & & \cdots & & -\rho & 1+\rho^2 & -\rho \\ \vdots & \vdots & \ddots & & \vdots & -\rho & 1 \end{pmatrix}.$$

Como se verá con más detalle en la siguiente sección, definir las distribuciones condicionales completas permite una especificación alternativa, pero equivalente, para derivar la densidad conjunta de $\boldsymbol{Y}$. Esto no es tan obvio como usar las densidades condicionales para formar las densidades conjuntas como un producto de estas densidades condicionales por la densidad marginal de $Y_1$, lo que requiere el nivel de detalle proporcionado por el Lema de Brook en campos aleatorios de Markov.

### 2.1.5. Lema de Brook y campos aleatorios de Markov (MRF)

Un resultado técnico útil para obtener la distribución conjunta de la $Y$ en algunos de los modelos que discutimos a continuación es el lema de Brook (Brook, 1964). La utilidad de este lema se expone en el artículo de Besag (1974) sobre modelos condicionalmente autorregresivos.

Sea un vector de variables aleatorias $\boldsymbol{Y}=(Y_1,\ldots,Y_n)'$. Según el lema de Brook se dice que el conjunto de las condicionales completas son compatibles, si se verifica que:

$$f\left(y_1,\ldots,y_n\right)=\prod_{i=1}^{n}\frac{f\left(y_i|y_1,\ldots,y_{i-1},y_{i+1,0,\ldots,}y_{n,0}\right)}{f\left(y_{i,0}|y_1,\ldots,y_{i-1},y_{i+1,0,\ldots,}y_{n,0}\right)}f\left(y_{1,0},\ldots,y_{n,0}\right), \tag{2.1}$$

o en forma más desarrollada el Lema de Brook se puede rescribir como

$$f(y_1,\ldots,y_n) = \frac{f(y_1|y_2,\ldots,y_n)}{f(y_{1,0}|y_2,\ldots,y_n)} \cdot \frac{p(y_2|y_{1,0},y_3,\ldots,y_n)}{f(y_{2,0}|y_{1,0},y_3,\ldots,y_n)}$$

$$\cdots \frac{f(y_n|y_{1,0},\ldots,y_{n-1,0})}{f(y_{n,0}|y_{1,0},\cdots,y_{n-1,0})} f(y_{1,0},\ldots,y_{n,0}),$$

donde $(y_{1,0},\ldots,y_{n,0})$ es definida en el soporte de $\boldsymbol{Y}$. Es decir, $f(y_1,\ldots,y_n)$ está determinada por las distribuciones condicionales completas y la constante de proporcionalidad $f(y_{1,0},\ldots,y_{n,0})$.

Para probar el Lema de Brook se usa el resultado básico de fdp condicionales y luego se aplica el método de inducción. Por ejemplo, en el caso de dos variables aleatorias:

$$\begin{aligned} f(y_1,y_2) &= f(y_1|y_2)\, f(y_2) = f(y_1|y_2)\, \frac{p(y_2|y_{1,0})}{f(y_{1,0}|y_2)} \\ &= \frac{f(y_1|y_2)}{f(y_{1,0}|y_2)} \frac{f(y_2|y_{1,0})}{f(y_{2,0}|y_{1,0})} f(y_{2,0}|y_{1,0})\, p(y_{1,0}) \\ &= \frac{f(y_1|y_2)}{f(y_{1,0}|y_2)} \frac{f(y_2|y_{1,0})}{f(y_{2,0}|y_{1,0})} f(y_{1,0},y_{2,0}). \end{aligned}$$

Luego es evidente que se puede construir una distribución conjunta para $\boldsymbol{Y} = (Y_1, Y_2)$, dado un conjunto completo de distribuciones condicionales (univariadas) completas.

Aunque Brook (1964) ilustra cómo crear la densidad conjunta a partir las distribuciones condicionales completas hasta una constante de proporcionalidad, a menudo es engorroso para un gran número de áreas geográficas. Así el campo aleatorio de Markov es la clave para pasar de las distribuciones condicionales completas a una distribución conjunta para $\boldsymbol{Y}$.

Un campo aleatorio de Markov (MRF) es un proceso estocástico. A cualquier MRF de este tipo le corresponde un grafo acíclico con aristas no dirigidas [1] (Whittaker 1990). Antes de que podamos definir el MRF en el campo espacial, primero debemos definir el concepto de un sistema de vecindad. Sea $S$ un conjunto de áreas de lattice con centroides $s \in S$. Luego usamos la notación $\partial s$ para denotar los vecinos de $s$.

**Definición (Markov random field):** Sea $X_s$ un campo aleatorio de valores discretos definido en la lattice $S$ con un sistema de vecindad $\partial s$. Además, suponga que $\boldsymbol{X}$ es una realización del campo aleatorio de Markov (MRF) $X_s$ y tiene una función de probabilidad o densidad (fdp) $f(x)$. Entonces decimos que $X_s$ es un MRF, si la fdp de $\boldsymbol{X}$ tiene la propiedad de que para todo $x \in \Omega$:

[1] Un grafo no dirigido $\mathcal{G}$ es un par ordenado $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ que está sujeto a las siguientes condiciones: (i) $\mathcal{V}$ es un conjunto, cuyos elementos se denominan vértices o nodos, y (ii) $\mathcal{E}$ es un multiconjunto de pares desordenados de vértices (no necesariamente distintos), llamados aristas o líneas.

$$f\left(x_s|x_r \text{ para } r \neq s\right) = f\left(x_s|x_{\partial r}\right).$$

En el contexto del modelo espacial, se espera que la distribución condicional completa para $Y_i$ dependa solo de los vecinos del área $i$ es decir del sistema vecindad $\partial_i$ tal que se puedan definir las distribuciones condicionales completas de la siguiente forma:

$$f\left(y_i|y_{-i}\right) = f\left(y_i|\left\{y_j : j\epsilon\partial_i\right\}\right),$$

donde $y_{-i}$ representa al vector $\boldsymbol{Y}$ sin $y_i$.

El Teorema de Hammersley-Clifford establece que para la fdp de $\boldsymbol{Y}$ definida sobre el lattice $S$ y el sistema de vecindad $\partial s$, entonces $\boldsymbol{Y}$ es un MRF. Los MRF que son gaussianos forman una clase de modelo introducido por Besag (1974).

## 2.2. Modelos de datos de áreas

Una vez definida la estructura de vecindad, se consideran modelos que se asemejan a los modelos autorregresivos en series de tiempo. Dos de estos modelos muy populares que incorporan esta información de vecinos se conocen como modelos autorregresivos condicional y simultáneo, es decir, los modelos CAR y SAR.

### 2.2.1. Modelos condicionales autoregresivos (CAR)

Primero se especifica directamente la distribución condicional de $Y_i$ condicionada por el resto de observaciones como promedio ponderado de los valores en tales observaciones y presencia de heterocedasticidad para cada observación. Así primero se asume que las distribuciones condicionales completas siguen una distribución normal:

$$y_i|y_j,\, j \neq i \sim N\left(\sum_j b_{ij}y_j, \tau_i^2\right),\, i = 1, \ldots, n. \tag{2.2}$$

Las condicionales completas definidas en la ecuación (2.2) son compatibles, es decir, a través del lema de Brook podemos obtener una "función de densidad conjunta" de la forma:

$$f\left(y_1, \ldots, y_n\right) \propto \exp\left\{-\frac{1}{2}\boldsymbol{y}'\boldsymbol{D}^{-1}\left(\boldsymbol{I}-\boldsymbol{B}\right)\boldsymbol{y}\right\}, \tag{2.3}$$

donde $\boldsymbol{B} = \{b_{ij}\}$ y $\boldsymbol{D}$ es diagonal con $D_{ii} = \tau_i^2$.

La expresión en la ecuación (2.3) sugiere una distribución normal multivariada para $\boldsymbol{Y}$

con media 0 y matriz de precisión $\boldsymbol{\Sigma}_y^{-1} = (I - B)^{-1} D$. Sin embargo, son necesarias condiciones que aseguren que la matriz de covarianza $\boldsymbol{D}^{-1}(I - \boldsymbol{B})$ es una matriz simétrica, específicamente:

$$\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2} \tag{2.4}$$

para todo $i$, $j$.

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{pmatrix}$$

$$\boldsymbol{D} = \begin{pmatrix} \tau_1^2 & 0 & \cdots & 0 \\ 0 & \tau_2^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \tau_n^2 \end{pmatrix}$$

A partir de la matriz de vecindad $\boldsymbol{W}$, las ecuaciones se definen por $b_{ij} = \omega_{ij}/\omega_{i+}$ y $\tau_i^2 = \tau^2/\omega_{i+}$, entonces resulta que $p(y_i|y_j, j \neq i) = N\left(\sum_j \omega_{ij} y_j/\omega_{i+}, \tau^2/\omega_{i+}\right)$. Luego se tiene que:

$$f(y_1, \ldots, y_n) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{y}'(\boldsymbol{D}_\omega - \boldsymbol{W})\boldsymbol{y}\right\}, \tag{2.5}$$

donde $\boldsymbol{D}_\omega$ es diagonal con $(\boldsymbol{D}_\omega)_{ii} = \omega_{i+}$.

Notemos un segundo aspecto $(\boldsymbol{D}_\omega - \boldsymbol{W})\mathbf{1} = 0$, es decir, $\boldsymbol{\Sigma}_y^{-1}$ es singular, por lo que su inversa $\boldsymbol{\Sigma}_y$ no existe y la distribución en 2.5 es llamada impropia.

Bajo las definiciones previas de $\boldsymbol{B}$ y $\boldsymbol{D}$, para abordar la impropiedad de la distribución en la ecuación (2.5), se redefine $\boldsymbol{\Sigma}_y^{-1} = \frac{1}{\tau^2}(\boldsymbol{D}_\omega - \rho\boldsymbol{W})$ donde para cierto parámetro de autocorrelación espacial $\rho$ se tenga que $\boldsymbol{\Sigma}_y^{-1}$ sea no singular. En particular, esto se cumple si $\rho \in \left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$, donde $\lambda_{(1)} < \lambda_{(2)} < \cdots < \lambda_{(n)}$ son los autovalores ordenados de $\boldsymbol{D}_\omega^{-1/2}\boldsymbol{W}\boldsymbol{D}_\omega^{-1/2}$.

Bajo la definición de $\boldsymbol{\Sigma}_y^{-1} = \boldsymbol{D}_\omega - \rho\boldsymbol{W}$, la distribución condicional completa $f(y_i|y_j, j \neq i)$ se convierte en $N\left(\rho\sum_j \omega_{ij} y_j/\omega_{i+}, \tau^2/\omega_{i+}\right)$. Por lo tanto, la media condicional de $Y_i$ es proporcional al promedio de los valores de sus vecinos. Esto permite una interpretación espacial sensata para el modelo CAR.

**Ejemplo** : CAR simple con dos vecinos:

$$Y_1|Y_2 \sim N\left(\rho w_{12} y_2, \tau^2/w_{1+}\right)$$

$$Y_1|Y_2 \sim N\left(\rho w_{21} y_1, \tau^2/w_{2+}\right)$$

$$f(y_1, y_2) \propto \frac{f\left(y_1|y_2\right) f\left(y_2|y_1 = 0\right)}{f\left(y_1 = 0|y_2\right)}$$

$$\propto \frac{\exp\left(-\frac{w_{1+}}{2\tau^2}\left(y_1 - \rho w_{12} y_2\right)^2\right) \exp\left(-\frac{w_{2+}}{2\tau^2}\left(y_2 - \rho w_{21} 0\right)^2\right)}{\exp\left(-\frac{1}{2\tau^2}\left(0 - \rho w_{12} y_2\right)^2\right)}$$

$$\propto -\frac{1}{2\tau^2}\left(\boldsymbol{y} - 0\right)' \begin{pmatrix} w_{1+} & -\rho w_{12} \\ -\rho w_{12} & w_{2+} \end{pmatrix} \left(\boldsymbol{y} - 0\right).$$

De este ejemplo se muestra que si:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\tau^2} \begin{pmatrix} w_{1+} & -\rho w_{12} \\ -\rho w_{12} & w_{2+} \end{pmatrix} = \frac{1}{\tau^2}(\boldsymbol{D}_w - \rho \boldsymbol{W}),$$

entonces

$$\boldsymbol{\Sigma} = \tau^2 \left(\boldsymbol{D}_w - \rho \boldsymbol{W}\right)^{-1}.$$

Y podemos concluir que $\boldsymbol{Y} = (Y_1, Y_2)^T$ sigue una distribución normal:

$$\boldsymbol{Y} \sim N\left(0, \tau^2 \left(\boldsymbol{D}_w - \rho \boldsymbol{W}\right)^{-1}\right).$$

Se pueden obtener límites más simples que los dados anteriormente para el parámetro de autocorrelación $\rho$ si reemplazamos la matriz de vecindad $\boldsymbol{W}$ por la matriz de vecindad "escalada" o "estandarizada" $\widetilde{\boldsymbol{W}} \equiv Diag\left(1/\omega_{i+}\right) \boldsymbol{W}$. Luego $\Sigma_y^{-1}$ puede escribirse entonces como $\frac{1}{\tau^2}\left(\boldsymbol{D}_w - \rho\widetilde{W}\right)$. Entonces si $|\rho| < 1$, $\boldsymbol{\Sigma}_y^{-1}$ es diagonalmente dominante y simétrica, por lo tanto es definida positiva.

### 2.2.2. Modelo Autoregresivo Simultáneo - SAR

El modelo autoregresivo simultáneo (SAR) fue introducido por Whittle (1954) y se define por un modelamiento simultáneo de la autocorrelacion espacial en la variable respuesta.

El modelo SAR asume que

$$Y_i = \sum_j b_{ij} Y_j + \epsilon_i; i = 1, 2, \ldots, n$$

donde $\epsilon_i \sim N\left(0, \tau^2\right)$ son errores independientes, o equivalentemente, $(\boldsymbol{I} - \boldsymbol{B})\boldsymbol{Y} = \boldsymbol{\epsilon}$ con $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. Por lo tanto, si $(\boldsymbol{I} - \boldsymbol{B})$ es de rango completo se tiene que:

$$\boldsymbol{Y} \sim N\left(0, (\boldsymbol{I} - \boldsymbol{B})^{-1} \widetilde{\boldsymbol{D}} \left((\boldsymbol{I} - \boldsymbol{B})^{-1}\right)^T\right), \tag{2.6}$$

donde $\widetilde{\boldsymbol{D}} = \tau^2 \boldsymbol{I}$, entonces la ecuación (2.6) se simplifica a

$$\boldsymbol{Y} \sim N\left(0, \tau^2 \left[(\boldsymbol{I} - \boldsymbol{B})(\boldsymbol{I} - \boldsymbol{B})^T\right]^{-1}\right).$$

Para garantizar una fdp normal multivariada en la ecuación (2.6), $(\boldsymbol{I} - \boldsymbol{B})$ debe ser de rango completo. Dos opciones se discuten con mayor frecuencia en la literatura. El primero enfoque asume $\boldsymbol{B} = \rho \boldsymbol{W}$ donde $\boldsymbol{W}$ es una matriz de vecindad. Aquí $\rho$ también se denomina parámetro de autorregresión espacial y, evidentemente, $Y_i = \rho \sum_j Y_j I\left(j \in \partial_i\right) + \epsilon_i$, donde $\partial_i$ denota el conjunto de vecinos de $i$. De hecho, se puede usar cualquier matriz de vecindad simétrica y, paralelamente como se explicó para el modelo CAR, se tiene que $\boldsymbol{I} - \rho \boldsymbol{W}$ será no singular si $\rho \in \left(\frac{1}{\lambda_{(1)}}, \frac{1}{\lambda_{(n)}}\right)$ donde $\lambda_{(1)} < \ldots < \lambda_{(n)}$ autovalores ordenados de $\boldsymbol{D}_\omega^{-1/2} \boldsymbol{W} \boldsymbol{D}_\omega^{-1/2}$.

De esta forma se puede puede definir el modelo SAR de forma matricial,

$$\boldsymbol{Y} = \rho \boldsymbol{W} \boldsymbol{Y} + \boldsymbol{\epsilon},$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} \omega_{11} & \omega_{12} & \cdots & \omega_{1N} \\ \omega_{21} & \omega_{22} & \cdots & \omega_{2N} \\ \vdots & & \ddots & \vdots \\ \omega_{N1} & \omega_{N2} & \cdots & \omega_{NN} \end{pmatrix} \begin{pmatrix} \rho_1 & 0 & \cdots & 0 \\ 0 & \rho_2 & 0 & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \rho_N \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

donde $\rho = diag\left(\rho_1, \ldots, \rho_N\right)$, el vector $\boldsymbol{\epsilon}$ tiene media $E\left(\boldsymbol{\epsilon}\right) = 0$ y covarianza $Cov\left(\boldsymbol{\epsilon}\right) = \tau^2 \boldsymbol{I}$.

Alternativamente, el segundo enfoque indica que $\boldsymbol{W}$ puede ser reemplazada por la matriz 'èstandarizada" $\widetilde{\boldsymbol{W}}$ donde para cada $i$, la i-ésima fila ha sido "normalizada" para sumar 1, es decir, $\widetilde{W}_{ij} = \omega_{ij}/\omega_{i+}$. Si establecemos que $\boldsymbol{B} = \rho \widetilde{\boldsymbol{W}}$, donde $\rho$ se llama parámetro de autocorrelación espacial, $\boldsymbol{I} - \rho \widetilde{\boldsymbol{W}}$ será no singular si $\rho \in (-1, 1)$. Luego

$$Y_i = \rho \sum Y_i I\left(j \in \partial_i\right) / \omega_{i+} + \epsilon_i,$$

$j$

y se garantiza la distribución normal multivariada

$$\boldsymbol{Y} \sim N\left(0, \tau^2 \left(I - \rho\widetilde{\boldsymbol{W}}\right)^{-1} \left(\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{-1}\right)^{\top}\right).$$

# Capítulo 3

# Estructura del modelo espacial

En esta tesis se propone estudiar cómo ha evolucionado el promedio del monto devengado de ejecución en inversiones públicas a nivel de provincias del Perú. Cabe mencionar que el gasto devengado hace mención al reconocimiento de una obligación de pago derivado del gasto comprometido previamente registrado [1] . Para ello, es importante entender primero las características de la variable de interés. Un modelo estadístico apropiado tomará en cuenta qué otras variables explican la variación del monto promedio de ejecución en las regiones. Entre las variables potenciales para el estudio se encuentran las siguientes:

- Si la inversión ha tenido declaratoria de Estado de Emergencia por Desastre: Hace mención al estado de excepción ante la condición de desastre ocasionado por un fenómeno de origen natural o inducido por la acción humana, con la finalidad de ejecutar acciones inmediatas y necesarias para la respuesta y rehabilitación.
- Latitud: Desde la perspectiva de la geografía, las latitudes son simplemente una medida de qué tan lejos está uno de Ecuador. Por lo que las ubicaciones que están lejos del Ecuador son propensas al clima frío.
- Longitud: Mide qué tan alto está uno sobre el nivel del mar, por lo que las ubicaciones que tienen una gran elevación son propensas al clima frío. Esto explica la distribución desigual de la precipitación en la parte norte y sur del Perú.
- Indice de Desarrollo Humano.
- Si la inversión ha presentado obras paralizadas.

[1] Se formaliza a través de la conformidad del área correspondiente en la entidad pública o Unidad Ejecutora que corresponda respecto de la recepción satisfactoria de los bienes y la prestación de los servicios solicitados y se registra sobre la base de la respectiva documentación sustentatoria.

- Si la inversión ha sido beneficiaria de fondos provenientes del Reconocimiento de Ejecución de Inversión de parte del MEF.
- Si la inversión pertenece a la cartera priorizada de inversiones de seguimiento por el MEF.
- Nivel de gobierno: Comprendida por las entidades adscritas al Sistema Nacional de Programación Multianual de Inversiones del gobierno nacional, regional y local.

Por otro lado, se espera que el monto (o logaritmo) promedio devengado en inversiones públicas sea similar en provincias vecinas. Esta dependencia espacial tendría que incorporarse en el modelo estadístico espacial para datos de áreas (Cressie, 1993; Banerjee et al., 2014). Diversos modelos para datos de áreas han sido propuestos, por ejemplo el modelo condicional autoregresivo (CAR), modelo autoregresivo espacial (SAR), modelo Besag-York (BYM), modelo direccionado aciclico autoregresivo (DAGAR) propuesto por Datta et al. (2019), entre otros.

Formalmente, se asume que la variable dependiente $Y_i$ representa el logaritmo del promedio de monto devengado de ejecución en inversiones públicas de la provincia i-ésima para $i = 1, \ldots, n$, donde $n$ es el número de provincias en Perú.

## 3.1. Modelo CAR

Primero, se plantea modelar $Y_i$ a través del modelo autoregresivo condicional (CAR) el cual asume que sigue la siguiente distribución condicional:

$$Y_i|Y_j \sim N(\rho \sum_j \frac{w_{ij}Y_j}{w_{i+}}, \frac{\tau^2}{w_{i+}}) \quad \text{para } j \neq i, \tag{3.1}$$

donde $\omega_{ij}$ corresponde a la relación de vecindad entre las provincias $i$ y $j$ de la matriz de pesos $\boldsymbol{W}$ de las provincias del Perú, $w_{i+}$ es el número de vecinos de la i-ésima provincia, $\tau^2$ es un parámetro de escala y $\rho$ es un parámetro de autocorrelación espacial. Esta definición asume que el monto de ejecución promedio en una provincia depende esencialmente de los montos de ejecución promedio de las provincias vecinas. A partir de estas funciones de densidad condicionales, se puede definir la función de densidad conjunta de $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ bajo condiciones específicas. En particular la distribución de $\boldsymbol{Y}$ es normal con vector de medias cero y matriz de precisión (inversa de la matriz de covarianza) conocida derivada de la ecuación (3.1). Generalizando este modelo, se puede incluir covariables a través de un vector de covariables $\boldsymbol{X}$, de tal forma que $\boldsymbol{Y}$ continua teniendo una distribución normal, pero ahora

el vector de medias está definido por $\boldsymbol{X\beta}$, y tiene la misma matriz de precisión derivada de la ecuación (3.1). Así, el modelo espacial usando CAR tiene la siguiente forma:

$$\boldsymbol{Y} \sim N\left(\boldsymbol{X\beta}, \boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}\right),$$

donde $\boldsymbol{\beta}$ es el vector de coeficientes de regresión y $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{D}_W - \rho\boldsymbol{W}\right)$ es llamada matriz de precisión. Para facilitar la interpretación del parámetro de autocorrelación espacial, se puede reemplazar $\boldsymbol{W}$ por $\widetilde{\boldsymbol{W}} \equiv Diag\left(1/\omega_{i+}\right)\boldsymbol{W}$, luego $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)$, con $\boldsymbol{D}_\omega$ es diagonal con $\left(\boldsymbol{D}_\omega\right)_{ii} = \omega_{i+}$.

Definido el modelo se deben estimar los parámetros $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \tau^2, \rho\}$, así como evaluar la precisión de los estimadores. En esta tesis se propone que la estimación se realice mediante máxima verosimilitud.

### 3.1.1. Inferencia clásica del modelo CAR

La función de verosimilitud del modelo CAR es definida por:

$$L\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) = \frac{1}{(2\pi)^{\frac{n}{2}}}\left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}}\exp\left[-\frac{1}{2}\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)^\top\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)\right].$$

Aplicando el logaritmo a la función de verosimilitud se tiene que:

$$\ell\left(\boldsymbol{\theta}\right) \propto -\frac{n}{2}\ln\left(\tau^2\right) + \frac{1}{2}\ln\left|\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right| - \frac{1}{2\tau^2}\left\{\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)^\top\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\left(Y-\boldsymbol{X\beta}\right)\right\}.$$

**Estimación de $\boldsymbol{\beta}$ y $\tau^2$**

Las derivadas de primer orden necesarias para hallar los estimadores de $\boldsymbol{\beta}$ y $\sigma^2$.

Primero se deriva la función de log-verosimilitud respecto a $\beta$:

$$\frac{d(\ell\left(\theta\right))}{d\boldsymbol{\beta}} = -\frac{d\left[-\frac{n}{2}\ln\left(\tau^2\right)\right]}{d\boldsymbol{\beta}} + \frac{1}{2}\frac{d\left[\ln\left|\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right|\right]}{d\boldsymbol{\beta}} - \frac{d\left[\frac{1}{2\tau^2}\left\{\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)^\top\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)\right\}\right]}{d\boldsymbol{\beta}}$$

$$\frac{d(\ell\left(\theta\right))}{d\boldsymbol{\beta}} \propto -\boldsymbol{X}^\top\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right) - \boldsymbol{X}^\top\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)$$

$$\propto -2\boldsymbol{X}^\top\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{Y}-\boldsymbol{X\beta}\right)$$

$$\frac{d(\ell(\theta))}{d\boldsymbol{\beta}} \propto -\boldsymbol{X}^\top \left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\boldsymbol{Y} + \boldsymbol{X}^\top \left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\boldsymbol{X}\boldsymbol{\beta}. \tag{3.2}$$

Luego en la ecuación 3.2 se iguala $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, se iguala a cero y despejando $\hat{\boldsymbol{\beta}}$ se obtiene el estimador de $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\boldsymbol{Y}. \tag{3.3}$$

De forma similar, se deriva la función de log-verosimilitud respecto a $\tau^2$:

$$\frac{d(\ell(\theta))}{d\tau^2} = \frac{d\left[-\frac{n}{2}\ln\left(\tau^2\right)\right]}{d\tau^2} + \frac{1}{2}\frac{d\left[\ln\left|\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right|\right]}{d\tau^2} - \frac{d\left[\frac{1}{2\tau^2}\left\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top \left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right\}\right]}{d\tau^2}.$$

Luego el estimador por máxima verosimilitud de $\tau^2$ es:

$$\hat{\tau}^2 = \frac{1}{n}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right) \tag{3.4}$$

Se observa que los estimadores para $\boldsymbol{\beta}$ y $\tau^2$ en las ecuaciones (3.3) y (3.4), respectivamente, dependen del parámetro $\rho$ que no es conocido. Y para el parámetro $\rho$ no se obtiene su estimador analíticamente por lo tanto se estima a través de otros métodos. Una opción es estimarlo numéricamente a través del algoritmo de Fisher-Scoring y reemplazando los EMV de $\boldsymbol{\beta}$ y $\tau^2$ iterativamente, hasta estimar los tres parámetros. Otra opción es usar los EMV de $\boldsymbol{\beta}$ y $\tau^2$ iterativamente y estimar $\rho$ por máxima verosimilitud restringida como se presenta a continuación.

**Estimación de $\rho$**

En la práctica los estimadores $\hat{\boldsymbol{\beta}}$ y $\hat{\tau}^2$ de las ecuaciones 3.3 y 3.4 se reemplazan en la función de verosimilitud restringida, definida por:

$$f(\rho^*) = -\frac{n}{2}\ln\left(\hat{\tau}^2\right) + \frac{1}{2}\ln\left|\boldsymbol{D}_W - \rho^*\widetilde{\boldsymbol{W}}\right| - \frac{1}{2\hat{\tau}^2}\left\{\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^\top \left(\boldsymbol{D}_W - \rho^*\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)\right\}.$$

Para estimar $\rho$ se maximiza numéricamente la función de verosimilitud restringida.

Sea $S(\rho^*) = \nabla f(\rho^*)$ la función score y $\mathcal{H}$ la matriz Hessiana correspondiente a las segunda derivada de $f(\rho^*)$. La información de Fisher esperada es $\boldsymbol{I}(\rho^*) = -E[\mathcal{H}(\rho^*)]$. La información de Fisher y función Score se emplean en el algoritmo de Fisher-Scoring para la

estimación de $\rho$,

$$\rho^{*(t+1)} = \rho^{*(t)} + \left[I\left(\rho^{*(t)}\right)\right]^{-1} S\left(\rho^{*(t)}\right),$$

donde $\rho^{*(t)}$ es el valor del parámetro en la i-ésima iteración. Comenzando con el valor inicial de $\rho^{*(0)}$ iteramos la ecuación hasta que cierto criterio de convergencia es alcanzado (es decir, $\rho^{*(t+1)} \approx \rho^{*(t)}$). La solución es $\hat{\rho}^* \approx \rho^{*(t)}$ para cierto valor de $t$ que cumple el criterio de convergencia.

## 3.2. Modelo SAR

Por otro lado, se plantea modelar $Y_i$ a través del modelo SAR:

$$Y_i = \rho \sum_{j \neq i} \omega_{ij} Y_j + \epsilon_i, \tag{3.5}$$

donde $\epsilon_i \sim N\left(0, \tau^2\right)$, $\omega_{ij}$ corresponde a la relación de vecindad entre las provincias $i$ y $j$ de la matriz de pesos $\boldsymbol{W}$ de las provincias del Perú, $\tau^2$ es un parámetro de escala y $\rho$ es un parámetro de autocorrelación espacial. Esta definición asume que el monto de ejecución promedio en una provincia depende de todos los montos de ejecución promedio de las provincias vecinas. A partir de esta definición, se puede definir el modelo para $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ de forma matricial tal que

$$\boldsymbol{Y} = \rho \boldsymbol{W} Y + \boldsymbol{\epsilon},$$

donde $E(\boldsymbol{\epsilon}) = 0$ y covarianza $Cov(\boldsymbol{\epsilon}) = \tau^2 \boldsymbol{I}$. Luego la distribución de $\boldsymbol{Y}$ es normal con vector de medias cero y matriz de precisión (inversa de la matriz de covarianza) conocida, es decir,

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

donde $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{I} - \rho \boldsymbol{W}\right)\left(\boldsymbol{I} - \rho \boldsymbol{W}\right)^\top$. Adicionalmente, se requieren condiciones específicas para que $(\boldsymbol{I} - \rho \boldsymbol{W})$ sea positiva definida, en particular, se reemplaza $\boldsymbol{W}$ por $\widetilde{\boldsymbol{W}} \equiv Diag\left(1/\omega_{i+}\right)\boldsymbol{W}$, luego $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{I} - \rho \widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{I} - \rho \widetilde{\boldsymbol{W}}\right)^\top$.

Generalizando este modelo (ecuación (3.5)), se puede incluir covariables a través de una matriz de covariables $\boldsymbol{X}$, así el modelo espacial SAR toma la siguiente forma matricial,

$$\boldsymbol{Y} = \rho \widetilde{\boldsymbol{W}} \boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

tiene distribución normal, tal que si $\boldsymbol{R} = (\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}})$ entonces $\boldsymbol{Y} \sim N\left(\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}\right)$, donde $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{\top}$. Los parámetros del modelo son: $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \tau^2, \rho\right)$.

### 3.2.1. Inferencia clásica del modelo SAR

Se tiene la siguiente expresión para la función de verosimilitud:

$$L\left(\boldsymbol{\theta}\right) = (2\pi)^{-\frac{n}{2}}\left|\boldsymbol{\Sigma}^{-1}\right|^{\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{Y} - \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta}\right)^{\top}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Y} - \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta}\right)\right\}.$$

La función de log-verosimilitud es:

$$\begin{aligned}\ell\left(\boldsymbol{\theta}\right) \propto & -\frac{n}{2}\ln\left(\tau^2\right) + \frac{1}{2}\ln\left|\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{\top}\right| \\ & - \frac{1}{2\tau^2}\left\{\left(\boldsymbol{Y} - \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta}\right)^{\top}\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{\top}\left(\boldsymbol{Y} - \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta}\right)\right\}.\end{aligned} \tag{3.6}$$

El procedimiento implica estimar analíticamente $\hat{\boldsymbol{\beta}}$ y $\hat{\tau}^2$ en función de $\rho$, mientras que la estimación de $\rho$ necesita de algún método numérico. A continuación se presentarán los estimadores para $\boldsymbol{\beta}$ y $\tau^2$.

### 3.2.2. Estimación de $\hat{\boldsymbol{\beta}}$ y $\hat{\tau^2}$

Derivando la funcion de log-verosimilitud en la ecuación (3.6) respecto a $\boldsymbol{\beta}$ e igualando a cero, obtenemos el estimador de máxima verosimilitud:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\top}\boldsymbol{R}^{-1}\boldsymbol{Y}\right)$$

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\top}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{X}^{\top}\left(\boldsymbol{I} - \widetilde{\boldsymbol{W}}\rho\right)\boldsymbol{Y}\right).$$

Derivando la función de log-verosimilitud en la ecuación (3.6) respecto a $\tau^2$ e igualando a cero, obtenemos el estimador de máxima verosimilitud:

$$\hat{\tau}^2 = \frac{1}{n}\left(\boldsymbol{R}\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^{\top}\left(\boldsymbol{R}\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^{\top}$$

$$\hat{\tau}^2 = \frac{1}{n}\left(\left(\boldsymbol{I} - \widetilde{\boldsymbol{W}}\rho\right)\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^{\top}\left(\left(\boldsymbol{I} - \widetilde{\boldsymbol{W}}\rho\right)\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}\right)^{\top}.$$

### 3.2.3. Estimación de $\rho$

El parámetro $\rho$ no se puede estimar analíticamente, por lo tanto se estima numéricamente como se ha explicado previamente. Para estimar $\rho$, se reemplaza los estimadores de $\hat{\boldsymbol{\beta}}$ y $\hat{\tau}^2$ en la función de verosimilitud restringida:

$$f(\rho^*) = -\frac{n}{2}\ln\left(2\pi\hat{\tau}^2\right) + \frac{1}{2}\ln\left|(\boldsymbol{I}-\rho^*\widetilde{\boldsymbol{W}})(\boldsymbol{I}-\rho^*\widetilde{\boldsymbol{W}})^\top\right| - \frac{1}{2\hat{\tau}^2}\left\{(\boldsymbol{Y}-\boldsymbol{R}^{-1}X\boldsymbol{\beta})^\top\left(\boldsymbol{I}-\rho^*\widetilde{\boldsymbol{W}}\right)^\top\left(\boldsymbol{I}-\rho^*\widetilde{\boldsymbol{W}}\right)(\boldsymbol{Y}-\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\beta})\right\}.$$

Sea $S(\rho^*) = \nabla f(\rho^*)$ la función score y $\mathcal{H}$ la matriz Hessiana correspondiente a las segunda derivada de $f(\rho^*)$. La información de Fisher esperada es $\boldsymbol{I}(\rho^*) = -E[\mathcal{H}(\rho^*)]$. La información de Fisher y función Score se emplean en el algoritmo de Fisher-Scoring para la estimación de $\rho$,

$$\rho^{*(t+1)} = \rho^{*(t)} + \left[I\left(\rho^{*(t)}\right)\right]^{-1} S\left(\rho^{*(t)}\right),$$

donde $\rho^{*(t)}$ es el valor del parámetro en la i-ésima iteración. Comenzando con el valor inicial de $\rho^{*(0)}$ iteramos la ecuación hasta que cierto criterio de convergencia es alcanzado (es decir, $\rho^{*(t+1)} \approx \rho^{*(t)}$). La solución es $\hat{\rho}^* \approx \rho^{*(t)}$ para cierto valor de $t$ que cumple el criterio de convergencia.

## 3.3. Cálculo de Intervalos de Confianza

Sea $\theta = (\theta_1, \ldots, \theta_d)$ un vector de parámetros a estimar y $l(\theta)$ la función de log-verosimilitud, entonces la matriz de información de Fisher Esperada es definida por:

$$I_E(\theta) = \begin{bmatrix} E\left[-\frac{\partial^2 l(\theta)}{\partial\theta_1^2}\right] & \cdots & \cdots & E\left[-\frac{\partial^2 l(\theta)}{\partial\theta_1\partial\theta_d}\right] \\ \vdots & \ddots & E\left[-\frac{\partial^2 l(\theta)}{\partial\theta_i\partial\theta_j}\right] & \vdots \\ \vdots & E\left[-\frac{\partial^2 l(\theta)}{\partial\theta_j\partial\theta_i}\right] & \ddots & \vdots \\ E\left[-\frac{\partial^2 l(\theta)}{\partial\theta_d\partial\theta_1}\right] & \cdots & \cdots & E\left[-\frac{\partial^2 l(\theta)}{\partial\theta_d^2}\right] \end{bmatrix}.$$

Bajo ciertas condiciones de regularidad, para tamaños de muestra grande, es decir cuando $n \to \infty$, entonces el estimador de máxima verosimilitud de $\theta$ sigue una distribución normal asintóticamente:

$$\hat{\theta} \sim N\left(\theta, I_E(\theta)^{-1}\right).$$

Luego la ecuación para el intervalo de confianza al $(1-\alpha)\,\%$ para un parámetro $\theta_i$ se define como:

$$\theta_{iL} = \hat{\theta}_i - z_{1-\alpha/2}J_{i,i}^{\frac{1}{2}}, \quad \theta_{iU} = \hat{\theta} + z_{1-\alpha/2}J_{i,i}^{\frac{1}{2}}$$

donde $\hat{\theta}_i$ es el estimador de máxima verosimilitud (EMV), $z_{1-\alpha/2}$ es el quantil de la distribución normal estándar y $J_{i,i}^{\frac{1}{2}}$ es la desviación estándar del EMV $\hat{\theta}_i$, el i-ésimo elemento de la diagonal de la matriz $J = I_E(\underline{\theta})^{-1}$.

# Capítulo 4

# Estudio de Simulación

En este capítulo, realizaremos un estudio de recuperación de parámetros con el objetivo de mostrar la idoneidad del método de estimación propuesto. Se desarrollan dos casos de estudios de simulación usando los modelos CAR y SAR planteados en el capítulo 3. Cada estudio se divide en un escenario para un conjunto prámetros inciales. Bajo los estudios de simulación buscamos simular la variables de respuesta del modelo espacial y verificar la correcta estimación de los parámetros e hiperparámetros usando inferencia clásica.

## 4.1. Simulación del modelo CAR

En esta sección se simulan datos provenientes del modelo CAR definido en la sección 3.1.

### 4.1.1. Generación de Datos

En la simulación de un modelo espacial autorregresivo condicional se genera datos espaciales que exhiben autocorrelación espacial basada en una estructura CAR específica. A continuación se describe la implementación realizada para simular los datos:

- **Delimitación del área de estudio**

Para la simulación de los datos espaciales se establece un tamaño de muestra $n = 171$ distritos del departamento de Lima (Figura 4.1). Donde se asume asume que $Y_i$ representa la variable aleatoria de interés en el i-ésimo distrito tal que $i = 1, \ldots, n$.

- **Especificación de las relaciones de vecindad y creación del grafo**

Para cada unidad geográfica del conjunto de datos espaciales de los distritos de Lima, se identifica sus unidades vecinas. Estas unidades se determinan por la contigüidad directa (es

Figura 4.1: Mapa de los distritos del departamento de Lima.

decir, compartir un límite). Con la identificación de estas relaciones espaciales se construye la matriz de pesos espaciales $\boldsymbol{W}$. Esta matriz codifica las relaciones espaciales entre las unidades, y es central en el modelo CAR. Las ponderaciones elegido en $\boldsymbol{W}$ es del tipo binaria (donde se asigna 1 si dos unidades son vecinas y 0 si no lo son).

En R, con el paquete spdep a través de la la función poly2nb() se identifica los vecinos basándose en la contigüidad de las unidades geográficas. En particular, se opta por la contigüidad de tipo "queen", que considera dos unidades como vecinas si comparten al menos un punto en común. A modo ilustrativo, para nuestra data de estudio para el distrito de Chorrillos en la provincia de Lima, las unidades vecinas identificados son Villa el Salvador, Barranco, San Juan de Miraflores y Santiago de Surco.

Una vez determinados los vecinos (Figura 4.2), se construye la matriz de vecindad $\boldsymbol{W}$. En nuestro contexto, esta matriz es de $171 \times 171$, reflejando el número total de distritos considerados. Cada entrada en $\boldsymbol{W}$ indica si un par de distritos son vecinos (representado por un 1) o no (representado por un 0).

La matriz $\boldsymbol{W}$ se puede visualizar de manera eficiente usando técnicas de mapeo, donde se muetran las relaciones de vecindad. En la Figura 4.3 podemos observar la matriz $\boldsymbol{W}$, donde los cuadrados rojos representan que los distritos $i$ y $j$ son vecinas, además de contar con la

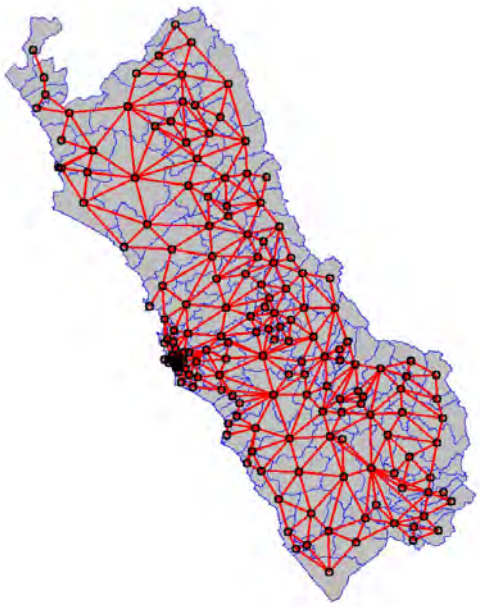


Figura 4.2: Grafo de los distritos del departamento de Lima.

diagonal con cuadrados rojos.

Finalmente, para convertir la lista de vecinos en una matriz binaria, se utiliza la función nb2mat() del paquete spdep. Esta matriz codifica relaciones espaciales y es instrumental en el cálculo de métricas de autocorrelación espacial. En este estudio particular, se encontró que, en promedio, cada distrito tiene 5.2 vecinos.

- **Definición del vector de medias y matriz de precisión**

Para la simulación de la covariable, $x_i$ es simulada de una variable aleatoria con distribución normal estándar $x_i \sim N(0,1)$, tal que $\boldsymbol{X}_i = (1, x_i)$. Los coeficientes de regresión se definen como $\beta = (\beta_0, \beta_1) = (0.5, 2)$. Se genera $\boldsymbol{X\beta}$, para $i = 1, \ldots, n$ a partir de $\boldsymbol{X}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_i$. Para la generación de la matriz de precisión (inversa de la covarianza) se asigna al parámetro $\rho$ un valor igual a 0.8 y para $\tau^2$ un valor igual a 1. Luego se obtiene $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)$. Finalmente se simulan datos de la v.a. $Y_i$ para $i = 1, ..., n$ a partir de la distribución del vector aleatorio $\boldsymbol{Y} \sim N\left(\boldsymbol{X\beta}, \boldsymbol{Q}^{-1}\right)$. En la Figura 4.4 se puede observar la simulación de la variable respuesta en los distritos del departamento de Lima.

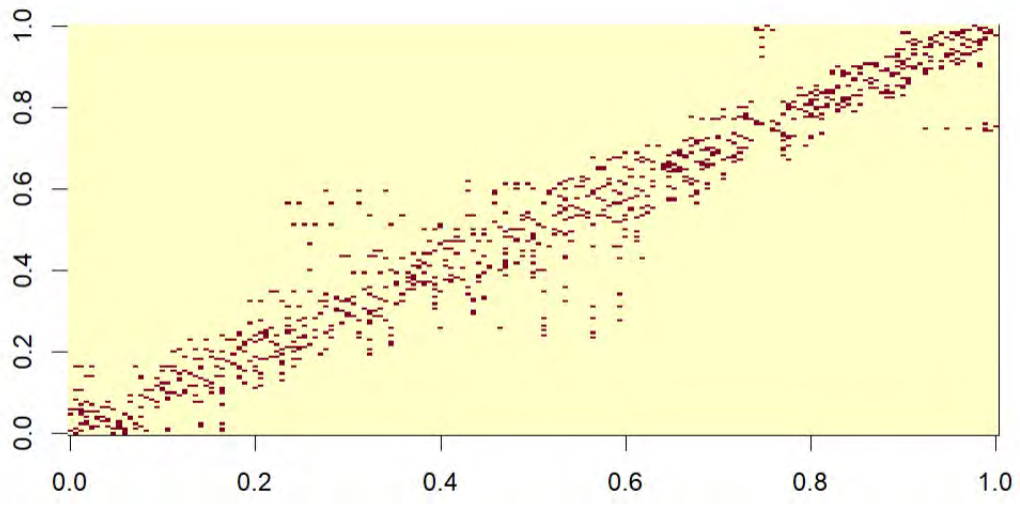



Figura 4.3: Matriz de vecindad de la provincia de Lima.

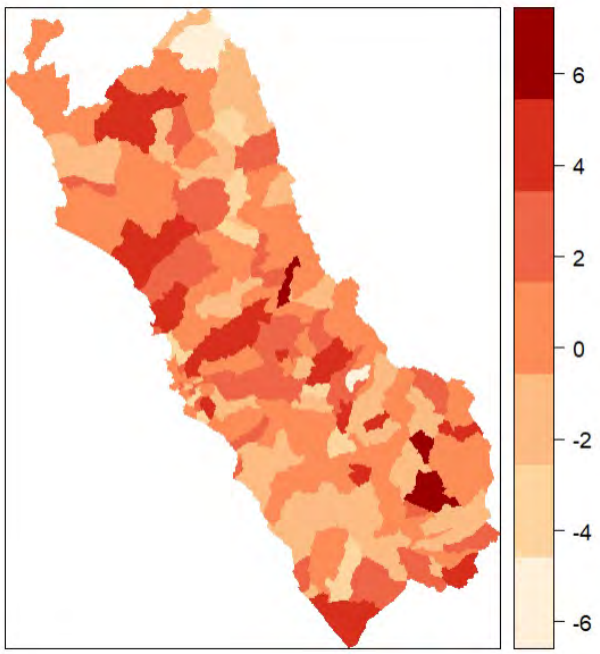



Figura 4.4: Simulación de la variable de estudio en cada distrito de la provincia de Lima.

### 4.1.2. Estimación

Se implementó y ajustó en R la inferencia del modelo CAR para datos de áreas. En el cuadro 4.1 se muestran las estimaciones por el método de máxima verosimilitud de los parámetros. Se encontraron resultados deseados pues se logró recuperar los parámetros predeterminados, es decir, se obtuvieron estimaciones aproximadas a los valores reales. De los resultados obtenidos también observamos que los intervalos de confianza (en adelante IC) contienen al valor real del parámetro.

Cuadro 4.1: Resultados de las estimaciones puntuales: media, desviación estándar e invervalos de confianza (IC) al 95 % del modelo CAR.

| Parámetro | Real | Estimado | Desviación Estándar | IC (95 %) |
|---|---|---|---|---|
| $\beta_0$ | 0.500 | 0.428 | 0.067 | (0.299, 0.558) |
| $\beta_1$ | 2.000 | 2.022 | 0.031 | (1.962, 2.082) |
| $\rho$ | 0.800 | 0.734 | 0.121 | (0.497, 0.970) |
| $\tau^2$ | 1.000 | 1.042 | 0.118 | (0.811, 1.272) |

En la Figura 4.5 se muestra que para una grid valores del $\rho$ y $\tau^2$, el valor máximo del logaritmo de la función de verosimilitud se obtiene aproximadamente para $\rho \in (0.5,1)$ y $\tau^2 \in (0.8, 1.3)$.

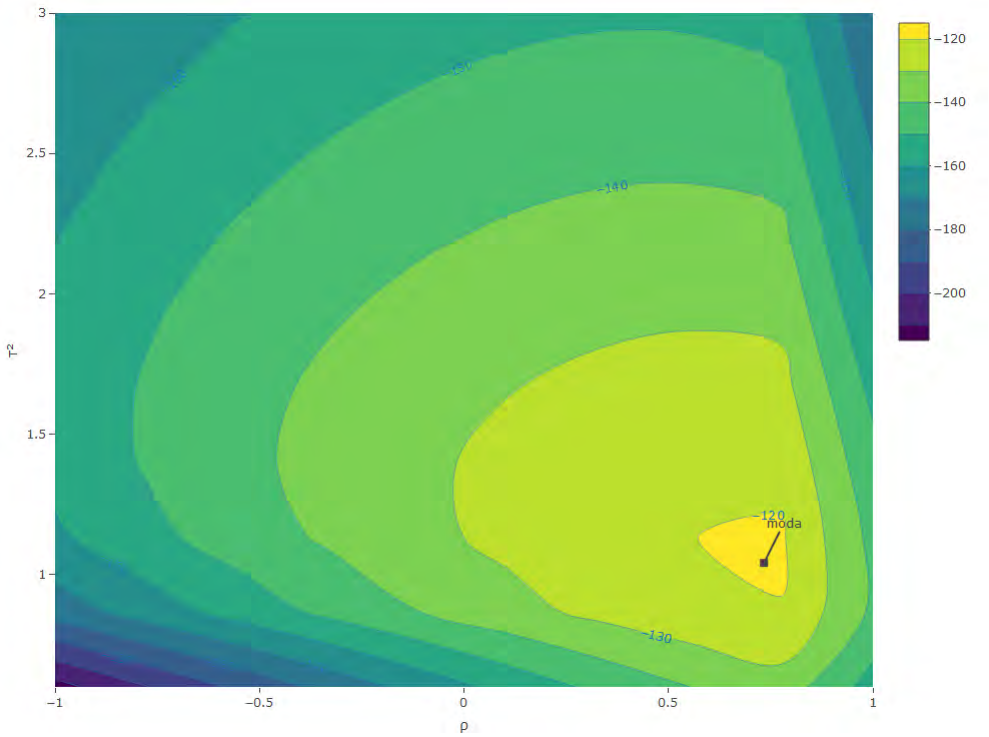



Figura 4.5: Gráfico de contorno de la función de log-verosimilitud para los parámetros espaciales $\rho$ e $\tau^2$.

## 4.2. Simulación del modelo SAR

En esta sección se simulan datos provenientes del modelo SAR definido en la sección 3.2. Se procede de forma similar a la simulación de datos del modelo CAR, se delimita el área de estudio, se especifica las relaciones de vecindad y se crea el grafo.

Luego se define el vector de medias y simula datos de simulación de la covariable, $x_i$ es simulada de una variable aleatoria con distribución normal estándar $x_i \sim N(0,1)$, tal que $\boldsymbol{X}_i = (1, x_i)$. Los coeficientes de regresión se definen como $\beta = (\beta_0, \beta_1) = (0.5, 2)$. Se genera $\boldsymbol{X\beta}$, para $i = 1, \ldots, n$ a partir de $\boldsymbol{X}_i^\top \boldsymbol{\beta} = \beta_0 + \beta_1 x_i$. Para la generación de la matriz de precisión (inversa de la covarianza) ahora del modelo SAR se asigna al parámetro $\rho$ un valor igual a 0.8 y para $\tau^2$ un valor igual a 1. Luego se obtiene $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^\top$. Finalmente se simulan datos (Figura 4.6) de la v.a. $Y_i$ para $i = 1, \ldots, n$ a partir de la distribución del vector aleatorio $\boldsymbol{Y} \sim N\left(\boldsymbol{R}^{-1}\boldsymbol{X\beta}, \boldsymbol{Q}^{-1}\right)$, donde $\boldsymbol{R} = \left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)$.

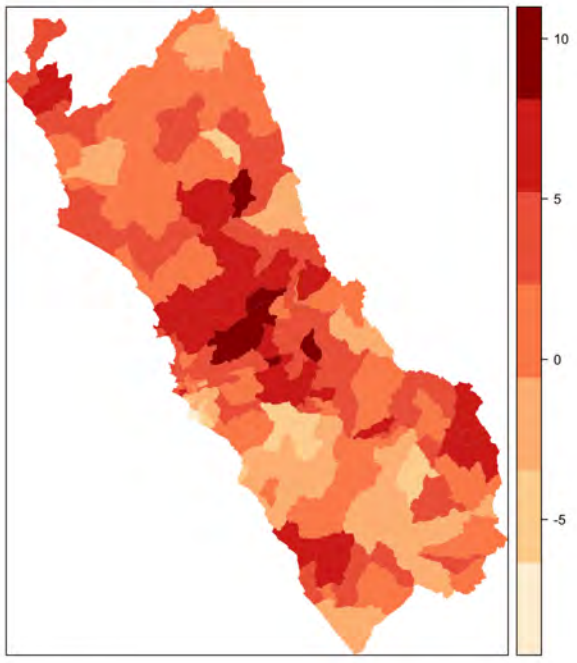



Figura 4.6: Simulación de la variable de estudio del modelo SAR en cada distrito de Lima.

Se implementó y ajustó en R la inferencia del modelo SAR para datos de áreas. El cuadro 4.2 muestra las estimaciones de los parámetros, los intervalos de confianza y la desviación estándar de cada parámetro de la simulación de una distribución SAR. Podemos observar que las estimaciones se acercan a los valores originales de los parámetros, y el valor original de los parámetros se encuentran dentro del intervalo de confianza al 95 %.

Cuadro 4.2: Resultados de las estimaciones puntuales: media, desviación estándar e invervalos de confianza (IC) al 95 % del modelo SAR

| Parámetro | Real | Estimado | Desviación Estándar | IC (95 %) |
|---|---|---|---|---|
| $\beta_0$ | 0.500 | 0.541 | 0.059 | $(0.425, 0.657)$ |
| $\beta_1$ | 2.000 | 2.112 | 0.078 | $(1.960, 2.265)$ |
| $\rho$ | 0.800 | 0.789 | 0.023 | $(0.744, 0.834)$ |
| $\tau^2$ | 1.000 | 1.043 | 0.114 | $(0.820, 1.266)$ |

En la Figura 4.7 se muestra que para una grid valores del $\rho$ y $\tau^2$ el valor máximo del logaritmo de la función de verosimilitud se obtiene aproximadamente para $\rho \in (0.5, 1)$ y $\tau^2 \in (1, 1.5)$.

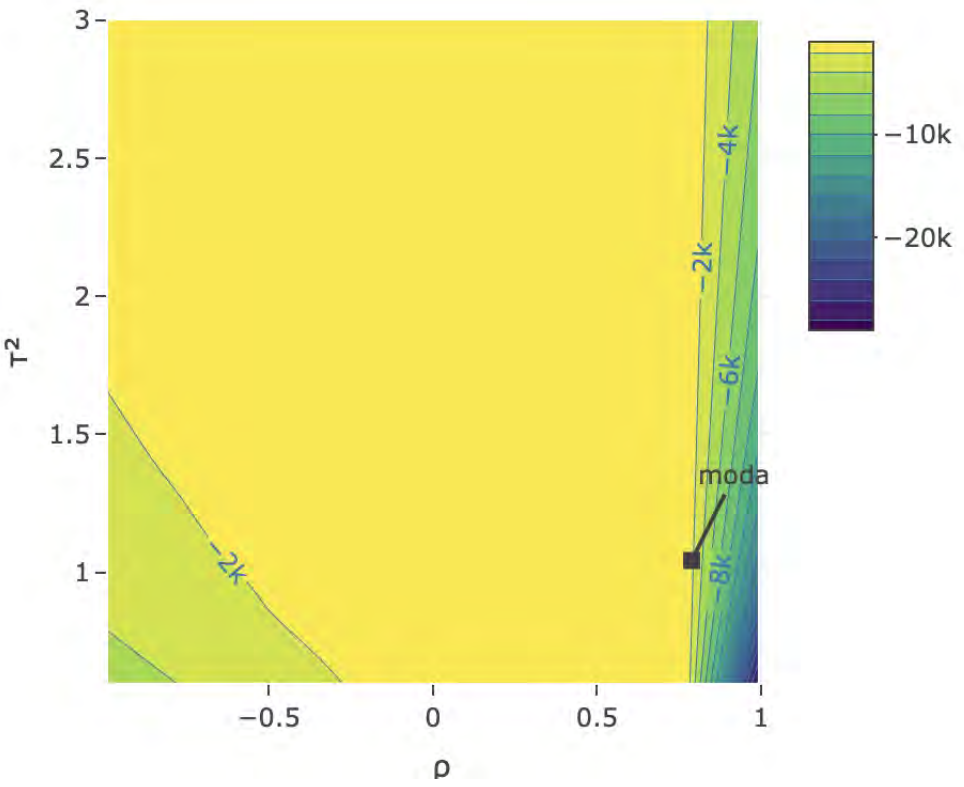



Figura 4.7: Gráfico de contorno de la función de log-verosimilitud para los parámetros espaciales $\rho$ e $\tau^2$.

## 4.3. Comparación de IC : CAR - SAR

El modelo autorregresivo condicional (CAR) y el modelo autorregresivo simultáneo (SAR) se emplean para simular datos correlacionados espacialmente. Bajo los supuestos del escenario del estudio de simulación se estudian los intervalos (IC) al 95 % de confianza (Figura 4.8). Observamos que el valor original de cada parámetro se encuentra dentro del IC con 95 % nivel de confianza.

En el modelo CAR, los intervalos de confianza para el parámetro $\beta_1$ es más estrecho en comparación con los del modelo SAR. En contraste, el intervalo de confianza para el parámetro $\rho$ muestra una mayor amplitud en el modelo CAR.

Los resultados indican una precisión diferencial en los modelos CAR y SAR con respecto a la autocorrelación espacial y las estimaciones de los coeficientes de regresión. Un intervalo de confianza mayor para $\rho$ en el modelo CAR sugiere menos certeza en la captura de la dependencia espacial, mientras que el intervalos más estrecho para el coeficiente de regresión $\beta_1$ sugiere una comprensión más clara de la relación de la covariable en este modelo.

Estos resultados sugieren una mayor precisión en torno a esas estimaciones de parámetros específicos. Sin embargo, la decisión sobre qué modelo utilizar debe tener en cuenta otros factores como el ajuste general del modelo y la naturaleza de los datos. Sería conveniente profundizar en otras métricas de diagnóstico y considerar el contexto y los objetivos específicos del análisis al decidir entre modelos.

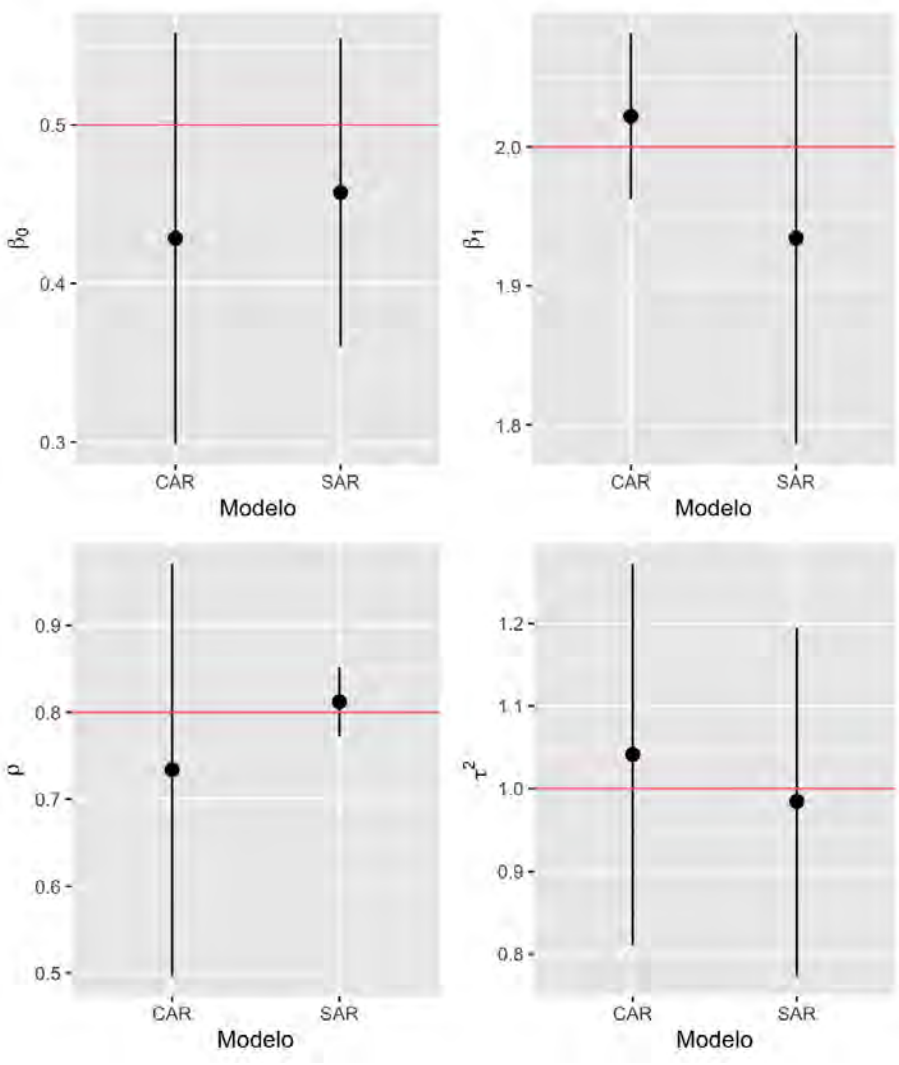



Figura 4.8: Intervalo de confianza al 95 % para $\beta_0$, $\beta_1$, $\rho$ y $\tau^2$ para los modelos ajustados.

## 4.4. Estimación de la variable respuesta

Finalmente, la Figura 4.9 compara los valores originales simulados de la variable respuesta (observado) con sus estimaciones usando el modelo CAR y el modelo SAR. En ambos casos, se observa que se ha recuperado bien los valores originales, presentándose una ligera mayor variabilidad en el modelo SAR. Para corroborar este resultado se calculó la raíz del error cuadrático medio estimado (RMSE), se obtuvieron los valores 0.5166268 y 1.432582, para el modelo CAR y SAR, respectivamente.

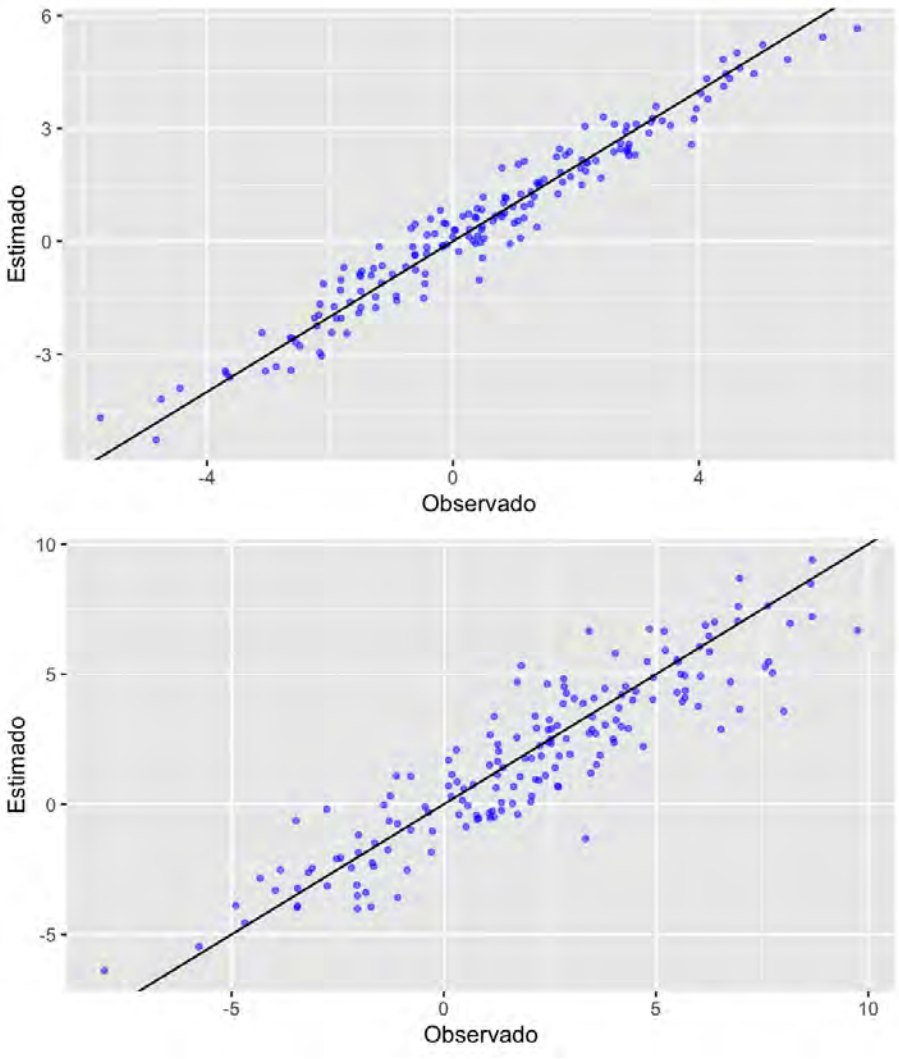


Figura 4.9: Estimación de la variable respuesta versus el valor original (observado) de la variable respuesta para los modelos ajustados: CAR (arriba) y SAR (abajo).

# Capítulo 5

# Aplicación

En este capítulo, se presenta los resultados del modelo espacial CAR, empleado para analizar datos del monto devengado en inversión pública, específicamente de las inversiones registradas en el aplicativo de Banco de invesiones del Sistema Nacional de Programación Multianual y Gestión de Inversiones- SNPMGI. El estudio abarca información recopilada de todos los departamentos del Perú correspondiente al año 2022. Se detalla la información y análisis empleado en la elaboración de esta investigación, abordando: las dependencias espaciales ligadas al desempeño de las inversiones, la pertinencia de los modelos espaciales, culminando con el proceso de modelado, estimación de parámetros y valoración del modelo seleccionado.

## 5.1. Descripción de la variable respuesta

La base de datos ha sido extraída del Sistema Nacional de Programación Multianual y Gestión de Inversiones - SNPMGI donde muestra datos de las inversiones públicas georeferenciadas de todo el Perú. En particular, se seleccionaron para esta tesis los datos correspondientes al año 2022, encontrándose dentro de ese período que los gobiernos locales han devengado de manera acumulada un total de S/ 21,245 millones, siendo los departamentos con mayor monto devengado: Cusco (S/. 2,752.6 millones), Ancash (S/. 2,143.6 millones), Lima (S/. 1,852.9 millones ), Piura (S/. 1422.8 millones), y Arequipa (S/. 1,375.8 millones).

Los datos se inversiones se encuentran organizados por ubicación geográfica (departamento, provincia y distrito) y para obtener los valores de interés de las variables a nivel de provincia, se realizó el promedio de las variables continuas. Asimismo, este conjunto de datos está compuesto por los campos descritos en el cuadro 5.1 correspondiente a la inversiones públicas del SNPMGI.

Cuadro 5.1: Organización de la data de las inversiones públicas del SNPMGI.

| CAMPO | DEFINICIÓN |
|---|---|
| Código de Inversiones | Clave única asignada a cada inversión para su identificación geográfica |
| Departamento | Identificación y localización de la inversión por departamento. |
| Provincia | Identificación y localización de la inversión por provincia. |
| Distrito | Identificación y localización de la inversión por distrito. |
| Nivel de gobierno | Clasificación de la inversión según el nivel de gobierno (nacional, regional, local). |
| Monto de devengado | Se refiere al monto devengado de los gastos derivados por la ejecución obra, la adquisición de los bienes, la prestación de los servicios y otros costos asociados a la inversión. |

## 5.2. Análisis exploratorio

En la Figura 5.1 se muestra el histograma de los datos de variable de estudio, el logaritmo del monto devengado promedio $Y_i$, en inversión pública por provincia. Podemos observar que los datos $y_i$ tienen una ligera asimetría hacia la derecha, pero sigue aproximadamente una distribución normal. Por ello como se explicó en el capítulo 3, la presente tesis utilizará la variable de respuesta $Y_i$ , la cual se asume que sigue una distribución normal.

El mapa de la Figura 5.2 muestra una distribución claramente desigual del monto devengado en inversión pública en las provincias, en la escala logarítmica. Las provincias con colores más oscuros, indican mayores montos de inversión. Mientras que otras áreas muestran significativamente menos inversión (colores más claros). Además se puede observar que cuando una provincia tiene un monto de inversión elevado, las provincias vecinas también tienen montos de inversión elevados, por ello a partir de este análisis exploratorio tenemos evidencia de autocorrelación espacial.

Para ajustar los modelos espaciales de regresión CAR y SAR, la matriz de vecindad $\boldsymbol{W}$ de las provincias en Perú fue construída asumiendo que dos provincias son vecinas si comparten algún límite geográfico. Según la matriz de vecindad de la Figura 5.3 se muestra que hay una cantidad moderada de vecindades, indicando una distribución razonable de conexiones entre provincias. Cada cuadrado rojo representa que dos provincias son vecinas.

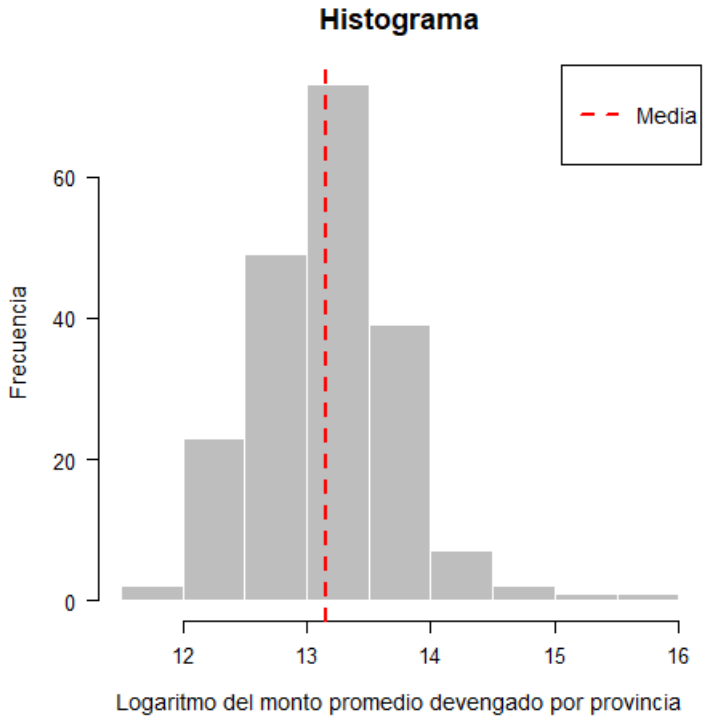



Figura 5.1: Histograma del logaritmo del monto devengado promedio en inversión pública.

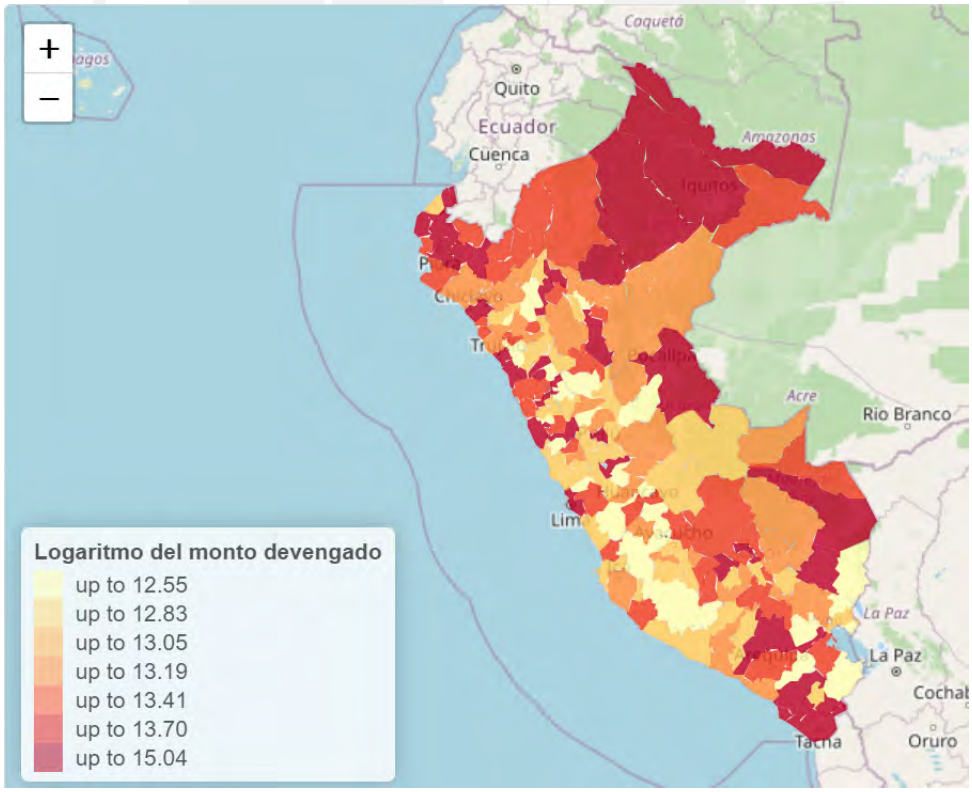



Figura 5.2: Mapa del monto devengado promedio en inversión pública a nivel de provincias en escala logarítmica.

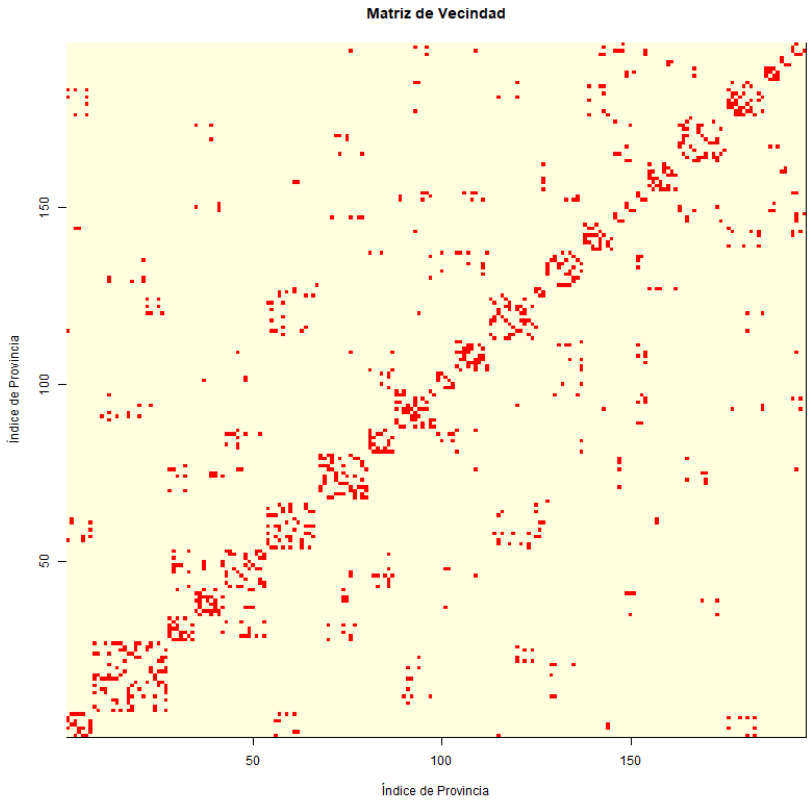



Figura 5.3: Matriz de vecindad de provincias en Perú.

### 5.2.1. Medidas de asociación espacial

En esta sección se describe la dependencia espacial encontrada a partir de estadísticas que nos resumen la autocorrelación espacial de la variable dependiente el logaritmo del monto devengado en inversión pública entre las provincias del Perú. Estas son el índice de I de Moran y la C de Geary. Las estadísticas mencionadas anteriormente miden la dependencia espacial, la cual supone la existencia de correlación entre los valores de la variable respuesta en provincias próximas entre sí.

Con el test de I de Moran se busca contrastar la hipótesis nula de distribución espacial aleatoria, comparando los valores de cada provincia del Perú. El cuadro 5.2 muestra los valores de las estadísticas mencionadas y valor-p del índice I de Moran. El valor del índice de Moran de 0.219 indica una autocorrelación espacial positiva leve en los datos. Esto significa que hay cierta tendencia de que las provincias cercanas tengan valores similares de montos de inversión, pero esta tendencia no es fuerte. Los resultados encontrados confirman evidencia estadística para afirmar que la variable de estudio ($Y_i$, monto promedio devengado en escala logarítmica, posee correlación espacial a un nivel de significancia del 5 %, dado que el valor-p es inferior a 0.05 rechazamos la hipótesis nula.

Cuadro 5.2: Test I de Moran.

| I-Moran | valor-p |
|---|---|
| 0.219 | 3.521e-07 |

El segundo estadístico estimado es el C de Geary, el resultado obtenido es menor a uno, ver cuadro 5.3. Cabe mencionar que el valor de C nunca es negativo y valores pequeños (entre 0 y 1) indican asociación espacial positiva. En particular, la estadística sugiere que hay una tendencia moderada a que las provincias vecinas tengan montos devengados promedios más similares entre sí de lo que se esperaría bajo una distribución aleatoria. Sin embargo, dado que el valor no está extremadamente cerca de cero, la tendencia no es muy fuerte.

Cuadro 5.3: Test C de Geary.

| C de Geary | valor-p |
|---|---|
| 0.757 | 4.852e-07 |

De los test de medidas de asociación se muestran indicaciones sobre la naturaleza de la autocorrelación espacial en los datos para $Y_i$. Estos datos pueden tener una tendencia general hacia la autocorrelación espacial.

En conclusión los tests de asociación espacial realizados indican que existe una autocorrelación espacial positiva en los datos de monto devengado en escala logarítmica $Y_i$. Es decir, las provincias con valores similares de $Y_i$ tienden a estar espacialmente cerca unas de otras. Este agrupamiento espacial de valores similares puede ser debido a factores geográficos, económicos, sociales u otros que afectan a las provincias de manera similar.

## 5.3. Descripción de las covariables

Se tiene un total de 16 covariables recolectadas principalmente del SNPMGI. En el cuadro 5.4 se muestra la estructura de datos que se ha utilizado.

La variable “Cartera Priorizada” denota el porcentaje de inversiones designadas como prioritarias de seguimiento por el Ministerio de Economía y Finanzas (MEF). Como podemos observar en la Figura 5.4, el histograma para esta covariable revela un sesgo positivo, indicando que una minoría de las inversiones ha sido clasificada como prioritaria. Adicionalmente, la visualización en el mapa resalta las provincias peruanas con los niveles más altos de esta variable, facilitando la identificación geográfica de las regiones donde se concentran estas inversiones.

La covariable “Inversión Gobierno Local” se refiere al porcentaje de inversiones ejecutadas directamente por entidades de gobierno local. El histograma asociado a esta covariable,

Cuadro 5.4: Descripción de las covariables.

| CAMPO | DEFINICIÓN | TIPO |
|---|---|---|
| Latitud | Georeferenciación de la inversión | Continua |
| Longitud | Georeferenciación de la inversión | Continua |
| Saneamiento | Inversión es de la tipología de saneamiento | Binaria |
| Tiene Formato N° F12B | Inversión tiene Formato 12 B: Seguimiento a la Ejecución de Inversiones | Binaria |
| Tiene avance físico | Inversión presenta avance físico de ejecución obra | Binaria |
| Beneficiario REI 2022 | Inversión cuenta con Reconocimiento de ejecución de inversiones x MEF | Binaria |
| Cartera priorizada | Inversión perteneció a cartera priorizada x MEF | Binaria |
| Tiene expediente técnico | Inversión tiene expediente técnico | Binaria |
| Inversión Gobierno Local | Responsable de ejecución de inversión Gobierno Local | Binaria |
| Transporte y Comunicaciones | Inversión es de la tipología de transporte y comunicaciones | Binaria |
| Tiene Formato N° 09 | La inversión tiene Formato N° 09: Registro de cierre de Inversión | Binaria |
| IDH | Índice de Desarrollo Humano para 2019 provincial (BID) | Continua |
| Índice de corrupción | Nivel de corrupción provincial, Contraloría General de República | Binaria |
| Tiempo programado ejecución | Menor a 1 año entre la fecha de inicio y fecha de fin de ejecución física | Binaria |
| Ratio sobrecosto | Mide la diferencia entre el monto viable y el costo actualizado: mayor 15 % | Continua |
| Costo de inversión entre 50 y 300 millones | El costo de inversión actualizado | Continua |

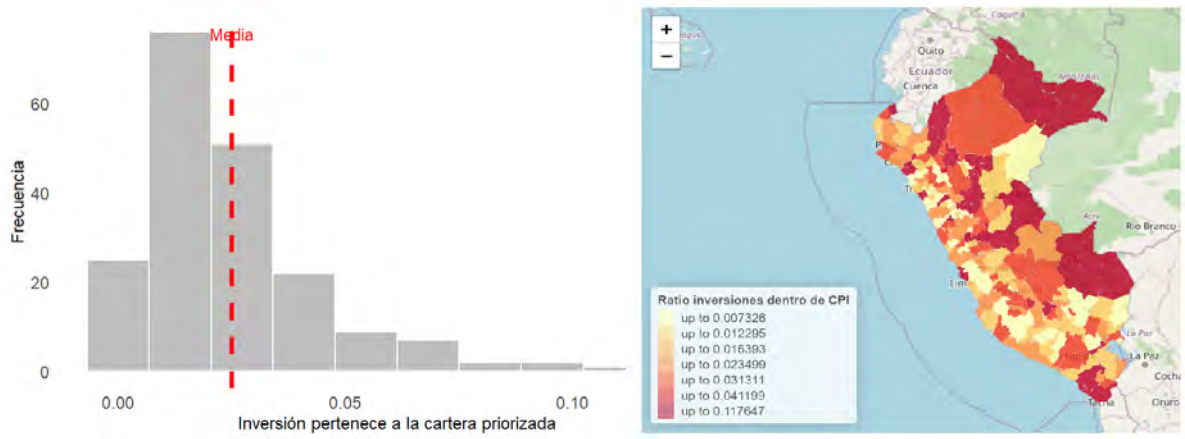



Figura 5.4: Inversiones públicas que estuvieron en Cartera priorizada en 2022.

presentado en la Figura 5.5, muestra una distribución sesgada hacia la izquierda, indicando que una proporción significativa de provincias ha realizado un alto porcentaje de estas inversiones. Por otro lado, el mapa ilustra visualmente las provincias con mayores ratios de inversión por parte de gobiernos locales, destacando las áreas con mayor actividad en este ámbito.

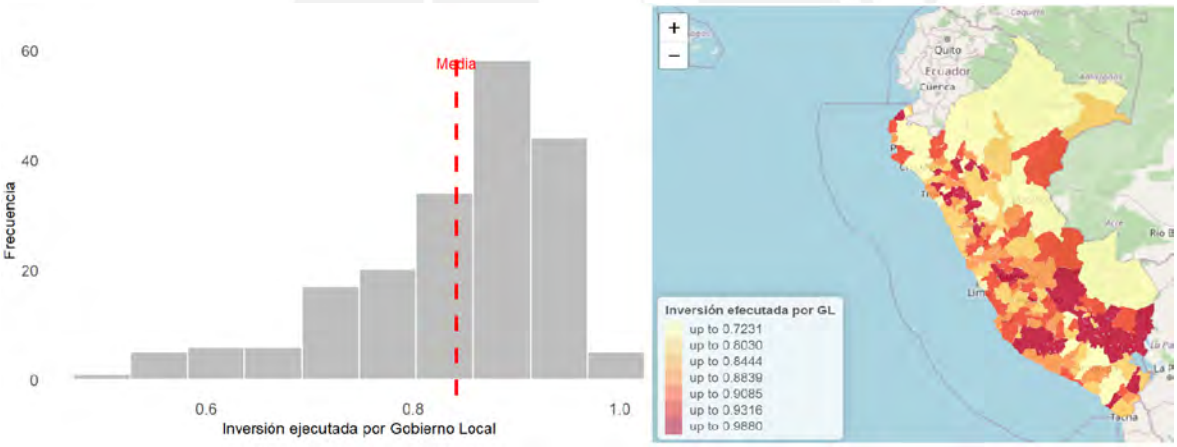



Figura 5.5: Inversiones públicas ejecutado por una entidad del Gobierno Local - GL 2022.

Otra de las covariables destacadas en el análisis es el "Avance físico de las inversiones", que indica el porcentaje de inversiones que han reportado un avance físico sustancial. El histograma de esta variable, como se observa en la Figura 5.6, muestra una distribución simétrica, lo que sugiere que, aunque variado, un número considerable de provincias ha informado de avances significativos. El mapa adjunto proporciona una visualización geográfica, destacando las provincias del Perú con mayores proporciones de este indicador.

La covariable "Indice de Corrupción" corresponde al Índice de Corrupción e Inconducta Funcional (INCO), una medida desarrollada por la Contraloría General de la República del Perú para estimar los niveles de corrupción en las provincias del país. El INCO se expresa en

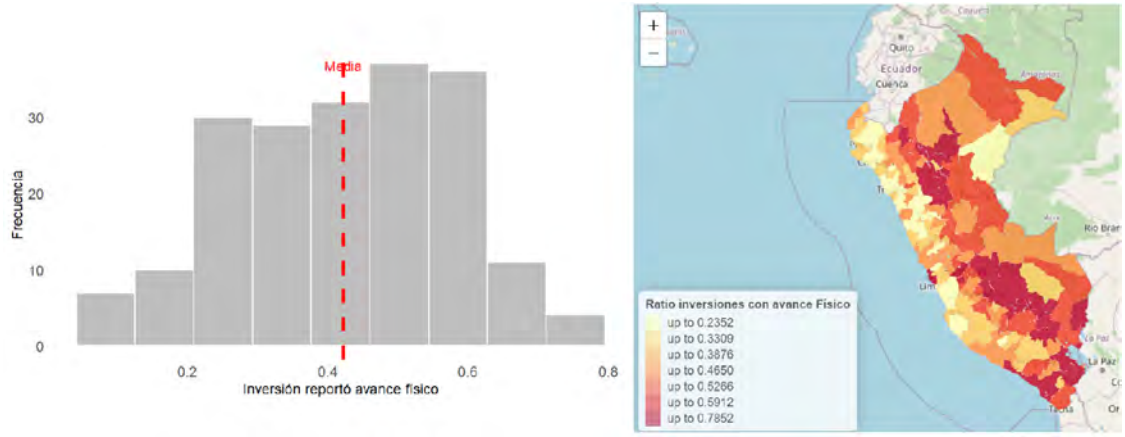



Figura 5.6: Inversiones públicas que reportaron avance físico en 2022.

una escala de 0 a 100 puntos, donde un valor más alto indica mayor percepción de corrupción. Esta covariable muestra una notable variabilidad en su distribución, como se observa en la Figura 5.7. El histograma de esta covariable revela una distribución asimétrica hacia la derecha, indicando que un número significativo de provincias presenta niveles de corrupción menores a la media (en comparación con otras provincias del país, los niveles de corrupción percibida son menores). Simultáneamente, el mapa de distribución espacial del Índice de Corrupción permite visualizar cómo se distribuyen geográficamente estos niveles de corrupción. Las áreas con tonos más oscuros representan provincias con índices más altos de corrupción.

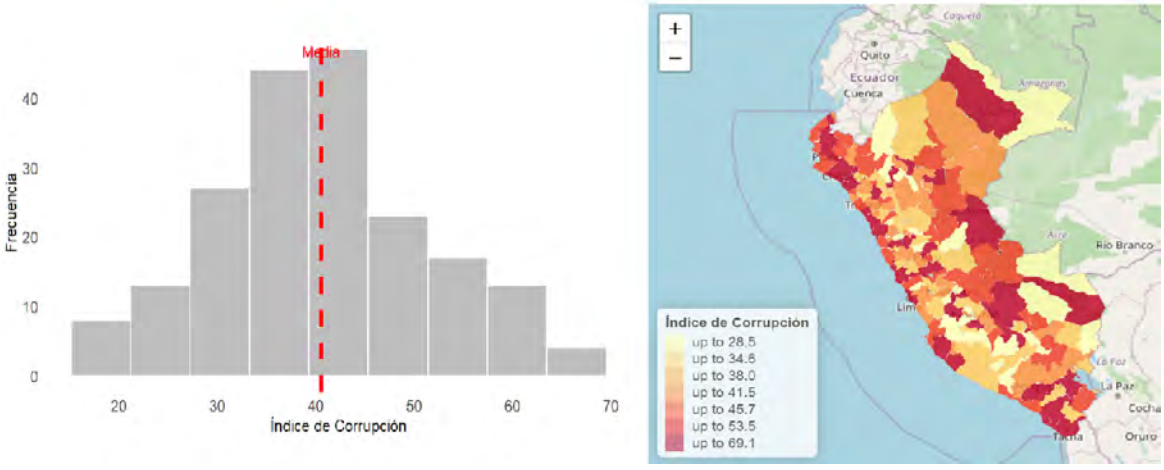



Figura 5.7: Indice de Corrupción a nivel provincial en 2022.

La matriz de correlación de Pearson en la Figura 5.8 muestra los coeficientes de correlación entre varias variables clave en el estudio del monto promedio devengado en escala logarítmica. El coeficiente de -0.57 entre "Inversión Gobierno Local" y "Cartera priorizada" sugiere una correlación negativa moderada, lo que implica que a medida que el nivel de gobierno aumenta, la probabilidad de que las inversiones pertenezcan a la cartera priorizada

disminuye. En contraste, el coeficiente de 0.39 entre "Cartera priorizada" y "Monto devengado" indica una correlación positiva, sugiriendo que las inversiones dentro de la cartera priorizada tienden a tener montos devengados más altos. Estos coeficientes ayudan a identificar posibles dependencias entre las variables. Es importante mencionar, que mientras que las correlaciones pueden informar sobre las relaciones entre variables, no sustituyen el análisis espacial detallado que considera la autocorrelación espacial y otras características únicas de los datos espaciales.

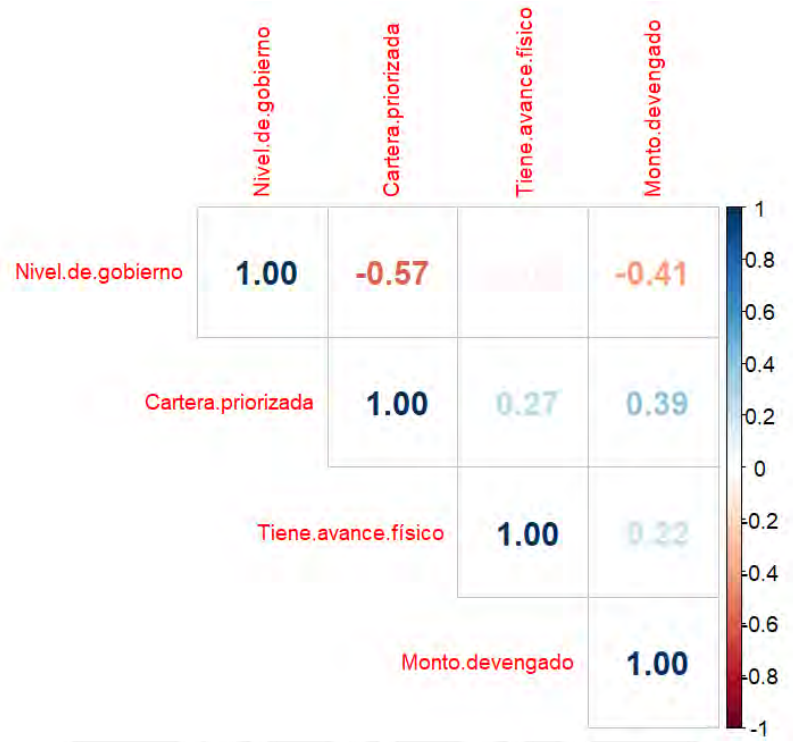



Figura 5.8: Matriz de correlación de Pearson entre las variables de estudio.

## 5.4. Modelos aplicados

En esta sección, se discuten los resultados obtenidos a partir de la implementación de las dos propuestas de modelos descritas en el Capítulo 3. Posteriormente, se realiza una comparativa entre ambas, empleando el criterio del error cuadrático medio (RMSE) para determinar cuál de los dos modelos ofrece un mejor ajuste.

En particular, $Y_i$ representa el logaritmo del monto devengado promedio en inversión pública en la provincia $i = 1, \ldots, n = 196$. A continuación se describen los modelos ajustados.

### 5.4.1. Modelo CAR

Se asume que el vector aleatorio

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \sim N\left(\boldsymbol{X\beta}, \boldsymbol{Q}^{-1}\right)$$

donde $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)$, $\widetilde{\boldsymbol{W}}$ es la matriz de vecindad estandarizada, $(\boldsymbol{D}_W)_{ii} = \omega_{i+}$. Los parámetros del modelo están definido por el vector $\boldsymbol{\theta} = \left(\beta, \tau^2, \rho\right)$.

- Modelo CAR-1 con tres covariables: Inversión gobierno local, Cartera priorizada, e Indice de corrupción.
- Modelo CAR-2 con cuatro covariables: Inversión gobierno local, Cartera priorizada, Tiene avance físico, e Indice de corrupción.

### 5.4.2. Modelo SAR

Se asume que el vector aleatorio

$$\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T \sim N\left(\boldsymbol{R}^{-1}\boldsymbol{X\beta}, \boldsymbol{Q}^{-1}\right),$$

donde $\boldsymbol{R} = \left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)$ y $\boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)\left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^\top$, $\widetilde{\boldsymbol{W}}$ es la matriz de vecindad estandarizada.

- Modelo SAR-1 con tres covariables: Inversión gobierno local, Cartera priorizada, e Indice de corrupción.
- Modelo SAR-2 con cuatro covariables: Inversión gobierno local, Cartera priorizada, Tiene avance físico, e Indice de corrupción.

## 5.5. Resultados

Los parámetros a estimar de todos los modelos están definidos por el vector $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \tau^2, \rho\right)$. Se ajustaron los modelos y como criterios de selección de modelos se usó el criterio de información de Akaike (AIC) y el criterio de información bayesiano (BIC). También se calculó la raíz del error cuadrático medio de estimación (RMSE) definido de la siguiente manera :

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{196}\left(y_i - \hat{y}_i\right)^2}{196}},$$

donde $y_i$ es el el logaritmo del monto promedio devengado para la i-ésima provincia, e $\hat{y}_i$ es su estimación, es decir la media de $Y_i$ de acuerdo a cada modelo.

Los resultados de estas medidas se muestran en el cuadro 5.5. Según estos resultados, el modelo SAR-2 fue el que mejor se ajustó a los datos pues tiene los menores valores de AIC, BIC y RMSE, seguido del modelo CAR-2. Según el criterio de BIC el mejor modelo resultó ser el modelo SAR-2, seguido del modelo SAR-1. Además, es notable que el modelo CAR-2, con un RMSE de 0.459, exhibe una precisión comparable al del modelo SAR-2, lo cual resalta su capacidad para proporcionar una estimación precisa de los valores observados, posiblemente debido a una mejor gestión de la variabilidad de los datos y a una especificación más acertada en términos de las variables incluidas y la estructura del modelo.

Cuadro 5.5: Criterios de selección de modelos.

| | Modelo CAR-1 | Modelo CAR-2 | Modelo SAR-1 | Modelo SAR-2 |
|---|---|---|---|---|
| AIC | 259.4401 | 253.4729 | 254.695 | **250.622** |
| BIC | 279.1088 | 276.419 | 274.363 | **273.569** |
| RMSE | 0.464 | **0.459** | 0.462 | **0.458** |

### 5.5.1. Estimación de parámetros

Los valores estimados de los parámetros de los tres mejores modelos seleccionados se muestran en el cuadro 5.6. Los modelos CAR-2 y SAR-2 incluyen cuatro covariables. A diferencia del modelo SAR-1, estos modelos incorporan la variable "Tiene avance físico", permitiendo una evaluación más detallada de cómo las inversiones que muestran progreso tangible en términos físicos influyen en el logaritmo del monto devengado promedio en inversión pública por provincia.

En el Cuadro 5.6 se muestran las estimaciones puntuales para el modelo SAR con tres covariables (modelo SAR-1) y las estimaciones para el modelo con cuatro covariables (modelo SAR-2). En relación a la significancia estadística de las covariables, la inversión de gobierno local es estadísticamente significativa en ambos modelos, aunque el Modelo SAR-2 presenta un valor-p más bajo, sugiriendo una menor probabilidad de error. La cartera priorizada es significativa en ambos modelos; sin embargo, el Modelo SAR-2 muestra un coeficiente menor. El índice de corrupción es altamente significativo en ambos modelos, indicando un efecto robusto en las predicciones. En relación al parámetro de autocorrelación espacial $\rho$ ambos modelos reconocen la importancia de modelar la autocorrelación espacial pues su valor estimado es significativamente diferente de cero. Mientras que el valor de $\tau^2$ al ser mucho menor que uno, indica evidencia de varianza espacial $(1/\tau^2)$ mayor que uno.

Según los resultados del Cuadro 5.6 correspondientes al modelo CAR-2, todas las variables incluidas muestran significancia estadística, indicando que influyen de manera confiable en la variable de respuesta, el logaritmo del monto devengado promedio. El valor estimado del parámetro de autocorrelación espacial $\rho$ es ligeramente mayor que en el modelo CAR-1 (ver anexo D). Además se observa que el modelo CAR tiende a estimar con un valor más alto el parámetro de autocorrelación espacial. Sin embargo los modelos SAR tienen menor valor de estimación para $\tau^2$ indicando que la varianza espacial ( $1/\tau^2$) es mayor para los modelos SAR, es decir la variabilidad espacial es capturada mejor por los modelos SAR.

Cuadro 5.6: Resultados de las estimaciones puntuales: media, desviación estándar e intervalos de confianza (IC) al 95 % de los modelos CAR-2, SAR-1 y SAR-2.

| Parámetro | Estimación | Desv. Estándar | valor-p | IC (95 %) |
|---|---|---|---|---|
| **Modelo CAR-2** | | | | |
| Intercepto | 13.633 | 0.375 | < 2.2e-16 *** | $(12.897, 14.370)$ |
| Inversión gobierno local | -1.054 | 0.394 | 0.0074951 ** | $(-1.827, -0.281)$ |
| Cartera priorizada | 5.434 | 1.985 | 0.0061975 ** | $(1.542, 9.325)$ |
| Tiene avance físico | 0.732 | 0.262 | 0.0052375 ** | $(0.218, 1.246)$ |
| Índice de corrupción | -0.321 | 0.087 | 0.0002148 *** | $(-0.492, -0.151)$ |
| $\rho$ | 0.642 | 0.145 | 1.01e-05 *** | $(0.357, 0.927)$ |
| $\tau^2$ | 0.943 | 0.098 | < 2.2e-16 *** | $(0.749, 1.136)$ |
| **Modelo SAR-1** | | | | |
| Intercepto | 10.034 | 1.190 | < 2.2e-16 *** | (7.700, 12.367) |
| inv gobierno local | -1.170 | 0.389 | 0.0026833 ** | (-1.933, -0.406) |
| cartera priorizada | 6.450 | 1.886 | 0.0006268 *** | (2.753, 10.147) |
| indice corrupción | -0.298 | 0.085 | 0.0004876 *** | (-0.466, -0.130) |
| $\rho$ | 0.302 | 0.083 | 0.0002789 *** | (0.139, 0.465) |
| $\tau^2$ | 0.198 | 0.020 | < 2.2e-16 *** | (0.158, 0.237) |
| **Modelo SAR-2** | | | | |
| Intercepto | 9.945 | 1.175 | < 2.2e-16 *** | (7.641, 12.250) |
| inv gobierno local | -1.281 | 0.386 | 0.0009073 *** | (-2.037, -0.524) |
| cartera priorizada | 5.061 | 1.931 | 0.0087596 ** | (1.277, 8.846) |
| tiene avance físico | 0.521 | 0.210 | 0.0130224 * | (0.109, 0.933) |
| indice corrupción | -0.297 | 0.084 | 0.0004190 *** | (-0.462, -0.132) |
| $\rho$ | 0.302 | 0.082 | 0.0002406 *** | (0.140, 0.463) |
| $\tau^2$ | 0.192 | 0.019 | < 2.2e-16 *** | (0.153, 0.230) |

Como el modelo SAR-2 fue el mejor en términos del AIC, BIC y RMSE por ello se interpretan los parámetros solo para este modelo. La inversión del gobierno local muestra un efecto negativo (-1.281) sobre el logaritmo del monto devengado, sugiriendo que un aumento en la gestión local de las inversiones podría asociarse con una reducción en el monto devengado. En contraste, la cartera priorizada exhibe un impacto positivo considerable (5.061), indicando que las inversiones catalogadas como prioritarias tienden a reportar montos devengados más

altos. El avance físico también muestra un efecto positivo (0.521), mientras que el índice de corrupción presenta un efecto negativo (-0.297).

### 5.5.2. Estimación de la variable respuesta

Con los valores estimados de los parámetros, se calcula el valor estimado de la variable respuesta $\hat{y}_i = E(Y_i)$ según cada modelo. La Figura 5.9 muestra la relación entre los valores observados y los valores estimados de la variable respuesta $Y_i$ utilizando los modelos CAR-2, SAR-1 y SAR-2. Los resultados muestran un ajuste razonable a los datos observados aunque como trabajo futuro se propone usar más covariables o usar la distribución de la variable respuesta sin transformaciones para mejorar dicha estimación.

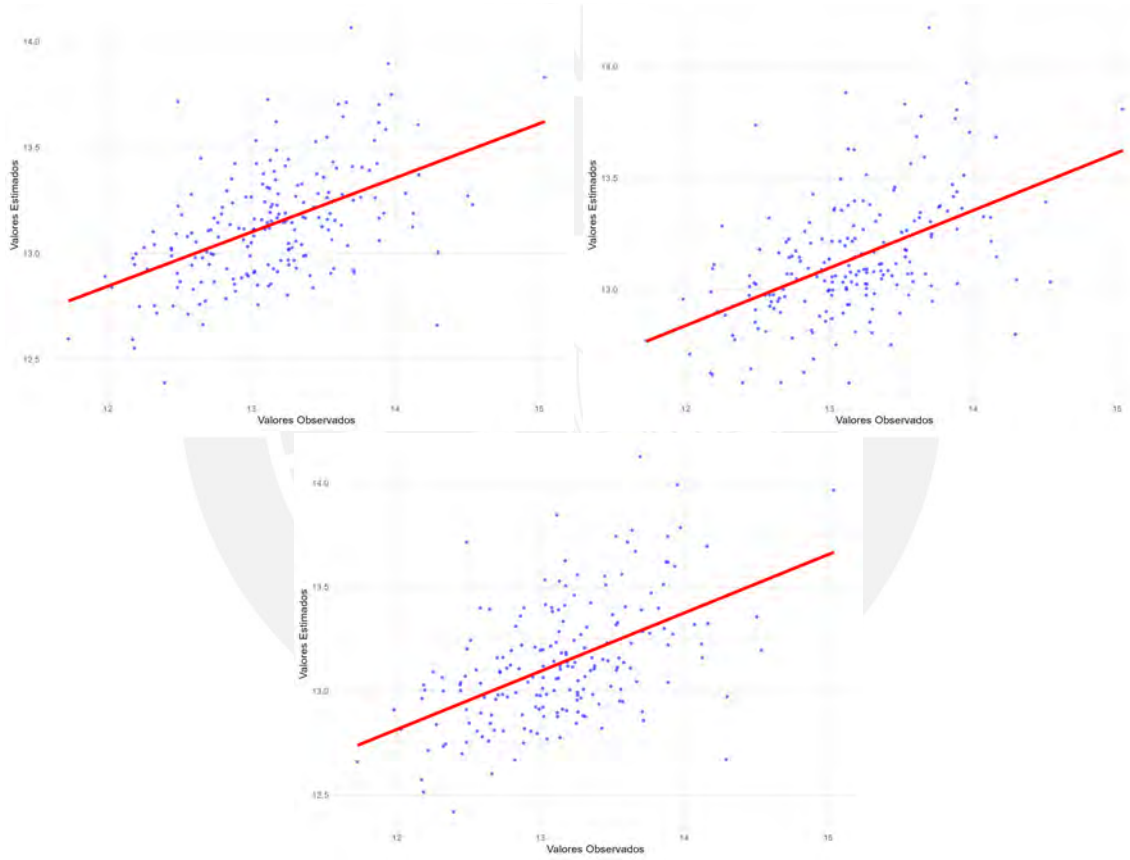
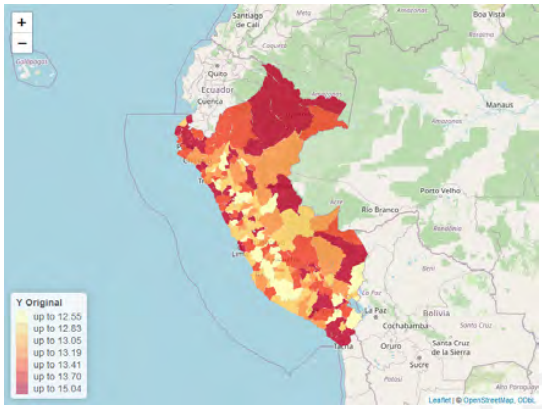


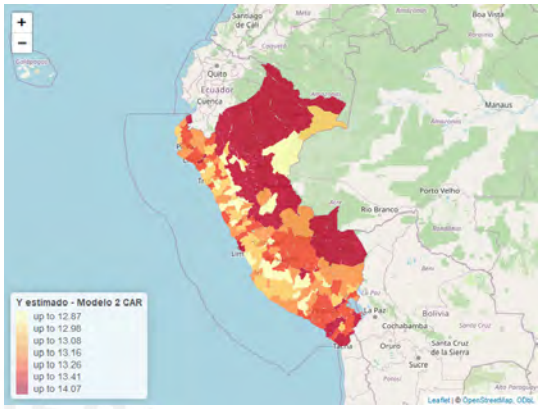

Figura 5.9: Gráfica de $Y_i$ versus las estimaciones de $Y_i$ con el Modelo CAR-2 (arriba izquierda), SAR-1 (arriba derecha) y SAR-2 (abajo). La línea roja representa $x = y$.

La Figura 5.10 presenta mapas de estimaciones de la variable $Y_i$ para los modelos ajustados. Estos mapas contrastan los valores originales observados de $Y_i$ con los valores estimados $\hat{Y}_i = E(Y_i)$ para cada modelo, con el objetivo de comparar la eficacia de la distribución espacial de la variable de interés según cada modelo ajustado. En general los mapas reflejan adecuadamente las estimaciones de $Y_i$, estimado valores altos del logaritmo del monto devengado en varias provincias de la selva y Lima metropolitana. Mientras que en la sierra se

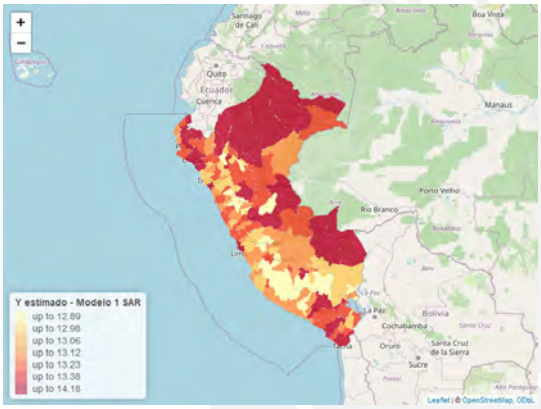
observan más provincias donde el logaritmo del monto devengado es menor. Estos resultados sugieren que los modelos son capaces de reflejar la estructura espacial de los datos de manera efectiva.



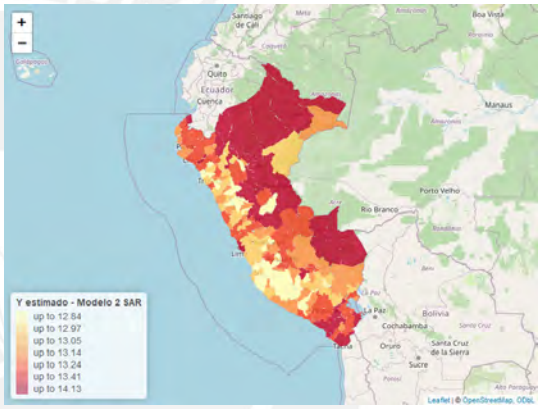

(a) $Y_i$ Original




(b) $Y_i$ estimado - Modelo CAR-2




(c) $Y_i$ estimado - Modelo SAR-1



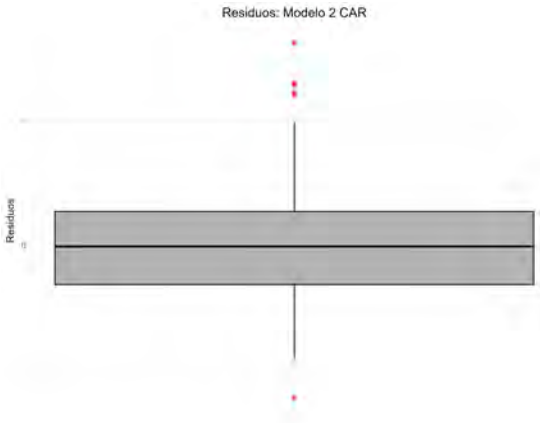

(d) $Y_i$ estimado - Modelo SAR-2

Figura 5.10: Mapa de valores observados $Y_i$ (a) y sus estimaciones obtenidas con los modelos CAR-2 (b) , SAR-1 (c), SAR-2 (d).

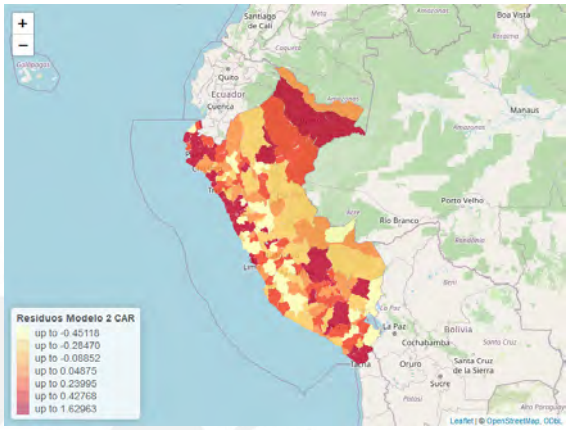
### 5.5.3. Análisis de residuos

La Figura 5.11 muestra la representación de los residuos de los modelos ajustados, utilizando dos métodos de visualización: diagramas de cajas y mapas espaciales. Estos gráficos tienen el propósito de evaluar la distribución y la dispersión de los residuos generados por cada modelo para determinar su ajuste y efectividad en modelar la variable dependiente en el contexto espacial. Los residuos en los diagramas de cajas (Figura 5.11) se concentran alrededor de cero, sin mostrar muchos residuos atípicos, lo que es un indicador positivo de la estimación del logaritmo del monto devengado por los modelos. En base al análisis de los mapas de residuos de la Figura 5.11, el Modelo SAR-2 parece ofrecer un mejor ajuste en términos de distribución más uniforme y menos variabilidad en los residuos. Los residuos del Modelo SAR-2 también muestran una distribución más dispersa y menos agrupada es-

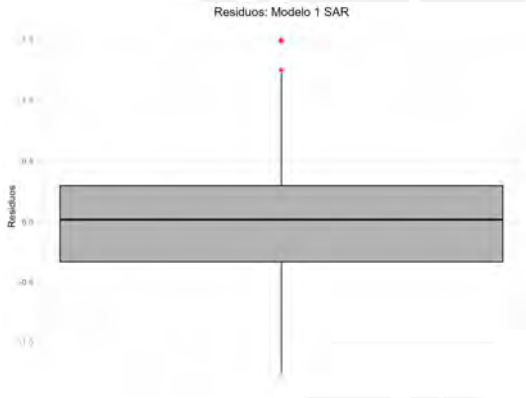
pacialmente, lo que indica que captura mejor la autocorrelación espacial en los datos. Esto sugiere que el Modelo SAR-2 podría ser preferible para utilizar en la estimación y análisis de datos donde la estructura espacial es una consideración importante. La presencia de residuos atípicos en los modelos resalta la necesidad de una evaluación más profunda, posiblemente ajustando los modelos o reconsiderando algunas de las covariables incluidas en el análisis exploratorio.



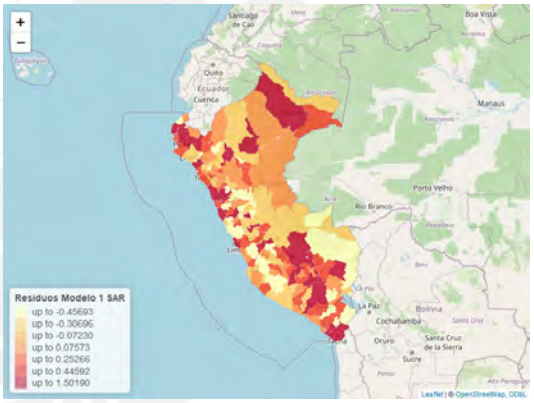

(a) Diagrama de cajas CAR-2



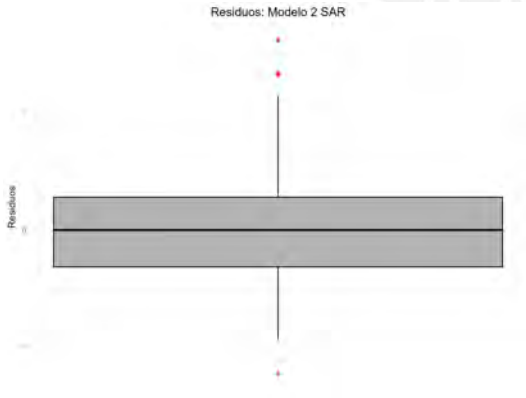

(b) Residuos est. - Modelo CAR-2



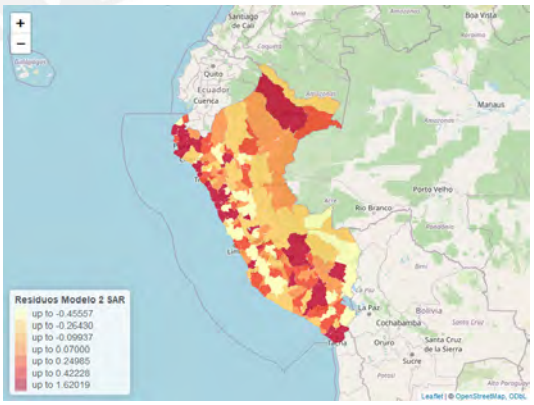

(c) Diagrama de cajas SAR-1




(d) Residuos est. - Modelo SAR-1




(e) Diagrama de cajas SAR-2




(f) Residuos est. - Modelo SAR-2

Figura 5.11: Representación de los residuos de los modelo CAR-2 (primera fila), SAR-1 (segunda fila) y SAR-2 (tercera fila) mediante un diagrama de cajas (izquierda) y un mapa (derecha).

# Capítulo 6

# Conclusiones

## 6.1. Conclusiones

- Al aplicar medidas de asociación espacial, como el índice de Moran y el test de Geary en los datos del logaritmo de monto devengado en inversión pública en el 2022, los resultados indicaron una leve autocorrelación espacial positiva (índice de Moran de 0.219), sugeriendo que las provincias vecinas tienden a tener valores de monto devengado similares. Este hallazgo subraya la importancia de utilizar modelos que capturen la dependencia espacial.

- Se realizó una comparación entre los modelos CAR y SAR con covariables. Ambos modelos demostraron ser eficaces para manejar la autocorrelación espacial en los datos. Sin embargo, es importante resaltar que en los modelos CAR la autocorrelación espacial ($\rho$) resultó ser mayor a 0.5, indicando una posible sobreestimación de la dependencia espacial. Asimismo, el modelo SAR mostró un ajuste general similar al modelo CAR, sin embargo este último obtuvo valores más bajos en los criterios de información como AIC y BIC y RSME, lo que sugiere que el modelo SAR, en particular, el Modelo SAR-2 podría ser preferido para la interpretación y estimación en contextos similares.

- Las covariables consideradas en el modelo de mejor ajuste (Modelo CAR-2) mantuvieron su significancia estadística en los diferentes modelos, con la inversión del gobierno local y el índice de corrupción mostrando efectos negativos, mientras que la cartera priorizada y el avance físico mostraron un impacto positivo.

- Al analizar los intervalos de confianza de las covariables, podemos concluir que la estimación del parámetro para el índice de corrupción es la más precisa y confiable, mientras que la estimación para la cartera priorizada es la menos precisa debido a la

mayor amplitud del intervalo de confianza. Esta información es crucial para entender la estabilidad y la significancia de los predictores en el modelo.

- Los residuos también fueron evaluados para verificar la adecuación de los modelos a los datos.

- Las visualizaciones a través de mapas proporcionaron una representación clara de cómo ambos modelos estimaban los valores de $Y_i$ en comparación con los valores observados. Esto subraya la utilidad de los modelos espaciales en la interpretación de la distribución espacial del gasto público.

## 6.2. Sugerencias para investigaciones futuras

Se plantea varias mejoras para la evaluación de los modelos CAR y SAR utilizados en el análisis de la distribución del gasto devengado de la inversión pública a nivel provincial en Perú. Algunas mejoras sugeridas se mencionan a continuación:

- Se sugiere incluir más variables socioeconómicas y políticas que puedan tener un impacto en la inversión pública para capturar mejor la complejidad de los factores que influyen en el gasto devengado.

- Se recomienda explorar diferentes especificaciones de la matriz de pesos espaciales para mejorar la incorporación de la autocorrelación espacial en los modelos, lo que podría mejorar la precisión de las estimaciones y proporcionar una visión más detallada de las dependencias espaciales. En Earnest et al. (2007), se ha evaluado el efecto de diferentes matrices de peso de vecindario en los modelos CAR. Para modelos bayesianos y su impacto en la inferencia matriz de pesos se puede consultar Duncan et al. (2017).

  En este estudio, se seleccionó el estilo de ponderación “Binaria” por ser simple y efectiva, pero no tiene en cuenta diferencias de tamaño o forma entre las unidades. Al respecto se realizó un re-escalamiento de la matriz para controlar la influencia de unidades con muchos o pocos vecinos.

# Apéndice A

## Modelo CAR: Esperanza Condicional $E\left(Y_i|Y_{-i}\right)$

Para calcular la esperanza condicional $E\left(Y_i|Y_{-i}\right)$ en un modelo de regresión espacial condicional autoregresivo (CAR) , necesitamos usar la estructura del modelo CAR junto con la teoría de distribuciones condicionales en modelos multivariados normales. En un modelo CAR, como el presentado, la variable $Y$ sigue una distribución normal multivariada $\boldsymbol{Y} \sim N\left(\boldsymbol{X\beta}, \boldsymbol{\Sigma}\right)$ donde $\boldsymbol{\Sigma}$ es la matriz de covarianza que depende de la estructura espacial del modelo.

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q} = \frac{1}{\tau^2}\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)$$

La esperanza condicional $E\left(Y_i|Y_{-i}\right)$ se expresa como

$$E\left(Y_i|Y_{-i}\right) = \mu_i + \boldsymbol{\Sigma}_{i,-i}\boldsymbol{\Sigma}_{-i,-i}^{-1}\left(\boldsymbol{Y}_{-i} - \boldsymbol{\mu_{-i}}\right),$$

donde

$\mu_i$ es la i-ésima entrada del vector de medias $\boldsymbol{X\beta}$,

$\boldsymbol{\mu_{-i}}$ es el vector de medias sin la i-ésima entrada,

$\boldsymbol{\Sigma_{i,-i}}$ es la fila $i$ de la matriz $\boldsymbol{\Sigma}$ excluyendo la columna $i$,

$\boldsymbol{\Sigma_{-i,-i}}$ es la matriz $\boldsymbol{\Sigma}$ excluyendo la fila $i$ y la columna $i$,

$\boldsymbol{Y}_{-i}$ es el vector de observaciones excluyendo $Y_i$.

Luego, la media condicional de $Y_i$ estimada es dada por:

$$\hat{\mu}_{i|-i} = X_i^\top\hat{\beta} + \hat{\boldsymbol{\Sigma}}_{i,-i}\hat{\boldsymbol{\Sigma}}_{-i,-i}^{-1}\left(\boldsymbol{Y_{-i}} - \boldsymbol{X}_{-i}^\top\hat{\boldsymbol{\beta}}\right).$$

En el modelo de regresión condicional autoregresivo (CAR), la predicción de valores en una nueva localización no requiere directamente la estimación de la esperanza condicional de manera explícita como en otros modelos estadísticos.Esto se debe a cómo está estructurado

el modelo CAR y la forma en que maneja la dependencia espacial.

El modelo ya incorpora la estructura de dependencia mediante la inclusión directa de los valores de los vecinos en la fórmula del modelo, no es necesario realizar un paso adicional para estimar la esperanza condicional de $Y_i$ dado $Y_j$ .

Esta característica simplifica la implementación del modelo y la hace computacionalmente eficiente, especialmente en escenarios donde la red de dependencias espaciales es compleja y extensa.

# Apéndice B

## Apéndice B: Modelo SAR: Estimación de $\boldsymbol{Y}$

Consideramos la estimación en el modelo SAR autorregresivo espacial. Dada una matriz de peso espacial $\boldsymbol{W}$ y variables exógenas $\boldsymbol{X}$, este modelo se puede escribir:

$$\boldsymbol{Y} = \rho\widetilde{\boldsymbol{W}}\boldsymbol{Y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} = \left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{X}\boldsymbol{\beta} + \left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{\epsilon}.$$

Como $\boldsymbol{\epsilon} \sim N(0, \tau^2 I)$ entonces la media condicional de $\boldsymbol{Y}$ en este modelo viene dada por

$$\boldsymbol{\mu} = E\left(\boldsymbol{Y}\right) = \left(\boldsymbol{I} - \rho\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{X}\boldsymbol{\beta}.$$

Luego el valor estimado de la media de $\boldsymbol{Y}$ es:

$$\hat{\boldsymbol{\mu}} = \left(\boldsymbol{I} - \hat{\rho}\widetilde{\boldsymbol{W}}\right)^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

ET LUX IN TENEBRIS LUCET
MCMXVII

# Apéndice C

## Apéndice C: Estudio de Simulación de Montecarlo

El propósito de utilizar la simulación de Monte Carlo es estimar la distribución de $\hat{Y}_i$ bajo el modelo de regresión condicional autoregresivo CAR de cuatro covariables (modelo CAR-2), teniendo en cuenta la incertidumbre y la autocorrelación espacial.

**Pasos para la Simulación de Monte Carlo en el Modelo CAR**

- Se establece una semilla para garantizar la reproducibilidad de las simulaciones.
- Se define 1000 simulaciones para obtener una estimación robusta de las predicciones de $\hat{Y}_i$
- Se extraen los coeficientes estimados del modelo ajustado modelo CAR-2. Esto incluye interceptos y coeficientes para cada covariable, además de los parámetros de la estructura espacial ($\rho$ y $\tau^2$).
- Se construye la matriz de diseño $X$ con las covariables de interés extraídas del conjunto de datos.
- Para cada simulación, se generan errores espaciales a partir de una distribución multivariante normal con media cero y una matriz de covarianza derivada de la inversa de $\boldsymbol{Q}$, donde $Q = \frac{1}{\tau^2}\left(\boldsymbol{D}_W - \rho\widetilde{\boldsymbol{W}}\right)$
- Para cada simulación, se calcula $\hat{Y}_i$ utilizando la combinación lineal de covariables y los errores espaciales simulados.
- Se calcula la media de las predicciones de todas las simulaciones para obtener una estimación estable de $\hat{Y}_i$.
- Se calculan los intervalos de confianza al 95 % para las estimaciones, proporcionando una medida de la incertidumbre asociada con $\hat{Y}_i$.

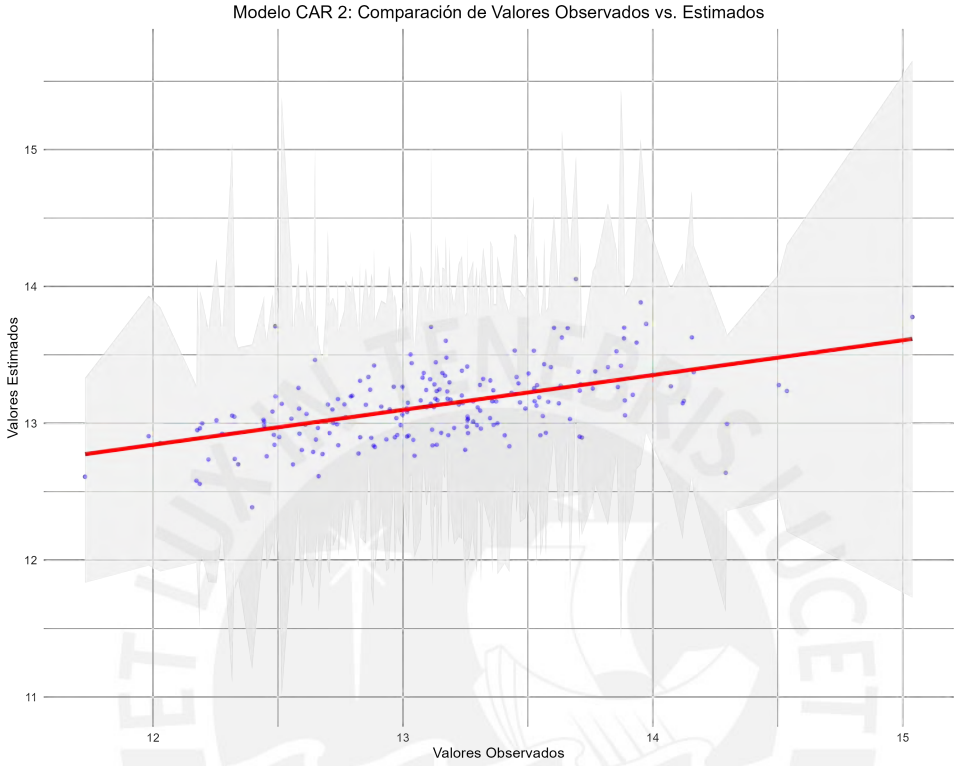



Figura 1: Gráfica de estimaciones de $Y_i$ con Modelo CAR-2.

Cuadro 1: Error del modelo predictivo Modelo CAR-2.

| | Modelo 2 |
|---|---|
| RMSE | 0.4588012 |

# Apéndice D

## Apéndice D: Modelo CAR-1 en aplicación

En el cuadro 2 se muestran las estimaciones puntuales para el modelo CAR-1 con tres covariables.

El análisis de las estimaciones puntuales en el Modelo CAR-1 revela que las variables *Inversión gobierno local*, *Cartera priorizada*, e *Índice de corrupción* poseen significancia estadística, lo que indica que son predictores relevantes del logaritmo del monto devengado promedio en inversión pública por provincia. Con respecto a los parámetros de autocorrelación espacial ($\rho$ y $\tau^2$), el valor estimado de la autocorrelación espacial igual a 0.6 indica que hay autocorrelación espacial entre los valores de $Y_i$ en provincias vecinas, mientras que el valor estimado de la varianza es aproximadamente 1 siendo significativamente diferente de cero, estos resultados subrayan la presencia de dependencia espacial en los datos y refuerza la validez del modelo en capturar la interacción entre observaciones cercanas geográficamente.

Cuadro 2: Resultados de las estimaciones puntuales: media, desviación estándar e intervalos de confianza (IC) al 95 % del modelo CAR-1.

| Parámetro | Estimación | Desv. Estándar | valor-p | IC (95 %) |
|---|---|---|---|---|
| Intercepto | 13.871 | 0.368 | < 2.2e-16 *** | (13.150, 14.592) |
| inv gobierno local | -1.006 | 0.399 | 0.0118 * | (-1.790, -0.222) |
| cartera priorizada | 6.888 | 1.954 | 0.0004 *** | (3.057, 10.719) |
| indice corrupción | -0.342 | 0.088 | 9.962e-05 *** | (-0.515, -0.170) |
| $\rho$ | 0.599 | 0.149 | 6.064e-05 *** | (0.306, 0.892) |
| $\tau^2$ | 0.989 | 0.103 | < 2.2e-16 *** | (0.787, 1.191) |

La inversión del gobierno local muestra un efecto negativo significativo con un coeficiente de -1.006, sugiriendo que un incremento en la gestión local de las inversiones puede estar relacionado con menores montos devengados. Por otro lado, la cartera priorizada, con un coeficiente de 6.888, tiene un impacto positivo considerable, indicando que las inversiones clasificadas como prioritarias tienden a asociarse con montos devengados más altos. El índice de corrupción, con un coeficiente de -0.342, revela un efecto negativo, donde niveles más altos

de corrupción se asocian con reducciones en los montos devengados.

La Figura 2 correspondiente al modelo CAR-1 sugiere que mientras el modelo maneja la dependencia espacial a un grado satisfactorio, podría haber una variabilidad en los residuos que indica la posible mejora en la especificación del modelo o en la inclusión de otras variables explicativas.

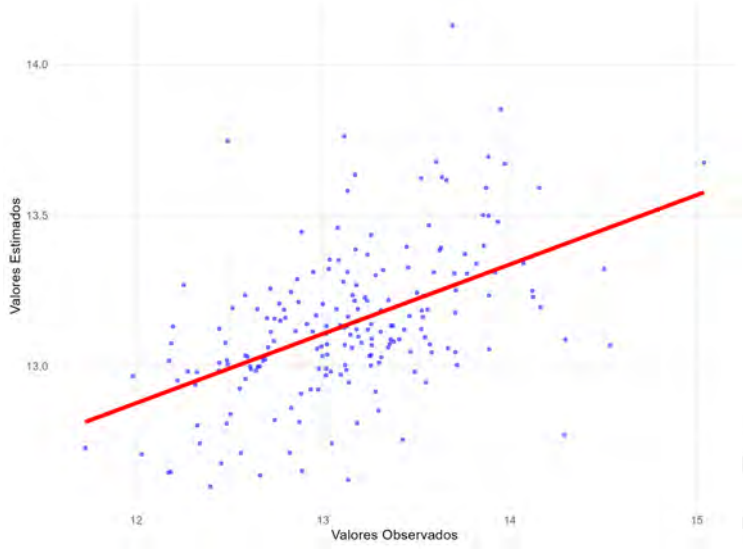
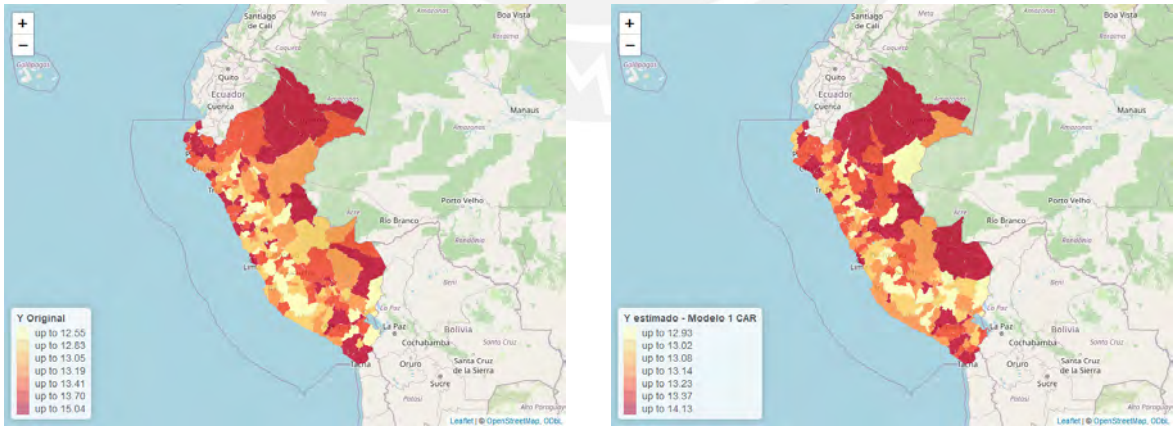



Figura 2: Gráfica de estimaciones de $Y_i$ con Modelo CAR-1. La línea roja representa $x = y$.

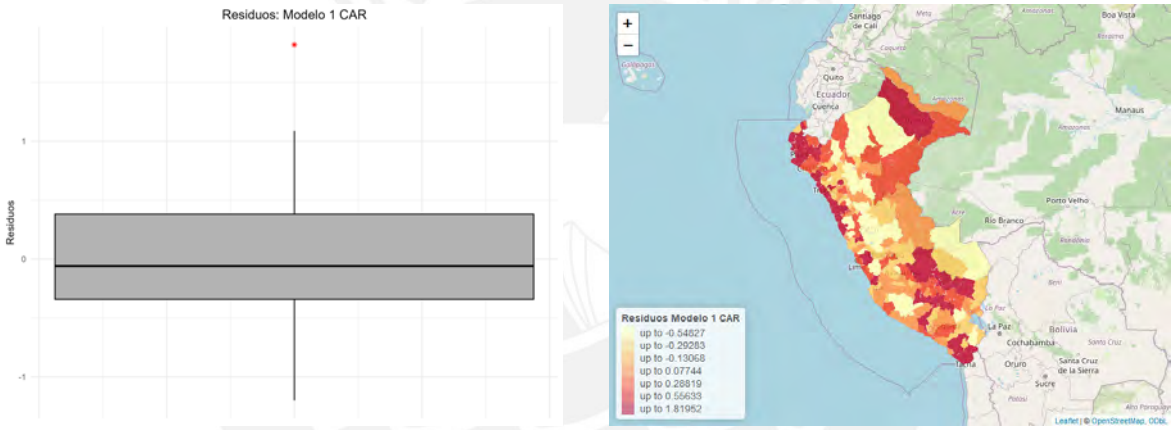
El mapa de estimación de $Y_i$ usando el modelo CAR-1 se muestra en la Figura 3. En el mapa del observamos una concentración más marcada de valores altos en regiones específicas, indicando una variabilidad más amplia en las estimaciones de $Y_i$.




(a) $Y_i$ Original (b) $Y_i$ estimado - Modelo CAR-1

Figura 3: Mapa de estimaciones de $Y_i$ con Modelo CAR-1.

El mapa de residuos para al modelo CAR-1 se muestra en la Figura 4.




(a) Diagrama de cajas (b) Residuos est. - Modelo CAR-1

Figura 4: Representación de los residuos del modelo CAR-1 mediante un diagrama de cajas (izquierda) y un mapa (derecha).

ET LUX IN TENEBRIS LUCET
MCMXVII

# Apéndice E

## Apéndice E: Número de Vecinos por Provincia

En la Figura 5, se presenta un histograma que muestra la distribución del número de vecinos por provincia en Perú cuando se asume que dos provincias son vecinas si comparten algún límite geográfico.

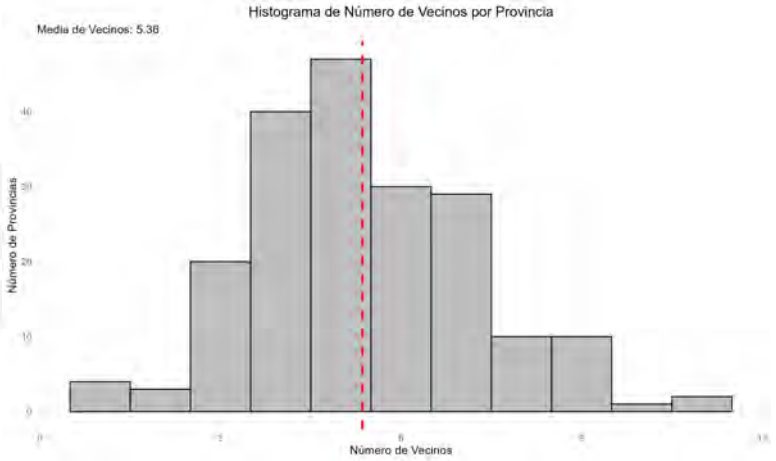



Figura 5: Perú: Histograma de Número de Vecinos por Provincia.

La Figura 6 presenta un grafo que muestra las relaciones de vecindad entre las provincias de Maynas, Requena, Putumayo, Loreto, y Mariscal Ramón Castilla. Cada nodo del grafo representa una provincia, mientras que las aristas indican la conexión de vecindad entre ellas.

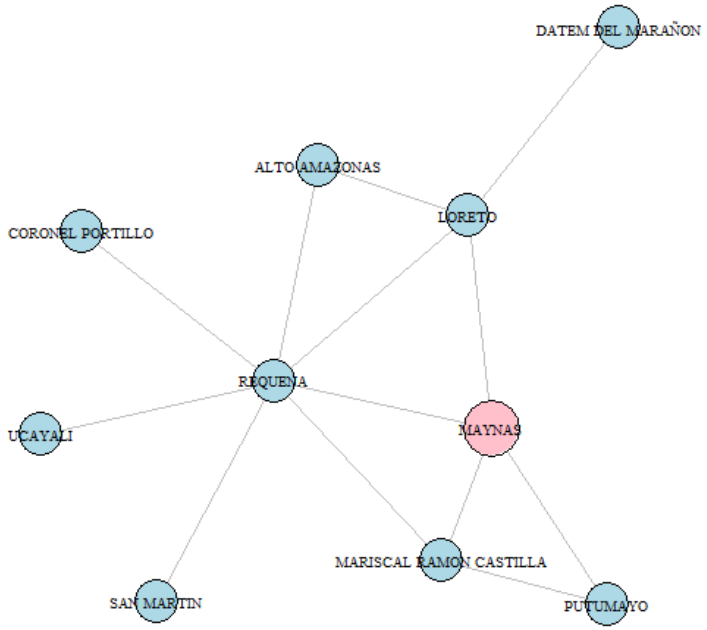



Figura 6: Grafo de las provincias de Maynas, Requena, Putumayo, Loreto, y Mariscal Ramon Castilla .

# Apéndice F

## Apéndice F: Comparación de Intervalos de Confianza

El gráfico presentado en la Figura 7 muestra una comparación de los intervalos de confianza al 95 % para los parámetros estimados en los modelos CAR-2 y SAR-2. Cada punto en el gráfico representa la estimación puntual del parámetro, mientras que las líneas que se extienden desde los puntos indican los intervalos de confianza, reflejando la precisión de estas estimaciones.

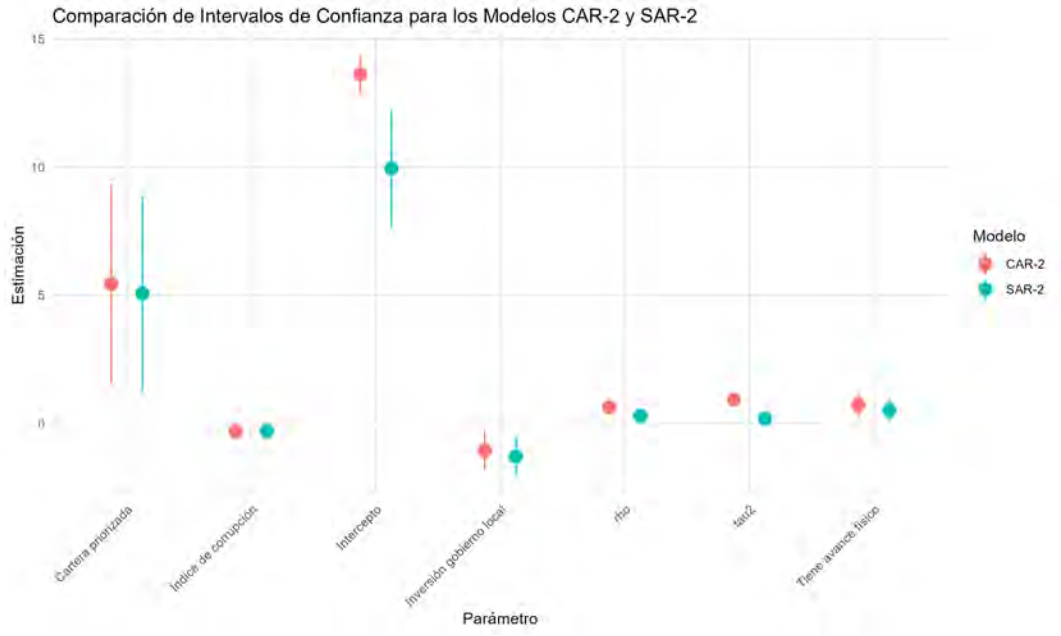



Figura 7: Comparación de Intervalos de Confianza de los parámetros estimados para el Modelo CAR-2 y SAR-2.

ET LUX IN TENEBRIS LUCET
MCMXVII

# Bibliografía


Arslan, O. y Akyürek, (2018). Spatial modelling of air pollution from $PM_{10}$ and $SO_2$ concentrations during winter season in marmara region (2013-2014), *International Journal of Environment and Geoinformatics* **5**(1): 1 – 16.

Assunção, R. y Krainski, E. (2009). Neighborhood dependence in Bayesian spatial models, *Biometrical journal* **51**(5): 851–69.

Banerjee, S., Carlin, B. P. y Gelfand, A. E. (2014). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman and Hall/CRC.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **36**: 192–236.

Besag, J. (1986). On the statistical analysis of dirty pictures, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **48**(3): 259–302.

Besag, J. y Higdon, D. M. (1999). Bayesian analysis of agricultural field experiments, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61**.

Costa-i Fon, J. y Rodriguez-Oreggia, E. (2005). Is the impact of public investment neutral across the regional income distribution? evidence from Mexico, *Economic Geography* **81**(3): 305–322.

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

Datta, A., Banerjee, S., Hodges, J. S. y 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*

Duncan, E. W., White, N. M. y Mengersen, K. (2017). Spatial smoothing in bayesian models: a comparison of weights matrix specifications and their impact on inference, *International*

*Journal of Health Geographics* **16**(47).
**URL:** *https://doi.org/10.1186/s12942-017-0120-x*

Earnest, A., Morgan, G., Mengersen, K., Ryan, L., Summerhayes, R. y Beard, J. (2007). Evaluating the effect of neighbourhood weight matrices on smoothing properties of conditional autoregressive (car) models, *International Journal of Health Geographics* **6**(54).
**URL:** *http://www.ij-healthgeographics.com/content/6/1/54*

Elliott, P. y Wartenberg, D. (2004). Spatial epidemiology: current approaches and future challenges, *Environ Health Perspect.* **112**(9): 998–1006.

Jimenez, A., Merino, C. y Sosa, J. (2020). Perú: factores determinantes de la inversión pública en los gobiernos locales, periodo 2008-201, *Economía* .

Lastra, J. (2017). Perú: factores determinantes de la inversión pública en los gobiernos locales, periodo 2008-2014. Tesis de maestría. Pontificia Universidad Católica del Perú.

LeSage, J. P. y Thomas-Agnan, C. (2015). Interpreting Spatial Econometric Origin-Destination Flow Models, *Journal of Regional Science* **55**(2): 188–208.
**URL:** *https://ideas.repec.org/a/bla/jregsc/v55y2015i2p188-208.html*

Pichihua Serna, Z. (2022). Boletín anual de ejecución de la inversión pública 2022, *Technical Report 1-42*, Ministerio de Economía y Finanzas.

Rodríguez-Pose, A., Psycharis, Y. y Tselios, V. (2012). Public investment and regional growth and convergence: Evidence from Greece., *Papers in Regional Science* **91**: 543–568.

Rue, H. y Held, L. (2005). *Gaussian Markov Random Fields*, Chapman and Hall/CRC.

Wall, M. M. (2004). A close look at the spatial structure implied by the CAR and SAR models, *Journal of Statistical Planning and Inference* **121**(2): 311–324.

Whittle, P. (1954). On stationary processes in the plane, *Biometrika* **41**(3/4): 434–449.
**URL:** *http://www.jstor.org/stable/2332724*