Universidad de la República

Facultad de Ciencias Económicas y de Administración

Modelos Logísticos aplicados a Tarifas de Saneamiento de Montevideo

Trabajo final de grado presentado como requisito para la obtención del título

Licenciado en Estadística

Estudiante:
Ignacio Campón Villamayor

Tutor:
Fernando Massa Mandagarán

Montevideo, Octubre 2024
Uruguay

Resumen

En una ciudad ideal, donde todos los ciudadanos tienen acceso a una red de saneamiento general, y donde los vertimientos de agua son gestionados de manera que no generan contaminación en el entorno, se esperaría una mejora significativa en la calidad de vida. Sin embargo, este escenario dista de la realidad, especialmente cuando consideramos la infraestructura y el mantenimiento necesarios para operar una red de saneamiento que atienda a una población de más de un millón de habitantes. La gestión de una infraestructura de esta magnitud implica costos financieros sustanciales. Por esta razón, es crucial estudiar los patrones de pago de los usuarios en áreas donde ya existe una red de saneamiento, con el fin de identificar estrategias que aseguren un mayor porcentaje de pago de estos servicios y asegurar la sostenibilidad financiera de los sistemas de saneamiento urbano.

Esta investigación estudia el comportamiento de pago de los usuarios existentes desde mayo 2001 hasta diciembre 2022 de **tarifas de saneamiento** en Montevideo, Uruguay. Se construye una secuencia de **modelos logísticos**, los cuáles tienen el objetivo de predecir la probabilidad de pago de las tarifas de saneamiento. Se demuestra la presencia de **autocorrelación espacial** con el índice de Moran presente en los datos, por lo que se acude a modelar la variable de respuesta, con modelos que capturan la variabilidad espacial. Los modelos **CAR** (condicionales autorregresivos) y los modelos **ICAR** (intrínsecos condicionales autorregresivos) resultan ser las mejores opciones para capturar la variabilidad espacial presente y realizar predicciones sobre nuevas observaciones.

Palabras clave: Modelos Logísticos, Modelos Mixtos, Estadística Espacial, Datos de área, Autocorrelación espacial, Modelos Autorregresivos Condicionales e Intrínsecos Autorregresivos Condicionales.

Índice

Resumen I

Índice de figuras V

1. Introducción **1**

1.1. Saneamiento 1

1.2. Tarifa de Saneamiento 3

1.3. Problema en General 4

2. Marco Teórico **5**

2.1. Modelos Lineales Generalizados 5

2.1.1. Regresión Logística 7

2.1.2. Devianza 8

2.1.3. Inferencia 9

2.2. Inferencia Bayesiana 10

2.2.1. Aproximación Monte Carlo 12

2.3. Indicadores de Desempeño 14

2.3.1. Matriz de Confusión 14

2.3.2. Widely Applicable Information Criterion 15

2.4. Estadística Espacial: Datos de Área 18

2.4.1. Polígonos 19

2.4.2. Matriz de Vecinos y Pesos Espaciales 19

2.4.3. Autocorrelación Espacial 20

2.4.4. Índice de Moran 21

2.5. Modelos Mixtos 22

2.5.1. Modelo Autorregresivo Condicional (CAR) 23

2.5.2. Modelo Intrínseco Autorregresivo Condicional (ICAR) 25

2.6. Metodología 26

3. Datos **28**

3.1. Base de datos 28

3.2. Procesamiento de datos 29

3.2.1. Filtrado de observaciones 30
3.2.2. Transformación de variables 32
3.2.3. Referenciación geográfica 32
3.3. Análisis descriptivo 34

4. Resultados **43**
4.1. Modelo Logístico I 44
4.2. Modelo Logístico II 50
4.3. Modelo Logístico III 53
4.4. Modelo Logístico IV 54
4.5. Estructura espacial 55
4.6. Modelo Logístico - CAR 61
4.7. Modelo Logístico - ICAR 64

5. Conclusiones **69**

A. Apéndices **71**
A. Resultados Complementarios 71
A.1. Visualizaciones y diagnóstico 71

Referencias **82**

Índice de figuras

3.1. Gráfico de barras de la variable CATEGORIA. 31
3.2. Padrón 42563 `multipolygon`, imagen extraída del SIG. 34
3.3. Gráfico de barras de las TDS por CCZ y CATEGORÍA. 35
3.4. Mapa de Montevideo, con cuentas totales por CCZ. 36
3.5. Mapa de Montevideo, con porcentaje de CORTADAS por CCZ. 37
3.6. Mapa de Montevideo, con porcentaje de cuentas con deuda ≥ 1 año. . . . 38
3.7. Cuentas por padrón del CCZ 1. 38
3.8. Promedio de bimestres impagos del CCZ 1 por padrón. 39
3.9. Porcentaje de bimestres impagos por CCZ. 40
3.10. Promedio de pagos del último bimestre por CCZ. 41
4.1. Coeficientes estimados asociados a la variable *BIM5.1* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha). 47
4.2. Coeficientes estimados asociados a la variable *BIM1.1* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha). 47
4.3. Mapa de Montevideo de los CCZ y su estructura de vecindad según el criterio de contigüidad. 56
4.4. Mapa de Montevideo según la *media* de los residuos de devianza del modelo 4.4 por CCZ. 57
4.5. Mapa de Montevideo según la *mediana* de los residuos de devianza del modelo 4.4 por CCZ 58
4.6. Mapa de Montevideo de los CCZ según la media de la variable dependiente *BIM6*. 60
4.7. Posterior plot del coeficiente α del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad. 62
4.8. $e^{\phi} - 1$ de los efectos aleatorios, del modelo 4.6. 63
4.9. Posterior plot de la estimación del desvío σ_{CAR} del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad. 66
4.10. $e^{\phi} - 1$ de los efectos aleatorios, del modelo 4.8. 68
A.1. Coeficientes estimados asociados a la variable *CATEG.DOM* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha). 71
A.2. Coeficientes estimados asociados a la variable *CANT_UN* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha). 72

A.3. Coeficientes estimados asociados a la variable *PROP_BIM_IMP* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha). 72
A.4. Trace plots del modelo 4.4. 73
A.5. Posterior plot del *Intercept* del modelo 4.4, el área sombreada es el intervalo al 95 % de probabilidad. 73
A.6. Posterior plot del modelo 4.4, el área sombreada es el intervalo al 95 % de probabilidad. 74
A.7. Trace plots del modelo 4.6. 74
A.8. Trace plots del modelo 4.6. 75
A.9. Trace plots del modelo 4.6. 75
A.10. Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad. 76
A.11. Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad. 76
A.12. Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad. 77
A.13. Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad. 77
A.14. Trace plots del modelo 4.8. 78
A.15. Trace plots del modelo 4.8. 78
A.16. Trace plots del modelo 4.8. 79
A.17. Posterior plot del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad. 79
A.18. Posterior plot del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad. 80
A.19. Posterior plot del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad. 80

1. Introducción

1.1. Saneamiento

El *saneamiento*, según la Real Academia Española, se define como el conjunto de técnicas y sistemas destinados a mejorar las condiciones higiénicas de una comunidad o una ciudad. Es un factor crucial para mejorar la calidad de vida de las personas en todo el mundo. Según datos de la Organización Mundial de la Salud (OMS), en 2022, el 57 % de la población mundial utilizaba un servicio de saneamiento gestionado de forma segura. Los servicios de saneamiento deficientes no solo reducen el bienestar humano y el desarrollo social y económico, sino que también están asociados con la transmisión de enfermedades como el cólera, la disentería, la fiebre tifoidea, y contribuyen a la malnutrición y el retraso del crecimiento (Organización Mundial de la Salud, 2023).

Es fundamental distinguir entre el saneamiento de red general y otros tipos de saneamiento. El saneamiento de red general se refiere a un sistema integrado de tuberías y plantas de tratamiento que gestionan de manera centralizada las aguas residuales de una comunidad o ciudad. Este tipo de saneamiento es más eficiente y sostenible a largo plazo, en comparación con sistemas individuales como pozos sépticos o letrinas, que, aunque útiles en áreas rurales o con baja densidad de población, no ofrecen los mismos niveles de protección sanitaria y ambiental. El saneamiento de red general permite un tratamiento y disposición más controlados y efectivos de las aguas residuales, reduciendo significativamente el riesgo de enfermedades y mejorando la calidad del agua.

Muchas investigaciones tal y como Regueira (2015), Prindle y Salas (1967), Peña Barreto (2016)y más, han aportado conocimiento sobre la importancia del saneamiento en la Salud de la población, y entre ellas promovidas por la OMS. El acceso a un saneamiento seguro y eficiente tiene múltiples beneficios, más allá de la reducción del riesgo de diarrea. Entre estos beneficios se incluyen la promoción de la dignidad y el aumento de la seguridad, especialmente para mujeres y niñas, la mejora en la asistencia escolar, la reducción de la propagación de la resistencia a los antimicrobianos, y el potencial para aumentar la resiliencia general de la comunidad a las perturbaciones climáticas. Además, se estima que cada dólar invertido en saneamiento produce un retorno de $5.50 en menores costos de atención de salud, mayor productividad y menos muertes prematuras (Organización Mundial de la Salud, 2023).

Según la Organización Panamericana de la Salud (OPS), alrededor del 50,8 % de la

población en América Latina y el Caribe no tiene acceso a un saneamiento gestionado de manera segura. Esta falta de saneamiento adecuado está directamente relacionada con varios problemas de salud pública (Organización Panamericana de la Salud, 2023).

Además, la Comisión Económica para América Latina y el Caribe (CEPAL) destaca en su publicación *Estudio Económico de América Latina y el Caribe* (Comisión Económica para América Latina y el Caribe, 2024) la necesidad de una inversión significativa en infraestructuras hidráulicas y saneamiento para impulsar la economía y crear empleo en la región. Según estimaciones de la CEPAL, una inversión equivalente al 1,3 % del Producto Interior Bruto (PIB) regional durante diez años no solo mejoraría el bienestar y la calidad de vida de la población, sino que también generaría alrededor de 3,6 millones de empleos que contribuyen a preservar y restaurar el medio ambiente (empleos verdes) anualmente. Estos esfuerzos son esenciales para lograr una transición hídrica justa y sostenible (Organización Naciones Unidas, 2023).

En conclusión, el manejo adecuado de una red de saneamiento y el tratamiento y procesamiento adecuados de las aguas residuales garantizarán una calidad adecuada del agua, lo que previene la propagación de enfermedades.

Este trabajo busca promover la mejora de la infraestructura de saneamiento de red general en Montevideo, y ser tomada como ejemplo para el resto de departamentos en el país, asi como destacar la importancia de su uso.

Actualmente, la Intendencia de Montevideo es la encargada de gestionar al saneamiento en dicho departamento. El pago bimestral de los usuarios, mediante el impuesto al saneamiento, también conocido como Tarifa de Saneamiento (TDS), permite el mantenimiento y la realización de nuevas obras de infraestructura en la red general.

La morosidad en el pago de la TDS, representa un problema, impidiendo el mantenimiento y expansión adecuada de la infraestructura de saneamiento en Montevideo. Este trabajo tiene como objetivo principal abordar esta problemática mediante la predicción de la morosidad, permitiendo así, identificar usuarios y variables relevantes para generar una mejora en la recaudación. De esta manera, se busca asegurar que el sistema de saneamiento abarque eficientemente todo el departamento, mejorando la calidad de vida y salud pública de sus habitantes.

1.2. Tarifa de Saneamiento

La Tarifa de Saneamiento fue creada por los Artículos 89 al 95 del Decreto Nº 29434 de la Junta Departamental de Montevideo (2001), promulgado por la Resolución Nº 1594/01 del Intendente Municipal de Montevideo el 10 de mayo de 2001 y reglamentado por la Resolución Nº 2377/01 del 2 de julio de 2001.

Este impuesto, es el precio que debe abonarse por hacer uso de la red de saneamiento en Montevideo. Se calcula respecto a los vertimientos de agua realizados (cargo variable) y la cantidad de unidades ocupacionales que la conforman (cargo fijo). Más adelante se profundizan los conceptos de cargo fijo y cargo variable.

La TDS está conformada por al menos una unidad ocupacional y al menos un medidor de agua. A partir de ello es que se asignan TDS para cada unidad ocupacional.

Una unidad ocupacional puede ser una casa, un comercio, una escuela, un edificio, una fábrica, etcétera. Por ejemplo, un edificio puede tener tarifas de saneamiento individuales para cada propietario o inquilino, es decir una para cada apartamento, en donde cada uno debe hacerse cargo de abonarla; o por otro lado, puede tener una TDS única para todo el edificio la cual suele ser gestionada por la administración del edificio, encargada de dividir dicho monto en partes iguales respecto a la cantidad de unidades (apartamentos) y siendo incluida en los gastos comunes.

Los medidores de agua son gestionados por las Obras Sanitarias del Estado (OSE), su función es medir el volumen de consumo de agua en m^3; cada unidad ocupacional cuenta con un medidor, dichos medidores pueden abastecer a una unidad o más de una. Un edificio suele contar con un único medidor que abastece a todos los apartamentos del edifico, en cambio un conjunto de casas suelen tener un medidor de agua individual para cada vivienda.

La TDS se abona bimestralmente y su monto se calcula en base a los cargos fijos y cargos variables que varían de acuerdo a como esté compuesta cada tarifa. El cargo fijo es un monto fijo por cada unidad ocupacional que componga la tarifa, el cual no varía a lo largo de los períodos siempre y cuando la composición de la tarifa no cambie. Por otra parte, el cargo variable se calcula en base al total de volumen de agua consumido por las unidades en la tarifa. El promedio de consumo por unidad ocupacional del tipo casa familiar es de aproximadamente 5 m^3. En base al decreto Nº 29434, la Intendencia de Montevideo calcula que el 80 % del agua consumida se vierte a los colectores de saneamiento de red

general, en base a ello se fija un monto fijo por unidad de m^3 consumido, y de acuerdo al volumen total, que determinan los medidores de agua, se calcula el cargo variable.

Estos dos cargos establecen el monto a abonarse en el bimestre. Cuando los pagos no son realizados en tiempo y forma, a dichos montos se les suma las multas y recargos que deban realizarse, los cuales han de irse incrementando a medida que pasa el tiempo.

1.3. Problema en General

La Tarifa de Saneamiento, es el tercer impuesto de mayor recaudación detrás de la Patente de Rodados y la Contribución Inmobiliaria que regula la Intendencia de Montevideo. A fines de 2022, la morosidad en términos relativos ha alcanzado el 7,98 %, es decir del total de bimestres generados para todas las observaciones desde su la fecha de creación, el 7,98 % de dichos bimestres no son pagados. Esta morosidad representa un problema a la hora de gestionar de manera eficiente y equitativa las redes de saneamiento en el departamento.

Este trabajo busca indagar una forma de contrarrestar esa morosidad creando un modelo estadístico que logre predecir la probabilidad de pago del bimestre próximo para cada usuario. De esta forma se busca identificar aquellas observaciones con menor probabilidad y encontrar una solución beneficiosa para ambas partes.

El documento se estructura en cuatro grandes secciones: Datos, Marco Teórico, Resultados y Conclusiones. La sección Marco Teórico, describe los conocimientos y métodos estadísticos aplicados a lo largo del trabajo. La sección Datos, contiene una descripción de la base de datos sin procesar, significado de las variables, procesamiento de datos realizado y una descripción estadística de las variables relevantes. La sección Resultados explicita los pasos realizados, los razonamientos implícitos y los resultados obtenidos para la obtención del modelo final. Por último, la sección Conclusiones, describe en resumen los resultados obtenidos, así como sus implicaciones en términos del problema y aspectos futuros a resolver.

2. Marco Teórico

El Marco Teórico abarca un conjunto de técnicas y enfoques analíticos esenciales para el desarrollo del estudio. Se comienza describiendo los *Modelos Lineales Generalizados*, ya que conforman el punto de partida en la creación de regresiones logísticas para la variable dependendiente de interés (pago de la TDS en el próximo semestre). Luego, se introducen conceptos propios de la *Inferencia Bayesiana* ya que el modelo final fue ajustado bajo este enfoque . Por otro lado, se describen los *Indicadores de Desempeño* utilizados en la comparación de modelos. Herramientas como la matriz de confusión y el widely applicable information criterion (WAIC), son fundamentales para evaluar la precisión, adecuación y comparación de los modelos. Otro punto a destacar es la *Estadística Espacial*, centrada en el análisis de Datos de Área, matrices de vecindad, y el estudio de autocorrelación espacial, proporcionando una visión integral de los patrones espaciales. Se incluyen conceptos propios de los *Modelos Mixtos*, más concretamente aquellos que incluyen efectos aleatorios con estructura espacial; los modelos Conditional Auto-Regressive (CAR) e Intrinsic Conditional Auto-Regressive (ICAR). Este apartado culmina con la sección *Metodología* la cual sintetiza los métodos empleados y su aplicación en el contexto del estudio.

2.1. Modelos Lineales Generalizados

Los *Modelos Lineales Generalizados* (GLM) son una familia de modelos estadísticos utilizados para describir la relación entre una variable de respuesta y una o más variables explicativas. A diferencia de los modelos lineales múltiples, los GLM permiten modelar una variable de respuesta que no necesariamente sigue una distribución gaussiana, sino que puede tener cualquier distribución perteneciente a la familia exponencial (McCullagh y Nelder, 1989).

Estos modelos están formados por 3 componentes principales:

- **Componente Aleatorio:** representa la variabilidad no explicada o aleatoria en los datos. Corresponde a aquellos factores que no están incluidos en el modelo y que generan fluctuaciones o variaciones en los datos observados. Se asume que esta variabilidad aleatoria puede ser modelada por una distribución probabilística de la familia exponencial que describe la variabilidad inherente de la variable respuesta.
- **Componente Sistemático:** representa la relación sistemática o determinística entre

las variables predictoras y la variable de respuesta. Típicamente se asume una relación lineal, en la que se especifican las variables predictoras y sus coeficientes.

- **Función de Enlace:** establece la relación entre la media de la variable de respuesta y el componente sistemático del modelo. Su función es transformar la escala lineal del componente sistemático en una escala adecuada para la media de la variable de respuesta.

El Modelo de Regresión Lineal Múltiple es un caso particular de los Modelos Lineales Generalizados. Consideremos en lugar de la variable de respuesta y_i, ahora la respuesta media μ_i, que para un valor dado de un predictor x_i es la media o valor esperado de y_i, entonces:

$$E(y_i) = \mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}, \qquad i = 1, \ldots, n. \tag{2.1}$$

$$y_i \overset{iid}{\sim} \mathcal{N}(\mu_i, \sigma^2) \tag{2.2}$$

La ecuación 2.1, refiere al Componente Sistemático mientras que la ecuación 2.2, al Componente Aleatorio. En este marco, el componente sistemático describe cómo las variables explicativas afectan la media. Una vez que se modela cómo cambia la media (μ_i) en función de las variables explicativas (x_i), el componente aleatorio describe la variabilidad que observamos en la respuesta dado el valor del parámetro.

De manera más amplia, un GLM tiene un Componente Sistemático que incluye p variables explicativas:

$$\begin{gathered} \eta_i = \beta_0 + \sum_{j=1}^{p} x_j \beta_j \\ g(\mu_i) = \eta_i \\ i = 1, \ldots, n. \quad j = 1, \ldots, p. \end{gathered} \tag{2.3}$$

donde, para generalizar el componente sistemático, se utiliza una función de enlace $g(\mu)$ que requiere que alguna función de la respuesta esté relacionada de manera lineal con las variables explicativas (Grossman, Marcus, Palmer, Pulham, y Rumments, 2021).

El Componente Aleatorio se puede escribir de manera general como:

$$y_i \overset{iid}{\sim} EDM(g(\mu_i)) \tag{2.4}$$

Donde EDM por sus siglas en inglés (Exponential Distribution Model) representa una distribución de probabilidad de la Familia Exponencial.

En los modelos lineales generalizados, se estiman los parámetros utilizando la estimación de Máxima Verosimilitud. En el caso del modelo de regresión lineal, esto es equivalente a minimizar la suma de los cuadrados. Sin embargo, para otros GLM's, no existe una solución de forma cerrada, lo que obliga a realizar un algoritmo iterativo para obtener las estimaciones de los parámetros (Grossman y cols., 2021).

2.1.1. Regresión Logística

La *Regresión Logística* es un caso particular de los GLM, donde el Componente Aleatorio es modelado mediante una distribución Bernoulli. En un modelo logístico, la variable dependiente es una variable binaria (cuyas categorías indican la ausencia o presencia de cierta característica), mientras que las variables independientes pueden ser de naturaleza cuantitativa o categórica. En estos casos, la variable de respuesta Y suele ser codificada mediante los valores 0, 1 donde es de interés predecir el resultado (o la probabilidad del resultado) en base a ciertos predictores.

$$\begin{aligned} \log\left(\frac{p_i}{1+p_i}\right) =& \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \\ & i = 1, \ldots, n. \end{aligned} \tag{2.5}$$

donde $p_i = E(y_i)$, limitado por $0 \leqslant p_i \leqslant 1$, representa la probabilidad de que y_i tome el valor 1 dado un conjunto de características x_i:

$$p_i = \frac{1}{1 + \exp\left\{-(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi})\right\}}$$

De esta forma, el modelo logístico utiliza la función sigmoide, *logit* para modelar la relación (función de enlace) entre un conjunto de variables explicativas y la probabilidad de que la variable dependiente adopte el valor 1. La función sigmoide transforma los valores continuos del predictor lineal $\eta_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_j x_{ji}$ (componente sistemático) en una escala de 0 a 1, lo que representa la probabilidad de que el evento binario ocurra (Rencher y Schaalje, 2008).

Los coeficientes que son estimados en 2.5, pertenecen a la escala *log-odds*. Para interpretar dicho valor, se realiza parte de la transformación inversa, esto permite obtener el resultado en una escala interpretable y conocida como *odds-ratio*, donde simplemente se exponencia el valor del coeficiente, es decir: e^{β}.

2.1.2. Devianza

La *devianza* es el estadístico de cociente de verosimilitudes utilizado en GLM binomiales, en donde se compara el modelo propuesto con el modelo ideal sin restricciones (modelo saturado). Intuitivamente decimos que la devianza mide que tan bien se ajusta un cierto GLM a los datos, con respecto a un modelo ideal que tiene un ajuste perfecto, es decir que tiene tantos parámetros como observaciones y al cual se lo conoce como modelo saturado. Este estadístico, es una medida de la bondad de ajuste de un modelo, utilizada en el análisis de GLM y la cual surge como alternativa al R^2 (Agresti, 2015).

Formalmente, la devianza se plantea como:

$$D = -2 \cdot \log \left[\frac{\mathcal{L}(Y|\mu, \phi)}{\mathcal{L}_S(Y|Y, \phi)} \right] \tag{2.6}$$

donde $\mathcal{L}(Y|\mu, \phi)$ es la verosimilitud del modelo propuesto , y $\mathcal{L}_S(Y|Y, \phi)$ es la verosimilitud del modelo saturado, la máxima verosimilitud posible. Este cociente puede interpretarse como la proporción de verosimilitud explicada por el modelo propuesto en comparación con la máxima verosimilitud alcanzable. De forma análoga puede plantearse para cada observación como:

$$d_i = -2 \cdot \log \left[\frac{\mathcal{L}(y_i|\mu, \phi)}{\mathcal{L}_S(y_i|y_i, \phi)} \right]$$

donde d_i es el valor de la devianza correspondiente a la i-ésima observación. La suma de los residuos de devianza para cada observación resulta en un estadístico de bondad de ajuste, que resulta ser la propia devianza:

$$D = \sum_{i=1}^{n} d_i^2$$

De esta manera, la devianza es una medida de error donde valores bajos corresponden a situaciones donde el ajuste del modelo está cerca del modelo ideal (saturado), mientras que, valores altos indicarán un peor ajuste.

La devianza residual (residual deviance) es la devianza del modelo actual, mientras que la devianza nula (null deviance) es la devianza de un modelo sin predictores y solo una constante.

2.1.3. Inferencia

El estadístico de Wald permite evaluar pruebas de hipótesis de la forma:

$$\mathbf{H_0} : \beta_j = 0$$
$$\mathbf{H_1} : \beta_j \neq 0$$

Pudiendo ser β_j un coeficiente del predictor lineal de un GLM como se describió en la sección 2.1.1. El estadístico de prueba empleado en estos cuantifica la diferencia entre el valor estimado del coeficiente con el valor especificado bajo H_0 (estandarizado por su desviación estándar), es decir:

$$t = \frac{\hat{\beta}_j - 0}{s_{\hat{\beta}_j}}$$

En un contexto de GLM, la estimación del desvío estándar de $\hat{\beta}_j$ suele obtenerse empleando los elementos de la diagonal del inverso de la matriz de información de Fisher. Si bien este estadístco cuenta con una distribución exacta en el caso del modelo lineal general, cuando el componente aleatorio del GLM es diferente del gaussiano (y los parámetros se estiman por máxima verosimilitud), el estadístico se distribuye asintóticamente normal con media 0 y varianza 1. La utilidad estadística de esta prueba indica si la variable asociada a β_j tiene una contribución significativa al ajuste del modelo.

El estadístico de razón de verosimilitud (*Likelihood Ratio Test, LRT*) es una herramienta utilizada en la comparación de modelos anidados. Permite evaluar si un modelo más complejo ($\mathbf{H_1}$) se ajusta a los datos de manera significativamente mejor que un modelo mas simple que incluye menos variables explicativas ($\mathbf{H_0}$). Se calcula mediante el doble de la diferencia en el logaritmo de la verosimilitud entre los dos modelos considerados:

$$\begin{aligned}\Lambda &= -2\log\left(\frac{\mathcal{L}_{completo}}{\mathcal{L}_{reducido}}\right) = -2\log\left(\frac{\mathcal{L}(Y|\hat{\beta}^{(0)})}{\mathcal{L}(Y|\hat{\beta}^{(1)})}\right) \\ &= -2\left[\log(\mathcal{L}(Y|\hat{\beta}^{(0)})) - \log(\mathcal{L}(Y|\hat{\beta}^{(1)}))\right]\end{aligned}$$

donde $\mathcal{L}$ hace referencia a la función de verosímilitud del modelo, y $\hat{\beta}^{(0)}$, $\hat{\beta}^{(1)}$ son los estimadores bajo el cumplimiento de las hipótesis $\mathbf{H_0}$) y $\mathbf{H_1}$) respectivamente.

Siendo Y una variable binaria y $p_i = E(y_i)$, es posible considerar el modelo completo de la siguiente manera:

$$\log\left(\frac{p_i}{1+p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_t x_t + \cdots + \beta_p x_p$$

Por otro lado, siendo $t < p$ (con p igual a la cantidad de predictores) el modelo reducido

es:

$$\log\left(\frac{p_i}{1+p_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_t x_t$$

De esta forma, el estadístico *LRT* permite contrastar la significancia del modelo completo, contra un modelo mas simple.

El contraste de hipótesis se plantea de la siguiente manera:

$$\mathbf{H_0} : \beta_j = 0, \text{ para } t < j \leq p$$
$$\mathbf{H_1} : \text{ algún } \beta_j \neq 0, \text{ para } t < j \leq p$$

Este estadístico, bajo el cumplimiento de $\mathbf{H_0}$), se distribuye asíntoticamente χ^2 con ν grados de libertad, los cuales corresponden a la diferencia entre las dimensiones de los espacios paramétricos de las hipótesis nula y alternativa (McCullagh y Nelder, 1989).

$$\Lambda \overset{a}{\sim} \chi^2_{p-t}$$

La significancia estadística obtenida de esta prueba indica si las variables adicionales consideradas bajo $\mathbf{H_1}$ tienen una contribución significativa al ajuste del modelo. Utilizar el LRT proporciona un marco riguroso para tomar decisiones informadas sobre la inclusión de variables en modelos estadísticos, asegurando que cada variable aporta información valiosa y mejora la capacidad predictiva del modelo.

Un caso particular de este estadístico surge al considerar la comparación de un modelo donde se emplean k variables explicativas y otro donde el ajuste es perfecto (el modelo se satura intencionalmente).

$$\Lambda = -2\log\Big[\mathcal{L}(Y|\hat{\beta}) - \mathcal{L}(Y|Y)\Big]$$

Así, el estadístico es igual a la devianza del modelo. En este caso los grados de libertad son $\nu = n-k$. De esta forma, se determina si el ajuste del modelo propuesto es razonable frente al modelo saturado. Valores altos del estadístico reflejan un ajuste insuficiente.

2.2. Inferencia Bayesiana

La *Inferencia Bayesiana*, es un enfoque estadístico que permite realizar inferencia sobre parámetros desconocidos o predicciones futuras basadas en la probabilidad y el teorema de Bayes. Este enfoque se basa en la idea de que es posible utilizar información previa o conocimientos a priori sobre los parámetros y actualizarlos a medida que se recopila nueva información. La incorporación de información previa puede ser especialmente útil

cuando los datos son limitados o cuando se cuenta con información relevante de estudios previos (Hoff, 2009).

A diferencia del enfoque frecuentista tradicional, que se centra en estimar parámetros como valores fijos, la inferencia bayesiana considera los parámetros como variables aleatorias con distribuciones de probabilidad. Esto proporciona una perspectiva más completa y flexible para el análisis de datos.

La inferencia estadística se ocupa de elaborar métodos para estimar características generales de un fenómeno de interés a partir de una cantidad finita de datos observados. Sea un conjunto de datos $y = (y_1, y_2, \ldots, y_n)$ los cuáles son observaciones de cierto fenómeno bajo estudio y por otro lado, el o los parámetros $(\theta_1, \theta_2, \ldots)$ que hacen referencia a las características relevantes de los datos. Los objetivos inferenciales son *explicar* características relevantes de y (estimando θ), *predecir* el valor de observaciones futuras y *comparar* modelizaciones alternativas.

Previo a observar los datos, existe incertidumbre sobre los mismos y también sobre los parámetros del fenómeno de estudio. La manera en la que se plantea la inferencia en el contexto bayesiano consiste en asignar una distribución a los datos, siendo esta la función de verosimilitud $p(y|\theta)$ y, por otro lado, asignar una distribución a los parámetros desconocidos $p(\theta)$ denominada distribución a *priori* (previa).

Tras la recopilación de datos relevantes para el problema en cuestión, ya no existe incertidumbre sobre los mismos. En este momento, se actualiza la distribución de los parámetros. Se actualiza la incertidumbre sobre θ mediante la regla de Bayes obteniéndose una nueva distribución, denominada distribución a *posteriori* (posterior) $p(\theta|y)$. Esta representa el conocimiento actualizada de los parámetros dado los datos observados. Como se mencionó anteriormente, la misma se obtiene aplicando el teorema de Bayes para densidades:

$$p(\theta|y) = \frac{p(y|\theta) \cdot p(\theta)}{p(y)} = \frac{p(y|\theta) \cdot p(\theta)}{\int p(y|\theta) \cdot p(\theta)\, \mathrm{d}\theta} \tag{2.7}$$

A partir de la distribución a *posteriori* es posible interpretar los resultados a través de alguna medida de resumen de esta distribución, típicamente se obtienen estimaciones puntuales y/o intervalos de probabilidad para los parámetros desconocidos.

La librería implementada para aplicar el enfoque bayesiano en este trabajo es la librería `brms` (Bürkner, 2017). Dicha librería esta basada en la librería `rstan` (Guo y cols., 2023), la cual crea una interfaz con *Stan* (Stan Development Team, 2024), el cual es

un lenguaje de programación y una plataforma para la estadística bayesiana que facilita la implementación de métodos de Monte Carlo Hamiltoniano basados en cadenas de Markov (MCMC). Esta plataforma permite especificar, compilar y realizar inferencia sobre modelos estadísticos complejos a partir de muestras generadas de una distribución posterior de los parámetros de un modelo.

2.2.1. Aproximación Monte Carlo

Se define una distribución *previa conjugada* como una distribución de probabilidad que al ser combinada con datos observados, produce una distribución posterior que tiene la misma forma funcional que la distribución previa. En otras palabras, la distribución previa y la posterior pertenecen a la misma familia paramétrica.

La ventaja de utilizar una distribución previa conjugada para un parámetro desconocido θ, es que el cálculo de la distribución posterior se vuelve mucho más simple y más manejable, ya que la forma funcional se mantiene constante. Sin embargo, a menudo se desea resumir otros aspectos de una distribución posterior. Por ejemplo, se puede estar interesado en las medias y desviaciones estándar de alguna función de θ, o en la distribución predictiva de datos faltantes o no observados. Al comparar dos o más poblaciones, se puede estar interesado en la distribución posterior de $|\theta_1 - \theta_2|$ o $\frac{\theta_1}{\theta_2}$, todas las cuales son funciones de más de un parámetro. Obtener propiedades exactas para estas cantidades posteriores puede ser difícil o imposible. No obstante, sí es posible (y generalmente más sencillo) generar una muestra de valores aleatorios de los parámetros a partir de sus distribuciones posteriores. De esta manera, la distribución de estas cantidades de interés pueden ser aproximadas con un grado de precisión arbitrario utilizando el método de Monte Carlo (Hoff, 2009).

El método, conocido como *aproximación de Monte Carlo*, se basa en el muestreo aleatorio y su implementación no requiere un profundo conocimiento de cálculo o análisis numérico.

En el contexto de la inferencia bayesiana, sea θ un parámetro de interés con distribución $p(\theta)$, sean $y_1, y_2, \ldots, y_n$ los valores de una muestra con distribución $p(y_1, \ldots, y_n|\theta)$ y $p(\theta|y_1, y_2, \ldots, y_n)$ la correspondiente distribución posterior, es posible muestrear una cierta cantidad S de valores de θ a partir de esta última. La distribución empírica de $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ se conoce como una aproximación de Monte Carlo de $p(\theta|y_1, \ldots, y_n)$. Muchos lenguajes de programación y entornos informáticos cuentan con procedimientos para simular este proceso de muestreo. Por ejemplo, `R` (R Core Team, 2023) cuenta con

funciones incorporadas para simular muestras *i.i.d.* de la mayoría de las distribuciones. Este trabajo utilizó funciones implementadas en la librería `brms`, basadas en el algorítmo No-U-Turn (NUTS), una variante del Monte Carlo Hamiltoniano (Guo y cols., 2023).

La distribución empírica de estos valores $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ aproxima a $p(\theta|y_1, \ldots, y_n)$, mejorando la aproximación a medida que S aumenta. La ley de los grandes números establece que si $\theta^{(1)}, \ldots, \theta^{(S)}$ conforman una muestra *i.i.d.* de $p(\theta|y_1, \ldots, y_n)$, entonces:

$$\frac{1}{S}\sum_{s=1}^{S}\theta^{(s)} \to E[\theta|y_1, \ldots, y_n], \quad S \to \infty$$

A partir de resultados similares es posible obtener que:

- $\sum_{s=1}^{S} \frac{(\theta^{(s)}-\bar{\theta})^2}{(S-1)} \to Var[\theta|y_1, \ldots, y_n]$.
- la mediana de $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{1/2}$.
- el percentil α de $\{\theta^{(1)}, \ldots, \theta^{(S)}\} \to \theta_{\alpha}$.

Casi cualquier aspecto de una distribución posterior en la que se esté interesado puede aproximarse de manera arbitrariamente precisa con una muestra de Monte Carlo lo suficientemente grande (Hoff, 2009).

Para lograr una aproximación precisa de cualquier aspecto de la distribución posterior, es fundamental contar con una muestra de Monte Carlo lo suficientemente grande. Sin embargo, para garantizar que estas muestras sean representativas de la distribución de equilibrio, es crucial monitorear la convergencia de las cadenas.

Una forma de monitorear si una cadena ha convergido a la distribución de equilibrio es comparar su comportamiento con respecto a otras cadenas inicializadas aleatoriamente. El $\hat{R}$, es un indicador que mide la relación entre la varianza promedio de las muestras dentro de cada cadena y la varianza de las iteraciones entre cadenas. Si todas las cadenas están en equilibrio, el valor de $\hat{R}$ será próximo a uno. En caso de que las cadenas no hayan convergido a una distribución común, el indicador será mayor que uno (Stan Development Team, 2024).

El *effective sample size*, *ESS*, es el tamaño de muestra efectivo de de una muestra. Este indicador indica cuántas realizaciones independientes contienen la misma cantidad de información que la muestra dependiente obtenida por el algoritmo MCMC. Cuanto más próximo sea el ESS al número de iteraciones, mejor. Para asegurar estimaciones confiables de las varianzas y autocorrelaciones necesarias para ESS, se recomienda que el

ESS normalizado por rango sea mayor a 400, un número basado en la experiencia práctica y simulaciones (Vehtari, Gelman, Simpson, Carpenter, y Bürkner, 2021), generalmente suficiente para obtener una estimación estable del error estándar de Monte Carlo. Similar al indicador anterior, el *Bulk ESS* es una métrica que evalúa la calidad de las muestras en la parte central de la distribución posterior, estimando cuántas muestras independientes equivalen a las muestras correlacionadas producidas por MCMC. Un Bulk ESS más alto indica que la cadena de MCMC está bien mezclada en el centro de la distribución, lo que significa que estimaciones como la media o la mediana son confiables.

2.3. Indicadores de Desempeño

A la hora de encontrar el modelo que mejor se adapte a los datos, es necesario acudir a indicadores, medidas cuantitativas que permitan evaluar el desempeño y comparar los modelos a la hora de clasificar. En el caso de los modelos de regresión logística, los indicadores comúnmente utilizados se basan en la matriz de confusión, mientras que, dentro del enfoque bayesianos, se suele utilizar el estimador WAIC.

2.3.1. Matriz de Confusión

Una *Matriz de Confusión*, es una tabla de contingencia, utilizada como una medida de desempeño adoptada para afrontar el problema de clasificación. La matriz refleja en las filas la realidad y en las columnas lo que predice el modelo, estos resultados dependen fuertemente del umbral que seleccionamos (Powers, 2008).

El cuadro 2.1 representa una matriz de confusión, donde los valores predichos se encuentran en las columnas y los observados en las filas:

Cuadro 2.1: Matriz de confusión.

		Predicho	
		Positivos	Negativos
Observado	Positivos	VP	FN
	Negativos	FP	VN

- VP: Verdaderos positivos (casos positivos correctamente clasificados).

- VN: Verdaderos negativos (casos negativos correctamente clasificados).
- FP: Falsos positivos (casos negativos incorrectamente clasificados como positivos).
- FN: Falsos negativos (casos positivos incorrectamente clasificados como negativos).

A partir de estos valores, se define el *error global* o tasa de error, como la suma de casos mal clasificados, es decir los falsos negativos más los falsos positivos sobre la suma de los casos totales.

$$Error\ Global = \frac{(FP + FN)}{Total} \tag{2.8}$$

Se busca que el resultado del error en el modelo sea lo más bajo posible.

Por otro lado, la *sensibilidad* o también conocido como *recall*, es la proporción de casos verdaderos positivos que son correctamente predichos como positivos. La sensibilidad se calcula como:

$$Sensibilidad = \frac{VP}{(VP + FN)} \tag{2.9}$$

En contraste, la *especificidad* o *inverse recall*, es la proporción de casos negativos que son correctamente predichos como negativos, también se conoce como tasa de verdaderos negativos y se puede calcular como:

$$Especificidad = \frac{VN}{(FP + VN)} \tag{2.10}$$

Por último, la *precisión* o confianza, denota la proporción de casos predichos como positivos que son correctamente clasificados como verdaderos positivos. Este indicador es en el que se suelen enfocar aplicaciones dentro del aprendizaje automático, la minería de datos y la recuperación de información. La precisión se define como:

$$Precisión = \frac{VP}{(VP + FP)} \tag{2.11}$$

2.3.2. Widely Applicable Information Criterion

El *Widely Applicable Information Criterion* (WAIC) introducido por Watanabe (2010), es un indicador propio del enfoque bayesiano cuya finalidad es estimar el valor esperado fuera de la muestra (ecuación 2.12) para un nuevo conjunto de datos producido por el verdadero proceso generador de datos. Llamamos f al verdadero proceso generador de datos, y a los datos observados, y $\tilde{y}$ a datos futuros o conjuntos de datos alternativos que podrían haberse observado. El ajuste predictivo fuera de la muestra para un nueva

observación $\tilde{y}_i$ utilizando el score logarítmico es,

$$\log p_{post}(\tilde{y}) = \log \mathbb{E}_{post}(p(\tilde{y}_i|\theta) = \log \int p(\tilde{y}_i|\theta)p_{post}(\theta)d\theta$$

donde $p_{\text{post}}(\tilde{y}_i)$ es la densidad predictiva para $\tilde{y}_i$ inducida por la distribución posterior $p_{\text{post}}(\theta)$. La distribución p_{post} y la $\mathbb{E}_{\text{post}}$ denotan cualquier probabilidad o esperanza que promedie sobre la distribución posterior de θ.

Los datos futuros $\tilde{y}_i$ son en sí mismos desconocidos, por lo que definimos la densidad predictiva logarítmica puntual esperada fuera de la muestra como,

$$\begin{aligned} \text{elpd} &= \textit{logaritmo de la densidad predictiva puntual} \\ &\qquad \text{para una nueva observación} \\ &= \mathbb{E}_f(\log p_{post}(\tilde{y}_i)) = \int (\log p_{post}(\tilde{y}_i) f(\tilde{y}_i) d\tilde{y} \end{aligned}$$

En la literatura de aprendizaje automático, esto a menudo se llama densidad predictiva logarítmica media.

Para mantener la comparabilidad con el conjunto de datos dado, se puede definir una medida de precisión predictiva para los n puntos de datos tomados individualmente:

$$\begin{aligned} \text{elppd} &= \textit{logaritmo de la densidad predictiva puntual esperada} \\ &\qquad \text{para un nuevo conjunto de datos} \\ &= \sum_{i=1}^{n} \mathbb{E}_f(\log p_{\text{post}}(\tilde{y}_i)) \end{aligned} \tag{2.12}$$

La ventaja de usar una medida puntual, en lugar de trabajar con la distribución predictiva posterior conjunta $p_{\text{post}}(\tilde{y})$, radica en la conexión del cálculo puntual con la validación cruzada, lo que permite un enfoque general para la aproximación del ajuste fuera de la muestra utilizando los datos disponibles.

En la práctica, el parámetro θ no se conoce, por lo que no podemos conocer la densidad predictiva logarítmica $\log p(y|\theta)$. Por las razones anteriores, se trabaja con la distribución posterior, $p_{\text{post}}(\theta) = p(\theta|y)$, para resumir la precisión predictiva del modelo ajustado a los datos mediante,

$$\begin{aligned} \text{lppd} &= \textit{densidad predictiva puntual logarítmica} \\ &= \log \prod_{i=1}^{n} p_{\text{post}}(y_i) = \sum_{i=1}^{n} \log \int p(y_i|\theta)p_{\text{post}}(\theta)d\theta. \end{aligned} \tag{2.13}$$

Para calcular esta densidad predictiva en la práctica, podemos evaluar la esperanza

utilizando muestras de $p_{\text{post}}(\theta)$, denominadas θ^s, $s = 1, \ldots, S$:

$$\begin{aligned} \text{lppd calculada} &= \textit{densidad predictiva puntual logarítmica calculada} \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right) \end{aligned} \tag{2.14}$$

Por lo general, asumimos que el número de simulaciones S es lo suficientemente grande como para capturar completamente la distribución posterior; por lo tanto, nos referiremos indistintamente al valor teórico en 2.13 y al cálculo en 2.14 como la densidad predictiva puntual logarítmica o lppd de los datos.

La lppd de los datos observados y es una sobreestimación de la elppd para datos futuros (ecuación 2.12). Por lo tanto, el procedimiento es comenzar calculando 2.14 y luego aplicar algún tipo de corrección de sesgo para obtener una estimación razonable de 2.12.

Se han propuesto dos ajustes en la literatura, ambos basados en cálculos puntuales y pueden considerarse aproximaciones a la validación cruzada. El primer enfoque es una diferencia, similar a la utilizada para construir DIC mientras que el segundo enfoque utiliza la varianza de los términos individuales en la densidad predictiva logarítmica sumada sobre los n puntos de datos, este es el enfoque que desarrollamos y utilizamos:

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{var}_{\text{post}} \left(\log p(y_i|\theta) \right)$$

Para calcularla, se determina la varianza posterior de la densidad predictiva logarítmica para cada punto de datos y_i, es decir, $V_{s=1}^{S} \log p(y_i|\theta_s)$, donde $V_{s=1}^{S}$ representa la varianza muestral. La suma de todos los puntos de datos y_i proporciona el número efectivo de parámetros:

$$p_{\text{WAIC}} \text{ calculado} = \sum_{i=1}^{n} V_{s=1}^{S} \left(\log p(y_i|\theta^s) \right).$$

Luego, se utiliza p_{WAIC} como una corrección de sesgo:

$$\text{elppd}_{WAIC} = \text{lppd} - p_{\text{WAIC}}$$

Al igual que con AIC y DIC, definimos WAIC como -2 veces la expresión anterior para que esté en la escala de devianza y sea comparable con AIC, DIC y otras medidas de devianza.

En comparación con AIC y DIC, WAIC tiene la propiedad deseable de promediar sobre la distribución posterior en lugar de condicionar sobre una estimación puntual. Esto es especialmente relevante en un contexto predictivo, ya que WAIC evalúa las predicciones

que realmente se utilizan para nuevos datos en un contexto bayesiano.

2.4. Estadística Espacial: Datos de Área

La *Estadística Espacial* es la rama de la estadística que analiza datos georeferenciados. El objeto de estudio es, a grandes rasgos, analizar la existencia de dependencia espacial, y su incorporación en modelos estadísticos. Según el tipo de datos con el que se trabaje se definen tres subdisciplinas:

- **Geoestadística:** corresponde al análisis de datos continuos, y se trata de predecir los valores de la variable de interés en una superficie continua. Es muy utilizado en las Ciencias Agrarias, Climatológicas y en la Ecología. Se suele trabajar con distancias euclideas.
- **Análisis de patrones de puntos:** el objeto de interés es el lugar de ocurrencia de los eventos. Se trata de procesos estocásticos en el espacio, y se utiliza ampliamente en el área de la Epidemiología.
- **Datos de Área:** en este caso los datos son discretos en el espacio. Se cuenta con polígonos (unidades administrativas) con información agregada de una variable de interés. Por lo general no existe una medida de distancia, y la "cercanía" se encuentra definida por el investigador. Es muy utilizado en economía y de allí también su denominación como "Econometría Espacial".

Además de las áreas de conocimiento anteriormente mencionadas, es fundamental destacar el papel de la *Econometría Espacial*, una subdisciplina que se entrelaza estrechamente con la *Estadística Espacial*. Mientras que la Estadística Espacial se enfoca en el análisis y modelado de datos geográficos para identificar patrones y dependencias espaciales, la Econometría Espacial se centra en incorporar la dimensión espacial dentro de modelos econométricos. Esto es particularmente relevante en el análisis de datos económicos, donde las interacciones espaciales y las dependencias entre unidades geográficas (como ciudades, regiones o países) pueden influir significativamente en fenómenos económicos. La Econometría Espacial permite entender cómo y en qué medida las variables económicas en una ubicación son afectadas por las de sus vecinos, abordando así problemas de autocorrelación espacial, heterogeneidad espacial y otros desafíos inherentes a los datos geoespaciales (Anselin, 1988).

2.4.1. Polígonos

Cuando se trabaja con datos de área las unidades de análisis son polígonos fijos, y la variable de estudio es una característica asociada a cada polígono. Un *polígono* es una figura plana cerrada con tres o más lados y ángulos. Los rectángulos son una de las formas poligonales más sencillas. Sin embargo, muchos de los polígonos del análisis espacial (por ejemplo, las manzanas, barrios y departamentos) son más complejos que simples rectángulos. En algunos casos, interesa definir el centro de un polígono para definir una noción de distancia entre polígonos. Una definición de centro es el centroide de un polígono definido por el centro de masa (o punto de equilibrio) del polígono (Waller y Gotway, 2004).

Esta noción de vecindad es primordial cuando se trata de cuantificar la autocorrelación espacial, ya que la estructura de esta autocorrelación se organiza en función de la proximidad entre polígonos en una superficie particionada. La "proximidad" de los mismos depende de la definición de vecinos.

La creación de pesos espaciales es un paso necesario al utilizar datos de área en tanto que estos permiten verificar la existencia patrones espaciales en los datos o incluso en los residuos de los modelos. En primer lugar, se definen las relaciones entre observaciones, se selecciona un criterio de vecindad, y luego se asignan pesos a los enlaces de vecinos identificados.

2.4.2. Matriz de Vecinos y Pesos Espaciales

El paquete `spdep` (R. Bivand y Wong, 2018) permite generar la matriz de vecinos y pesos espaciales en el software de trabajo. En este informe, los vecinos son definidos por una matriz de conectividad binaria, en donde el elemento b_{ij} de la matriz $\boldsymbol{B}$ es 1 si las regiones, i, j comparten un borde, entonces:

$$b_{ij} = \begin{cases} 1 & \text{si las regiones } i \text{ y } j \text{ comparten un borde} \\ 0 & \text{en otro caso} \end{cases}$$

Una vez establecida la matriz de vecinos, se procede a asignar pesos espaciales a cada relación y así crear la matriz de pesos espaciales.

La *matriz de pesos espaciales* $\boldsymbol{W}$, también conocida como matriz de proximidad espacial,

cuantifica en sus elementos w_{ij}, la cercanía espacial entre las regiones i, j y, en conjunto, los w_{ij} definen una estructura de vecindad en toda la zona de estudio.

Existen múltiples estilos de pesos o ponderaciones para asignar a las unidades de análisis, pero este informe utiliza las ponderaciones mas comunes, la estandarización por filas, la cual asegura la comparabilidad entre polígonos con diferente número de vecinos.

Estas ponderaciones, al estar estandarizadas por filas, hace que varíen entre el cociente de uno con el mayor número de vecinos y el cociente de uno con el menor número de vecinos que posee la unidad de análisis. Adicionalmente, las sumas de los pesos de cada unidad de área suman uno. Una consecuencia de esto último, es que las ponderaciones de los enlaces que se originan en zonas con pocos vecinos son mayores que los que se originan en zonas con muchos vecinos, lo que tal vez potencie involuntariamente las entidades de área situadas en el borde de la zona de estudio. Esta matriz no es simétrica, pero es cercana a la representación simétrica (R. S. Bivand, Pebesma, y Gómez-Rubio, 2008).

El elemento w_{ij} de la matriz $\boldsymbol{W}$, estandarizado por filas es:

$$w_{ij} = \frac{b_{ij}}{\sum_{j=i}^{n} b_{ij}}$$

2.4.3. Autocorrelación Espacial

"La Autocorrelación Espacial es la correlación entre valores de una misma variable, estrictamente atribuible a una ubicación en una superficie bidimensional, introduciendo una alteración del supuesto de independencia entre las observaciones, fundamento de la estadística clásica." (Diniz-Filho, Bini, y Hawkins, 2003)

La *Autocorrelación Espacial* implica la presencia de correlación entre el mismo tipo de medición registrada en diferentes unidades de área. Un índice global de autocorrelación espacial es utilizado para resumir el grado en que observaciones similares tienden a ocurrir cerca unas de otras. Normalmente, los valores extremos positivos sugieren una autocorrelación espacial positiva, mientras que los valores en la dirección opuesta sugieren una autocorrelación espacial negativa (Waller y Gotway, 2004).

La mayoría de los índices de autocorrelación comparten una estructura básica común. En este sentido, se calcula la similitud de los valores en las ubicaciones i y j, y luego se

pondera esta similitud por la proximidad de dichas ubicaciones. Similitudes con valores elevados y con un peso alto (es decir, valores similares próximos) dan lugar a valores altos del índice. Por otro lado, similitudes bajas con un peso elevado (es decir, valores diferentes muy próximos) dan lugar a valores bajos. La notación sim_{ij} denota la similitud entre los valores de los datos y_i y y_j , mientras que w_{ij} o b_{ij} denota un peso que describe la proximidad entre las ubicaciones i y j, para $i, j = 1, \ldots, n$. Se pueden utilizar ambas matrices de forma equivalente, de aca en adelante se utiliza la matriz $\boldsymbol{W}$ a modo de ejemplo. La mayoría de los índices globales de autocorrelación espacial tienen la siguiente forma:

$$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\, sim_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \tag{2.15}$$

Se observa que la medida de similitud sim_{ij} depende de variables aleatorias que definen las observaciones, y las w_{ij} (b_{ij}) son cantidades fijas basadas en la geografía subyacente de las regiones. Como tales, los valores sim_{ij} definen la estructura en la que se distribuye el índice, y las w_{ij} (b_{ij}) definen las estructuras espaciales de correlación. Por lo tanto diferentes medidas de similitud definen diferentes clases de índices. Los mas comunes son *I de Moran* y la *C de Geary*, mientras que diferentes medidas de proximidad (pesos espaciales) conducen a diferentes aplicaciones dentro de la clase de índice (Waller y Gotway, 2004).

2.4.4. Índice de Moran

Las ponderaciones espaciales obtenidas en la matriz de pesos espaciales $\boldsymbol{W}$ ($\boldsymbol{B}$), empiezan a ser de utilidad, con el principal motivo de comprobar la presencia de autocorrelación espacial.

El *Índice de Moran* utiliza como medida de similaridad a $sim_{ij} = (y_i - \overline{y})(y_i - \overline{y})$, siendo $\overline{y}$ la media de la variable de interés. El mismo, se calcula como una relación entre el producto de la variable de interés y su rezago espacial, con el producto cruzado de la variable de interés y su rezago espacial, ajustado por los pesos espaciales utilizados:

$$I = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \overline{y})(y_j - \overline{y})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \frac{n}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \tag{2.16}$$

siendo y_i la i-ésima observación, w_{ij} es el peso espacial del vínculo entre i y j y n es el

total de observaciones. Centrarse en la media equivale a afirmar que el modelo correcto tiene una media constante, y que cualquier patrón restante luego de corregir por $\bar{y}$ se debe a las relaciones espaciales codificadas en los pesos espaciales.(R. S. Bivand y cols., 2008).

El *Test de Moran* consiste en una prueba de hipótesis, donde la hipótesis nula considera que no existe autocorrelación espacial, mientras que en la alternativa, sí existe. Cliff y Ord (1973), proponen dos alternativas para especificar la distribución bajo la hipótesis nula. Asumir que las y_i son realizaciones independientes e idénticamente distribuidas (i.i.d.) de una distribución normal con μ y σ^2 constante para todos los polígonos; o asumir que en cada polígono los valores de y_i son equiprobables, es decir que las realizaciones son permutaciones al azar del vector de observaciones Y, este supuesto es conocido como *randomization assumption*. A partir de estos supuestos es posible obtener resultados inferenciales a partir de la estandarización del estadístico I.

El valor esperado para el índice de Moran es $-1/(n-1)$ para los casos centrados en la media. El contraste de hipótesis consiste en comparar si $I > \mathbb{E}(I)$. En ese caso, los resultados de la prueba reflejan que no hay rastro de dependencia espacial con las ponderaciones w_{ij} utilizadas.

2.5. Modelos Mixtos

Muchos modelos estadísticos comunes pueden expresarse como modelos lineales que incorporan tanto efectos fijos como aleatorios. Los primeros refieren a parámetros asociados a una población entera o a un cierto conjunto de niveles repetibles de factores experimentales, mientras que los segundos se asocian a unidades experimentales individuales extraídas al azar de una población. Un modelo que incorpora efectos de los dos tipos se denomina modelo de efectos mixtos (Pinheiro y Bates, 2006).

Los modelos de efectos mixtos se utilizan para describir relaciones entre una variable de respuesta y variables explicativas en datos que se agrupan jerárquicamente según uno o más factores de clasificación. Al asociar efectos aleatorios comunes a observaciones que comparten el mismo nivel de un factor de clasificación, los modelos mixtos representan de manera flexible la estructura de covarianza inducida por la agrupación de los datos. El caso mas simple de un modelo de efectos fijos y aleatorios se denota como:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + b_i + \epsilon_{ij}, \quad i = 1, \ldots, M, \quad j = 1, \ldots, n_i,$$

donde y_{ij} es la variable de respuesta observada para la observación j en el grupo i, x_{ij} es el valor de la variable explicativa para la observación j en el grupo i, β_0 y β_1 representan los parámetros poblacionales, comunes a los i grupos, b_i es una variable aleatoria propia de todas las unidades experimentales del grupo i que representa la desviación de la constante del grupo i respecto de la constante poblacional β_0 y ϵ_{ij} es otra variable aleatoria que representa la desviación en la observación j y el grupo i sobre la media del grupo i.

Los modelos mixtos generalizados (GLMMs) extienden los modelos mixtos lineales al permitir que la variable de respuesta tenga una distribución de la familia exponencial. En el contexto de la regresión logística, un GLMM para una variable de respuesta Bernoulli puede expresarse como:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + b_i + \beta_1 x_{ij}$$

donde p_{ij} es la probabilidad de éxito para la observación j en el grupo i, x_{ij} es la variable explicativa asociadas a la observación j en el grupo i, β_1 es el coeficiente asociado a la variable explicativa y b_i es el efecto aleatorio propio de las unidades pertenecientes al grupo i, el cual se interpreta como desviación respecto de la constante poblacional β_0.

2.5.1. Modelo Autorregresivo Condicional (CAR)

Cuando los datos de área tienen una estructura espacial a partir de la cual es natural pensar que observaciones de regiones vecinas presentan una correlación más alta que las regiones distantes, esta correlación puede ser tenida en cuenta utilizando la clase de modelos espaciales llamados modelos CAR (Conditional Auto-Regressive), introducidos por Besag (1974).

Un modelo *Autorregresivo Condicional* (CAR), es un tipo de modelo estadístico utilizado para analizar datos espaciales que exhiben correlación espacial. Este modelo asume que el valor de la variable de interés en una ubicación geográfica está condicionalmente correlacionado con los valores de la misma variable en ubicaciones vecinas. Por lo tanto permite capturar la dependencia espacial. Los modelos CAR se utilizan cuando los datos consisten en una única medida agregada por unidad de área (ya sea un valor binario, de conteo o continuo) o como distribución de efectos aleatorios asociados a datos jerárquicos.

Para un conjunto de n polígonos, la relación entre los mismos se representa en la matriz de adyacencia o vecinos $\boldsymbol{W}$, de dimensión $n \times n$ (capítulo 2.4.2). Para la relación de vecindad binaria, escrita $i \sim j$ donde $i \neq j$, las entradas en la matriz $\boldsymbol{W}$ son 1 si las

regiones i y j son vecinas, y 0 en caso contrario.

Dado un conjunto de observaciones obtenidas en n polígonos diferentes de una región, las interacciones espaciales entre un conjunto de unidades espaciales se pueden modelar condicionalmente como una variable aleatoria espacial $\mathbf{\Phi}$, que es un vector de longitud n, entonces $\mathbf{\Phi} = (\phi_1, ..., \phi_n)^T$.

En las distribuciones condicionales completas, cada ϕ_i está condicionado a la suma de los valores ponderados de sus vecinos ($w_{ij}\phi_j$) y tiene una varianza desconocida

$$\phi_i|\phi_j, j \neq i, \sim \mathcal{N}\left(\alpha \sum_{j=1}^{n} w_{ij}\phi_j, \sigma^2\right)$$

Besag (1974) demostró que la especificación conjunta correspondiente de $\mathbf{\Phi}$ es una variable aleatoria normal multivariada centrada en 0; y la matriz de covarianza se especifica como una matriz de precisión $\boldsymbol{Q}$, que es la inversa de la matriz de covarianza $\mathbf{\Sigma}$, es decir, $\mathbf{\Sigma} = \boldsymbol{Q}^{-1}$, de modo que

$$\mathbf{\Phi} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{Q}^{-1}\right) \tag{2.17}$$

Para que la variable aleatoria normal multivariada estándar $\mathbf{\Phi}$ tenga una densidad de probabilidad conjunta adecuada, la matriz de precisión $\boldsymbol{Q}$ debe ser simétrica y definida positiva. Esto se logra construyendo la matriz de precisión $\boldsymbol{Q}$ a partir de la matriz de adyacencia $\boldsymbol{W}$:

$$\boldsymbol{Q} = \left(\frac{1}{\sigma}\boldsymbol{D}(\mathbb{I} - \alpha\boldsymbol{B})\right) \tag{2.18}$$

donde $\boldsymbol{W}$ la matriz de adyacencia $n \times n$, cuyas entradas en la diagonal principal son cero y los elementos fuera de la diagonal son 1 si las regiones i y j son vecinas, y 0 en caso contrario. $\boldsymbol{D}$ es una matriz diagonal de dimensión $n \times n$ donde las entradas en la diagonal principal son el número de vecinos de la región i, α controla el grado de correlación espacial; $\alpha = 0$ implica independencia espacial y $\alpha = 1$ implica correlación espacial completa. $\boldsymbol{B}$ es la matriz de adyacencia escalada $\boldsymbol{D}^{-1}\boldsymbol{W}$. Y finalmente $\mathbb{I} = \mathbb{I}_n$ es una matriz identidad $n \times n$.

Cuando α pertenece al intervalo $(0, 1)$, la matriz de precisión $\boldsymbol{Q}$ es definida positiva, por lo tanto, la distribución conjunta $\mathbf{\Phi}$ es no degenerada.

2.5.2. Modelo Intrínseco Autorregresivo Condicional (ICAR)

Los modelos *Intrinsic Conditional Auto-Regressive* (ICAR), son un caso particular de los modelo CAR, donde $\alpha = 1$. Esto supone una correlación espacial completa entre regiones. La distribución conjunta del modelo ICAR se deriva de la distribución conjunta del modelo CAR de la siguiente manera:

Dado que $\boldsymbol{B} = \boldsymbol{D}^{-1}\boldsymbol{W}$, la expresión $[\boldsymbol{D}(\mathbb{I} - \alpha\boldsymbol{B})]$ se simplifica $[\boldsymbol{D} - \boldsymbol{W}]$, donde $\alpha = 1$ es omitido.

La matriz resultante $[\boldsymbol{D} - \boldsymbol{W}]$ es singular, lo que hace que la distribución conjunta asociada al modelo ICAR sea:

$$\boldsymbol{\Phi} \sim \mathcal{N}\left(0, \left(\frac{1}{\sigma}\boldsymbol{D} - \boldsymbol{W}\right)^{-1}\right) \tag{2.19}$$

Si bien el modelo ICAR no es un modelo generativo, en el sentido de que no existe un procedimiento automático para agregar un nuevo polígono, por este motivo no es posible utilizarlo como modelo para los datos. No obstante, si se puede utilizar como distribución a priori para los efectos aleatorios del modelo mixto.

La distribución condicional correspondiente es:

$$p(\phi_i|\phi_j) = \mathcal{N}\left(\frac{\sum_{i\neq j}\phi_i}{D_{i,i}}, \frac{\sigma_i^2}{D_{i,i}}\right), \quad j \neq i$$

donde $D_{i,i}$ es el número de vecinos de la región n_i. La variable aleatoria ϕ_i que representa la variabilidad espacial individual para la región n_i, se distribuye normal con una media igual al promedio de sus vecinos y su varianza disminuye a medida que aumenta el número de vecinos.

La distribución conjunta del vector Φ, puede reescribirse de la siguiente forma:

$$p(\Phi) \propto exp\left\{\frac{-1}{2}\sum_{i\neq j}(\phi_i - \phi_j)^2\right\}$$

De la derivación anterior, vemos que la distribución conjunta no es identificable; agregar cualquier constante a los elementos de $\boldsymbol{\Phi}$ deja la distribución conjunta sin cambios. Sin embargo, imponer la restricción $\sum_i \phi_i = 0$ resuelve este problema (Morris y cols., 2019).

A partir de las variantes CAR e $ICAR$ es posible construir un modelo logístico jerárquico

espacial de la siguiente forma:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = x_{ij}^T\beta + \phi_i$$

donde p_{ij} es la probabilidad de éxito para la observación j en el grupo i, x_{ij} es un vector de variables explicativas asociadas a la observación j en el grupo i, β es el vector de coeficientes y ϕ_i es la variable aleatoria que espacialmente correlacionada, propia de la región o grupo i (sujeta además a que $\sum_i \phi_i = 0$).

2.6. Metodología

En esta sección se explicita el flujo de trabajo que se utiliza para el análisis de los datos (3), comenzando con una descripción de los mismos y luego presentando los pasos necesarios para la obtención de resultados (4). Se realiza un análisis descriptivo sobre las variables relevantes del problema tal y como la proporción de bimestres impagos, la información de pago del último bimestre y la ubicación geográfica de las observaciones. Con el objetivo de encontrar el mejor modelo logístico que explique la probabilidad de pago del bimestre próximo, se utilizan como variables explicativas aquellas que se relacionan con las características de las unidades observadas, categoría, información de pago de bimestres anteriores, antigüedad, la ubicación geográfica de la observación y más.

En primer lugar, se incluyen todas las variables explicativas al modelo, evaluando su significancia individual y conjunta a través de pruebas de hipótesis, y midiendo su capacidad de predicción mediante indicadores de desempeño.

La variable dependiente Y es *BIM6* que solo toma dos valores, este es el motivo por el cual modelamos a Y como una variable aleatoria *Bernoulli* y la relacionamos con el conjunto de variables explicativas mediante un modelo de regresión logística:

$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta}$$

donde $\text{logit}(\mathbf{p})$ es un vector columna de dimensión $n\times 1$ de la forma $\left(\log(\frac{p_1}{1-p_1}), \ldots, \log(\frac{p_n}{1-p_n})\right)$ siendo $p_i, i = 1, \ldots, n$ las probabilidades de que Y_i tome el valor 1 (pago) para cada observación; $\mathbf{X}$ es una matriz de dimensión $n \times (p+1)$, con p variables explicativas (correspondientes a las n observaciones) donde la primera columna es un vector de unos para el término constante, $\boldsymbol{\beta}$ es un vector columna de dimensión $(p+1) \times 1$ que representa los coeficientes que cuantifican la relación entre cada predictor $X_{i,j}, j = 0, 1, \ldots, p$ y el

logaritmo de los *odd*.

El siguiente conjunto de modelos son evaluados desde el enfoque bayesiano. Se incluyen las variables que resultan significativas en el enfoque frecuentista y dado conocimientos previos, se asignan previas débilmente informativas sobre los parámetros que representa la incertidumbre sobre los mismos.

$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta}, \qquad \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

donde $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ representa la distribución previa para $\boldsymbol{\beta}$. La distribución $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ denota una distribución normal multivariada donde $\boldsymbol{\mu}_0$ es un vector de medias previas para $\boldsymbol{\beta}$ y $\boldsymbol{\Sigma}_0 = \text{diag}(\sigma^2_{0,1}, \sigma^2_{0,2}, \ldots, \sigma^2_{0,p})$ es la matriz de covarianza previa para $\boldsymbol{\beta}$. La matriz de covarianza $\boldsymbol{\Sigma}_0$ es diagonal, lo que significa que los coeficientes β_j, $j = 1, \ldots, p$ son independientes entre sí.

Esta representación captura la esencia del modelo logístico desde un enfoque bayesiano con previas normales, manteniendo la notación matricial. La inferencia bayesiana se realiza entonces para obtener la distribución posterior de $\boldsymbol{\beta}$ dado los datos observados.

El último conjunto de modelos, son modelos mixtos, que incorporan a la representación anterior los efectos aleatorios espacialmente correlacionados. El motivo para emplear estos modelos tiene que ver con la estructura jerárquica de las observaciones. Las unidades muestrales (TDS) se encuentran anidadas en los CCZ. Así, se asignará un efecto aleatorio a cada CCZ y los efectos fijos operarán a nivel de las TDS.

$$\text{logit}(\mathbf{p}) = \mathbf{X}\boldsymbol{\beta} + \Phi, \qquad \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad \boldsymbol{\Phi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{Q}^{-1})$$

donde $\mathbf{X}$ es la matriz $n \times (p+1)$ de variables explicativas, $\boldsymbol{\beta}$ es el vector de coeficientes que representa los efectos fijos del modelo y $\boldsymbol{\Phi}$ es el vector de efectos aleatorios de dimensión $M \times 1$ que capturan la variabilidad espacial entre los M CCZ. La distribución previa para $\boldsymbol{\Phi}$ viene dada por una distribución normal multivariada centrada en un vector de ceros y matriz de covarianza $\boldsymbol{Q}^{-1}$ (veánse ecuaciones 2.5.1 y 2.5.2). La bondad de ajuste de estos modelos (así como la de sus análogos no espaciales) es evaluada mediante el índice de Moran. En última instancia, se compara la capacidad predictiva de los modelos mediante los indicadores de desempeño descritos en la sección 2.3.

3. Datos

Esta sección describe la información recopilada y extraída desde la base de datos de la Intendencia de Montevideo el día 23 de Diciembre de 2022; se describe el procesamiento y la unificación de los datos. Finalmente, se realiza un análisis descriptivo.

3.1. Base de datos

La base de datos, se compone de 292.427 observaciones (TDS) y 14 variables. Las variables son: número identificatorio (ID) de cada TDS, padrón, fecha de apertura, centro comunal zonal, categoría, cantidad de unidades ocupacionales, información sobre su pago en los últimos 6 bimestres, estado de actividad y por último la suma de la totalidad de bimestres con deuda.

A continuación se hace una descripción de las variables:

- "CUENTA_TS" es el número de cuenta corriente, es un número identificador que distingue cada Tarifa de Saneamiento. Particularmente todas las cuentas se conforman por la combinación de 7 dígitos numéricos, los cuales han sido asignados en forma creciente, de forma que reflejan la antigüedad de cada cuenta. El primer N° de cuenta corriente es el 2749869 creado el 15/05/2001 (fecha de apertura de la Tarifa de Saneamiento); por otro lado el N° de cuenta corriente más alto es el 5760291 con fecha de apertura el 01/12/2022 (fecha de inicio del próximo bimestre a facturar).
- "PADRON" es el número de padrón que individualiza al inmueble ya gráficamente representado y determinado en un plano. A partir de este número es posible identificar en un mapa la ubicación geográfica de la TDS.
- "F_APERTURA" es la fecha de apertura de la TDS. Codificada en formato dia/mes/año. La fecha de apertura es realizada al inicio de los bimestres, siendo en general las fechas de apertura 01/02, 01/04, 01/06, 01/08, 01/10 o 01/12. La fecha de apertura no es lo mismo que la fecha de creación por el hecho de que la fecha de apertura considera la apertura retrospectiva.
- "CCZ" es el Centro Comunal Zonal al que pertenece la TDS, precisamente una variable discreta que puede tomar valores únicos del 1-18.
- "CANT_UNID" es el número de unidades ocupacionales que componen la TDS, esta variable toma valores enteros positivos.

- "CATEGORIA" es una variable categórica, que representa el tipo de categoría de una TDS; las categorías son "DOMICILIARIO", "COMERCIAL/INDUSTRIAL", "GUBERNAMENTAL" o "ARTÍCULO 91".
- "BIM1 - BIM2 - BIM3 - BIM4 - BIM5 - BIM6" representan información de los pagos bimestrales del año 2022 de las 292.427 TDS. Dichas variables se codificaron con los valores "-1" que significa que la deuda en el bimestre no fue generada, "0" quiere decir bimestre impago y "1" significa bimestre pago.
    - "BIM1" corresponde al pago de los meses Diciembre 2021 y Enero 2022.
    - "BIM2" corresponde al pago de los meses Febrero-Marzo 2022.
    - "BIM3" corresponde al pago de los meses Abril-Mayo 2022.
    - "BIM4" corresponde al pago de los meses Junio-Julio 2022.
    - "BIM5" corresponde al pago de los meses Agosto-Setiembre 2022.
    - "BIM6" corresponde al pago de los meses Octubre-Noviembre 2022.
- "TDS_CORTADA" es un variable categórica que sirve para reflejar el estado de actividad de la tarifa, cuyos estados pueden ser inactiva cuando presenta valor "0" y activa cuando presenta valor "1". El estado de actividad de una TDS permite distinguir si la cuenta está generando deuda. Esto se da cuando en el próximo bimestre se generará un monto a pagar (cargo fijo + cargo variable); si está inactiva, el siguiente bimestre no generará deuda.
- "CANT_BIM_IMPAGOS" es una variable numérica que cuenta la cantidad de bimestres que no se han pagado desde la fecha de apertura de la TDS. Una unidad significa 1 bimestre es decir 2 meses de deuda impaga.

Adicionalmente, desde el Sistema de Información Geográfica (SIG, 2024) se extrajo información catastral, precisamente, geometrías y coordenadas geográficas sobre padrones, padrones anteriores y centros comunales zonales (CCZ). Esto será profundizado en la próxima sección.

3.2. Procesamiento de datos

La transformación de los datos brutos, a un conjunto de datos procesados y reproducibles, conlleva un procesamiento exhaustivo y minucioso, que incluye transformación de variables, filtrado de observaciones y referenciación geográfica. Una vez procesados, se le

asigna a cada observación su posición geográfica en el espacio convirtiéndolos en datos de área (2.4).

La especificación de la clase de cada variable a la hora de trabajar en `R`, es fundamental para el uso de sus funciones. Tal y como se mencionó en la descripción de la base de datos (3.1), las clases de las variables fueron asignadas de la forma siguiente:

- `numeric()` : CUENTA_TS , PADRON , CANT_UNID , CANT_BIM_IMPAGOS
- `date()` : F_APERTURA
- `factor()` : CCZ, CATEGORIA, BIM1, BIM2, BIM3, BIM4, BIM5, BIM6, TS_CORTADA

3.2.1. Filtrado de observaciones

Dentro de la muestra, existen pequeños conjuntos que no son de interés para el objetivo de este trabajo, por este motivo, las observaciones detalladas a continuación no son consideradas en el análisis.

La variable TS_CORTADA, permite distinguir entre las TDS activas y cortadas. Para la realización de este trabajo, no interesa considerar aquellas cuentas que no están activas. Esto se debe a que, al no estar generando deuda, la probabilidad de pago será nula por el hecho de que no corresponderá un pago. Por dicha razón se decide filtrarlas, dejando de lado 17.406 observaciones.

Por otro lado la variable CATEGORIA, contiene entre ellas, las categorías ARTICULO 91 y GUBERNAMENTAL, las cuáles representan el 0.01 % y 0.55 % respectivamente del total de TDS en Montevideo, véase la figura 3.1.

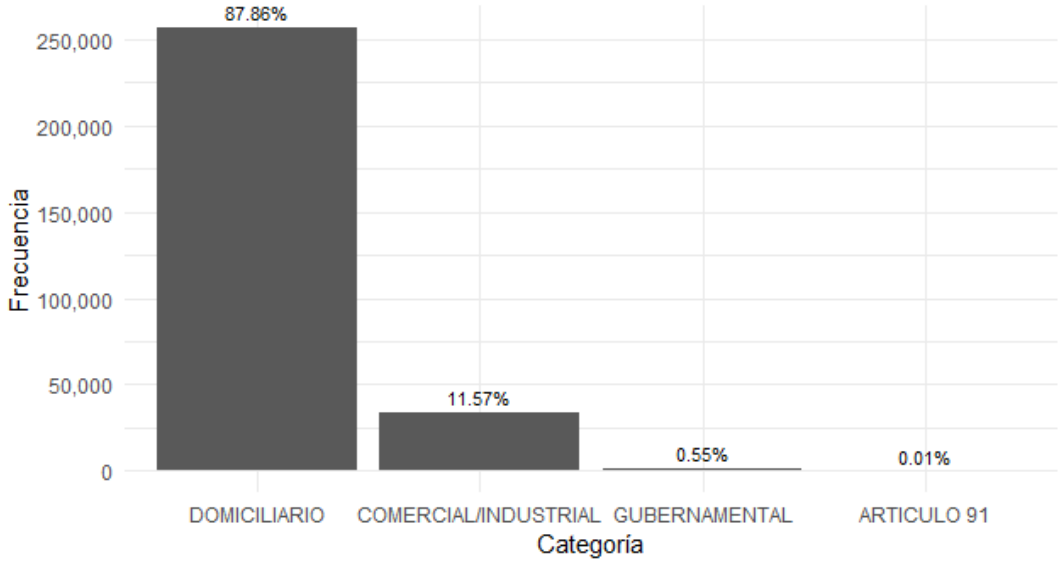

Figura 3.1: Gráfico de barras de la variable CATEGORIA.

Estas 2 categorías de TDS representan cuentas gubernamentales, tal y como, escuelas, ministerios, policlínicas, hospitales, jefaturas, etcétera, las cuáles tienen las particularidades de ser pagadas por sus entidades correspondientes. Por ejemplo, las escuelas son abonadas por ANEP, las policlínicas por el Ministerio de Salud Pública, las jefaturas o comisarías por el Ministerio del Interior. Cada entidad, genera sus pagos de la TDS de manera particular, ya sea abonando una vez al año, cada 2 años o incluso no abonando. Dadas las particularidades con las que generan los pagos y por tratase de la minoría, no son consideradas en el análisis. Luego de filtrar estas 1.434 TDS, permanecen 273.386 observaciones.

La variable BIM6, representa el pago del bimestre 6 de la tarifa, los valores que puede adoptar son *0* (bimestre impago), *1* (bimestre pago), o valor *-1* (sin generar deuda), este último valor lo contienen únicamente 155 de 273.386 observaciones. Estas tarifas de saneamiento, son consideradas cuentas bloqueadas en la Intendencia de Montevideo, un procesamiento manual de dichas TDS debe ser realizado por funcionarios para generar un desbloqueo y la generación de la deuda correspondiente. Por esta razón también son quitadas del análisis. De esta forma, la base de datos se compone por 273.231 observaciones.

3.2.2. Transformación de variables

Ciertas combinaciones de variables y transformaciones son realizadas para llevar a cabo el análisis y la creación del mejor modelo que se ajuste a los datos. A continuación se detallan los cambios realizados sobre las variables.

Tras el filtrado de observaciones en la variable BIM6, esta pasa a ser una variable binaria la cual será la variable a predecir.

La variable CANT_BIM_IMPAGOS, que representa la cantidad de bimestres impagos, debió ser transformada, por el hecho de que las TDS ya contaban con las deudas del próximo bimestre generadas; por tal razón, todas las cuentas que estaban al día, figuraban con un bimestre impago. Esto no parece correcto, debido a que dicho bimestre "impago" es el de diciembre-enero 2023 que correspondería pagar hasta el 01/02/23. A la fecha de recopilación de los datos (23/12/22), aún estaban en fecha de pagar. Para corregir este error, se seleccionaron las cuentas a las que se les generó deuda del próximo bimestre y se les restó una unidad, quedando con la cantidad de bimestres impagos correspondientes.

A partir de F_APERTURA se crea una nueva variable, ANTIGUEDAD, que representa la cantidad en años que tiene cada TDS a partir de su fecha de apertura.

Mediante la variable CANT_BIM_IMPAGOS y la variable ANTIGUEDAD, se crea la variable PROP_BIM_IMP, que representa el porcentaje de bimestres impagos, calculada como el cociente de las mismas, y multiplicada por 100. Esta variable, permite identificar observaciones con una de deuda considerable en relación a su antigüedad.

3.2.3. Referenciación geográfica

La referenciación geográfica es un proceso que consiste en identificar la posición en el espacio en la que cada observación se ubica. Cada observación contiene un padrón que ocupa un lugar en el espacio, esta sección explica como se asocia a cada padrón las coordenadas correspondientes.

Para ubicar geográficamente cada TDS en el mapa, fue necesario acudir a información adicional; para ello es que se descargaron archivos de extensión *.shp* (shapefiles) desde el Sistema de Información Geográfica de la Intendencia de Montevideo (SIG, 2024). El objetivo es asociar la geometría de los padrones que tengan TDS; dicha unión entre los

padrones de las TDS en la base de datos y los shapefiles de padrones descargados, fue realizada uniéndolos por número de padrón (variable en común entre ambos archivos) lo cual permitió georeferenciar cada TDS y asignarle su correspondiente geometría del padrón al que pertenecen.

El shapefile que contiene la información de los padrones, contiene información catastral sobre todos los padrones que existen a la fecha de descarga 28/12/2022 en Montevideo, incluyendo su geometría, mientras que por otro lado, un shapefile secundario contiene información catastral de padrones anteriores y que actualmente no existen por haber sufrido modificaciones prediales, ya sean fusiones o fraccionamientos.

La totalidad de observaciones, cuentan con un padrón asignado, sin embargo, un padrón puede contener numerosas TDS asignadas, por lo tanto, con el motivo de obtener la cantidad total de padrones únicos, se remueven los duplicados por padrón, obteniendo 170.991 padrones, a los que se busca asignar su correspondiente geometría.

A partir del shapefile de padrones, que contiene a la mayoría de los padrones que se estan buscando asociar su geometría, se logran seleccionar 170.674 polígonos (véase el capítulo 2.4.2) con padrones de interés. Mientras que en el shapefile secundario de padrones anteriores, se seleccionan 336 polígonos. Llama la atención, el hecho de que la suma de dichos resultados, resulte en 171.010, precisamente 19 más de los que se estaban buscando. Esto se debe a que hay padrones que se extienden a través de más de un polígono. Esto sucede cuando un padrón está compuesto por múltiple geometrías denominada `multipolygon` (Pebesma, 2018). Esta clase de padrones, está compuesta por más de un polígono.

De los 171.010 polígonos, 111 de ellos corresponden a 55 padrones de clase `multipolygon`; para hacer que dichos padrones tengan una única geometría, de forma que combine sus múltiples polígonos, se utilizó la función `st_union()` de la librería `sf` (Pebesma, 2018) contenida en el software `R` (R Core Team, 2023). Esta librería permite trabajar con datos espaciales. La función `st_union()` combina varias geometrías en una sola, creando así un único polígono a partir de la geometría de varios. A modo de ejemplo, la figura 3.2 refleja un padrón extraído del SIG. Este padrón esta compuesto por múltiples polígonos. Si se observa, el resultado de la búsqueda, devuelve 4 filas con el padrón 42563 las cuáles corresponden a cada uno de los puntos negros en la imagen haciendo referencia a cada polígono. Este padrón único cuenta con 4 polígonos, los cuales se fusionaron en uno solo.

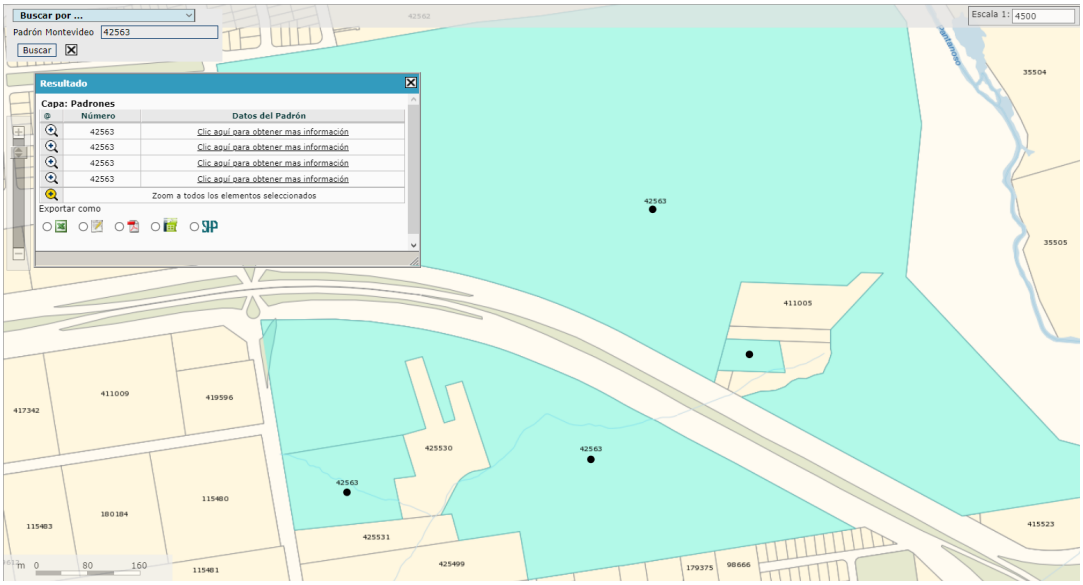

Figura 3.2: Padrón 42563 `multipolygon`, imagen extraída del SIG.

Tras lograr que todos los padrones cuenten con un único polígono de referencia como geometría, se pasó de una situación inicial con 171.010 polígonos a otra donde se contaba con 170.899 polígonos, que ahora corresponden, individualmente a un padrón.

No obstante, producto de los padrones de tipo `multipolygon`, se obtuvo un resultado inferior en 92 padrones en referencia a los 170.991 padrones iniciales que se empezaron buscando. Estos 92 padrones faltantes no figuran en los shapefiles descargadas del SIG, por el hecho de tratarse de padrones que estaban en proceso de modificaciones prediales como pudiesen ser fusiones o fraccionamientos. Estos procesos implican descartar el número de padrón original y terminan siendo asignados con un nuevo número de padrón, que por motivos de actualización aún no figuraban en los shapefiles.

La solución más conveniente fue excluir estos padrones que corresponden a 101 TDS. Contando finalmente con 292.231 TDS, con geometría asociada.

3.3. Análisis descriptivo

Las unidades de análisis son las TDS. Existen 292.427 observaciones distribuidas geográficamente por todo el departamento en 18 centros zonales comunales (CCZ). Dichos centros cuentan con distintas densidades de población y distintas infraestructuras de saneamiento lo cuál conlleva a cantidades heterogéneas de observaciones por centros comunales.

La figura 3.3 refleja la distribución de TDS por CCZ y CATEGORÍA.

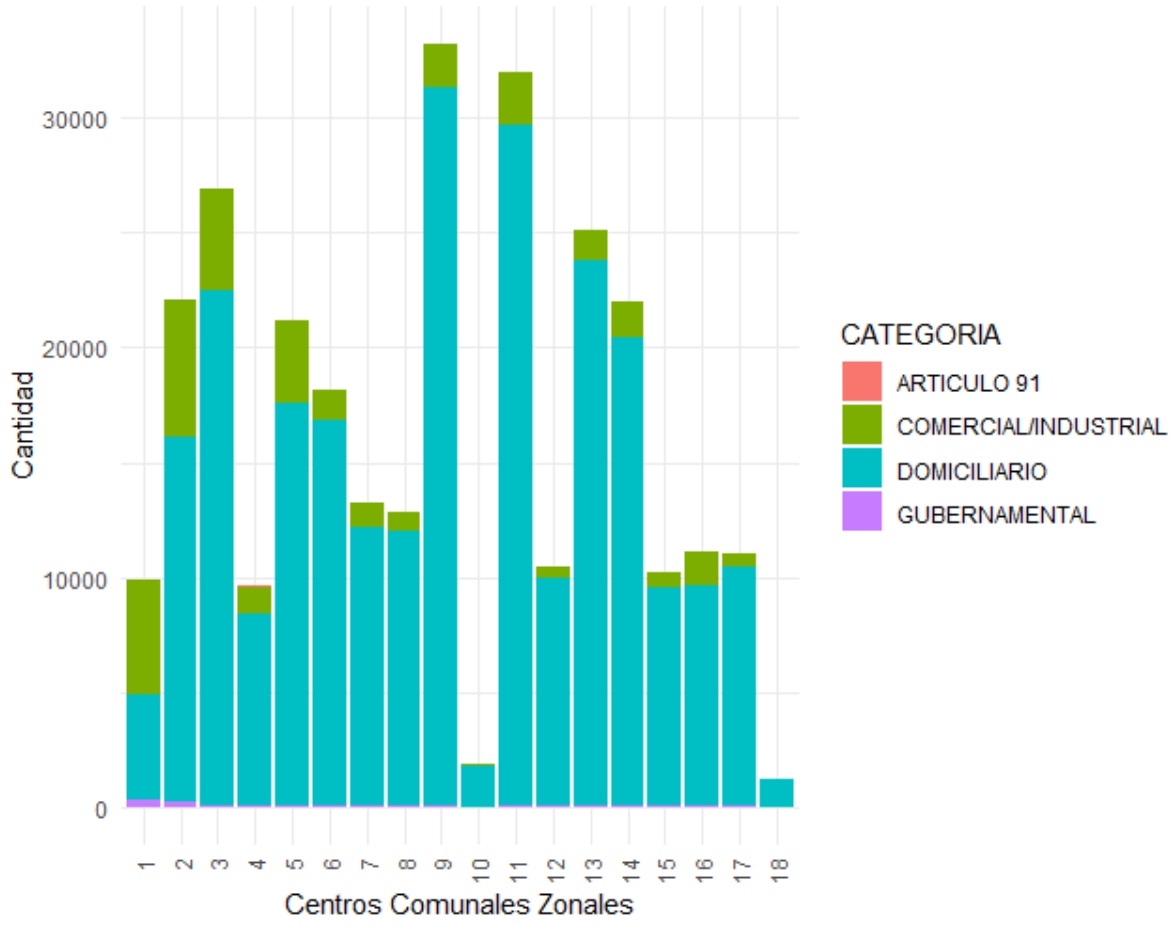

Figura 3.3: Gráfico de barras de las TDS por CCZ y CATEGORÍA.

Es inmediato observar que la cantidad de TDS por centro comunal zonal no es la misma para todo el departamento, al igual que la distribución de categorías. El centro comunal zonal 9 tiene la mayor cantidad con 33.192 cuentas. La categoría dominante es la "DOMICILIARIO" para todos los CCZ en excepción del CCZ 1, el cual posee una distribución equitativa de cuentas "COMERCIAL/INDUSTRIAL" y "DOMICILIARIO" con 4.939 y 4.578 respectivamente. Los CCZ 1,2 y 3 que incluyen barrios como Centro, Ciudad Vieja, Cordón, Parque Rodó, Tres Cruces, Aguada son los que poseen la mayor cantidad de cuentas "COMERCIAL/INDUSTRIAL". Por otro lado, las categorías "ARTICULO 91" y "GUBERNAMENTAL" son la minoría en todo el departamento. Más precisamente 1.619 observaciones pertenecen a la categoría "GUBERNAMENTAL" y tan solo 37 observaciones a la categoría "ARTICULO 91".

Otro de los factores que explica la distribución de cuentas totales por CCZ, ha de ser la

cercanía o lejanía a la zona metropolitana del departamento, al igual que el tamaño que pueda englobar cada CCZ. En la figura 3.4 podemos observar la misma distribución de cuentas por CCZ que en la figura 3.3, y es de notorio alcance, ver los distintos tamaños e irregularidades de los polígonos que representan cada CCZ. Centros como el 10 y el 18 son los que poseen menos cuentas y aunque sea contradictorio, son de los de mayor dimensión. Esta baja cantidad de TDS puede explicarse por su baja densidad de población, lejanía del área metropolitana del departamento o incluso por la presencia de zonas rurales, lo cual influye directamente en una baja infraestructura de saneamiento en dichos CCZ.

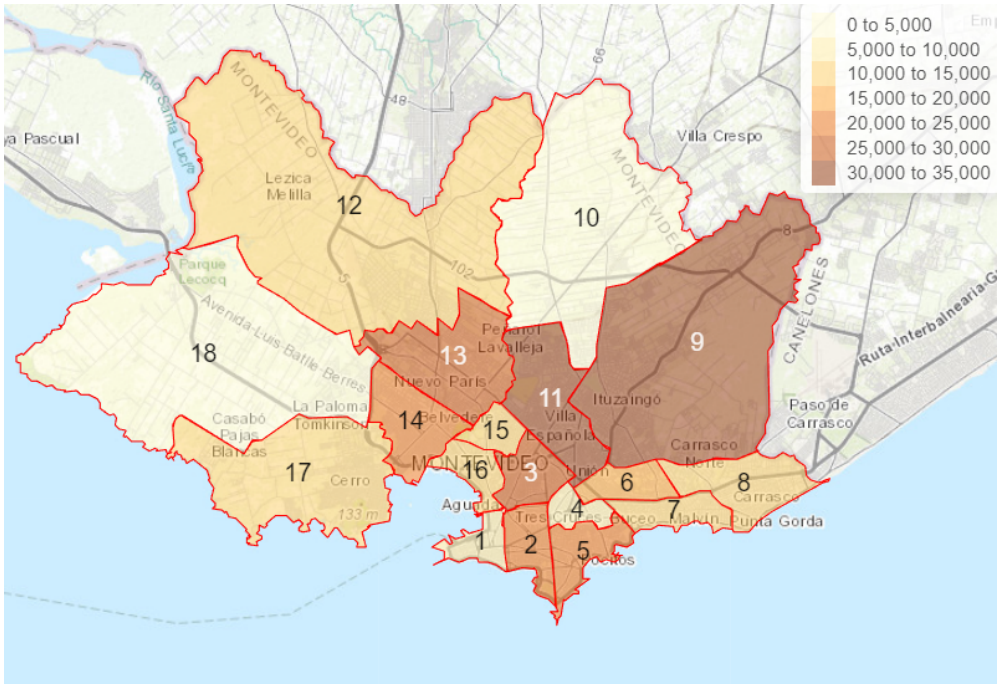

Figura 3.4: Mapa de Montevideo, con cuentas totales por CCZ.

La figura 3.5 refleja el porcentaje de cuentas con característica CORTADA por CCZ. Las cuentas *cortadas* son aquellas cuentas inactivas, que no están generando deuda monetaria. Esta información, es un primer indicador que se toma para saber si dicho usuario pueda o no estar haciendo fraude. Cuentas cortadas/inactivas no generan deuda y por lo tanto no deben abonar bimestres de tarifa de saneamiento. Las cuentas que poseen inactividad tienden a poseer deuda anterior, deuda impaga. Para este trabajo, estas cuentas inactivas no son consideradas por el hecho de que no es posible predecir el momento en el que se puedan reactivar y volver a generar deuda. Para esta investigación es de interés las cuentas que estén activas, sobre las cuales se busca inferir la probabilidad de pago del bimestre próximo.

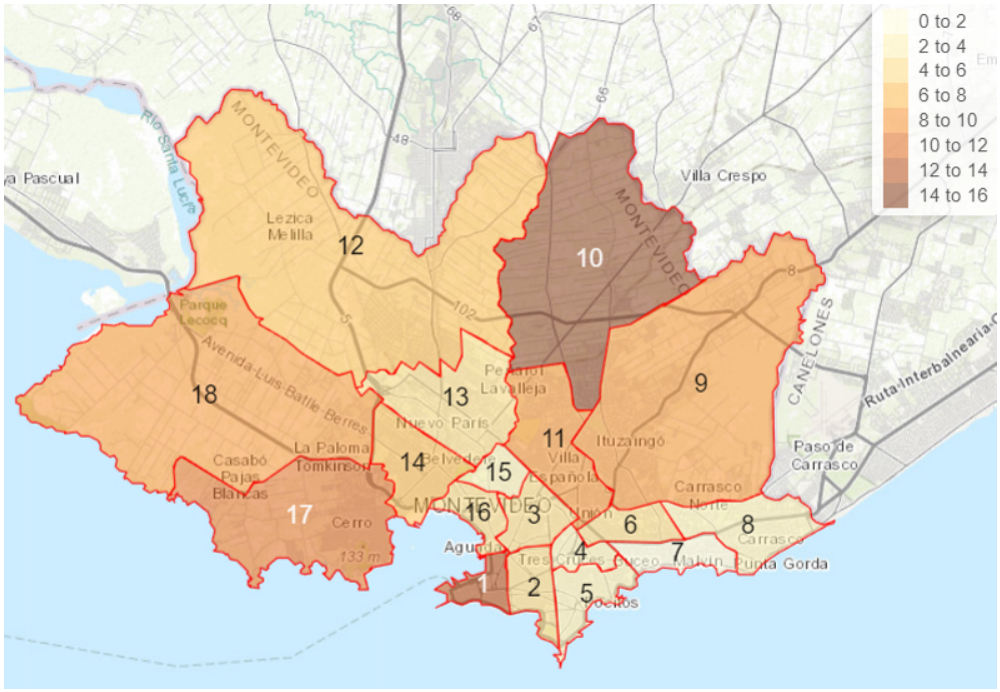

Figura 3.5: Mapa de Montevideo, con porcentaje de CORTADAS por CCZ.

Por otro lado, el porcentaje de cuentas activas con deuda impaga mayor a 6 bimestres (1 año) por CCZ puede visualizarse en la figura 3.6. Obsérvese como los CCZ 5, 7 y 8, son los que poseen mayor porcentaje de cuentas activas (y por lo tanto menor porcentaje de cuentas cortadas/inactivas), y a su vez presentan, menor morosidad en términos de cantidad de bimestres impagos según la figura 3.6. Este trabajo no tiene en cuenta el valor monetario que pueda o no significar esa deuda, si no que solo se considera en términos absolutos la cantidad de bimestres de deuda que tiene cada TDS.

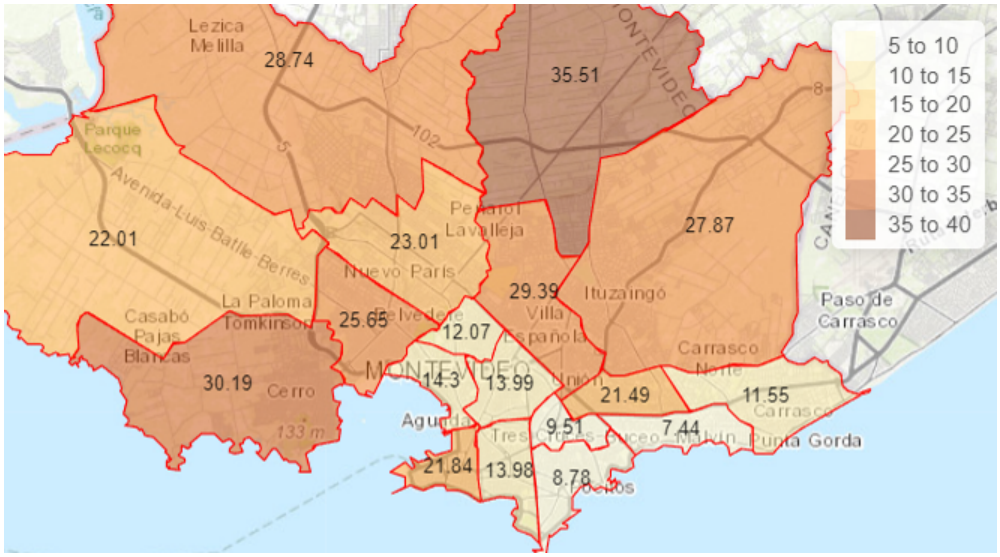

Figura 3.6: Mapa de Montevideo, con porcentaje de cuentas con deuda ≥ 1 año.

Cada unidad de análisis, cuenta con un padrón asociado, el cual permite asignarle sus coordenadas geográficas. En la figura 3.7 se observa la cantidad de cuentas totales por padrones de gran parte del CCZ 1. Este centro comunal abarca barrios como Ciudad Vieja, Aguada y Centro, barrios con alta densidad de población, que incluyen edificios y gran cantidad de comercios, por tal razón es esperable una gran cantidad de TDS por padrón.

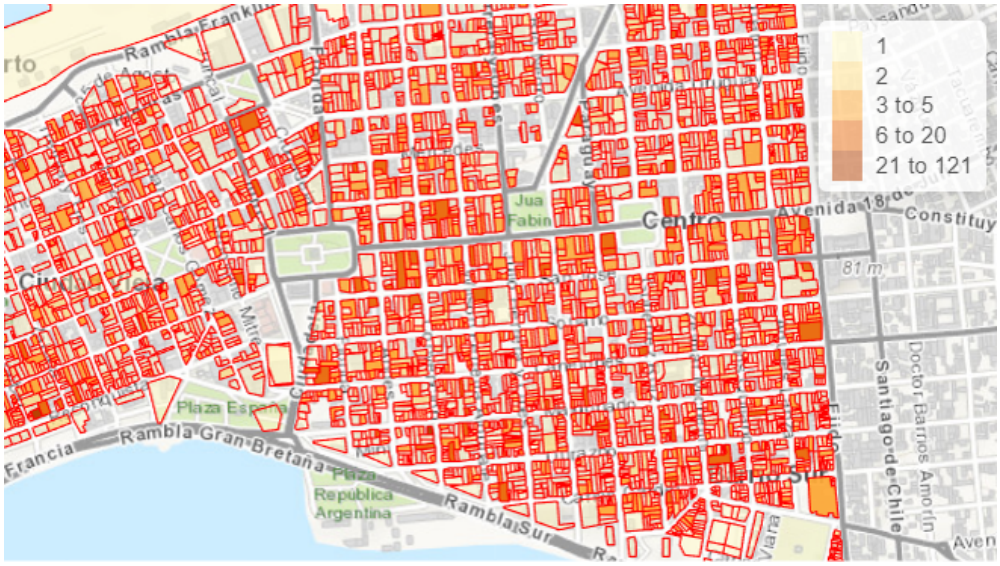

Figura 3.7: Cuentas por padrón del CCZ 1.

Por otro lado, en la figura 3.8 se observa el promedio de bimestres impagos por padrón,

de parte del CCZ 1. Cabe recordar que cada padrón puede incluir 1 o mas TDS, por tal razón se decide reflejar el promedio de bimestres con deuda por padrón.

Figura 3.8: Promedio de bimestres impagos del CCZ 1 por padrón.

A nivel de padrón, y a simple vista, no parece existir una tendencia, un patrón claro que refleje zonas de morosidad concentradas en particulares puntos del mapa.

En cambio, a nivel de CCZ el comportamiento cambia. Si se promedia la variable PROP_BIM_IMP por centro comunal zonal es posible observar una tendencia espacial. Al observar la figura 3.9, se puede apreciar que en la zona norte del departamento suele haber un mayor porcentaje de bimestres impagos, lo cual se traduce en mayor morosidad. Centros comunales vecinos suelen tener un comportamiento similar, la ubicación geográfica aparenta determinar el grado de morosidad que las TDS presentan en promedio a nivel de CCZ.

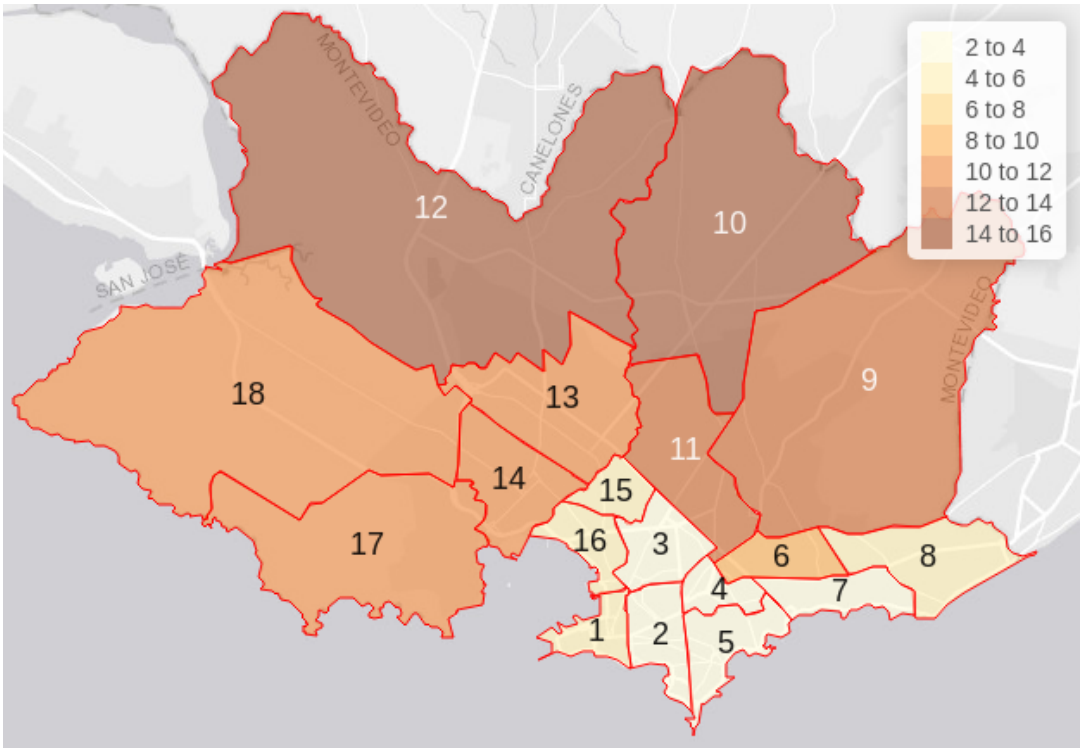

Figura 3.9: Porcentaje de bimestres impagos por CCZ.

Finalmente, si se promedia la variable BIM6, la cual representa el pago del último bimestre, describe la proporción de usuarios que en promedio abonan la TDS. Véase la figura 3.10. Zonas con alta proporción corresponden a zonas de altos ingresos monetarios, en cambio zonas con baja proporción coinciden con zona de menores ingresos monetarios.

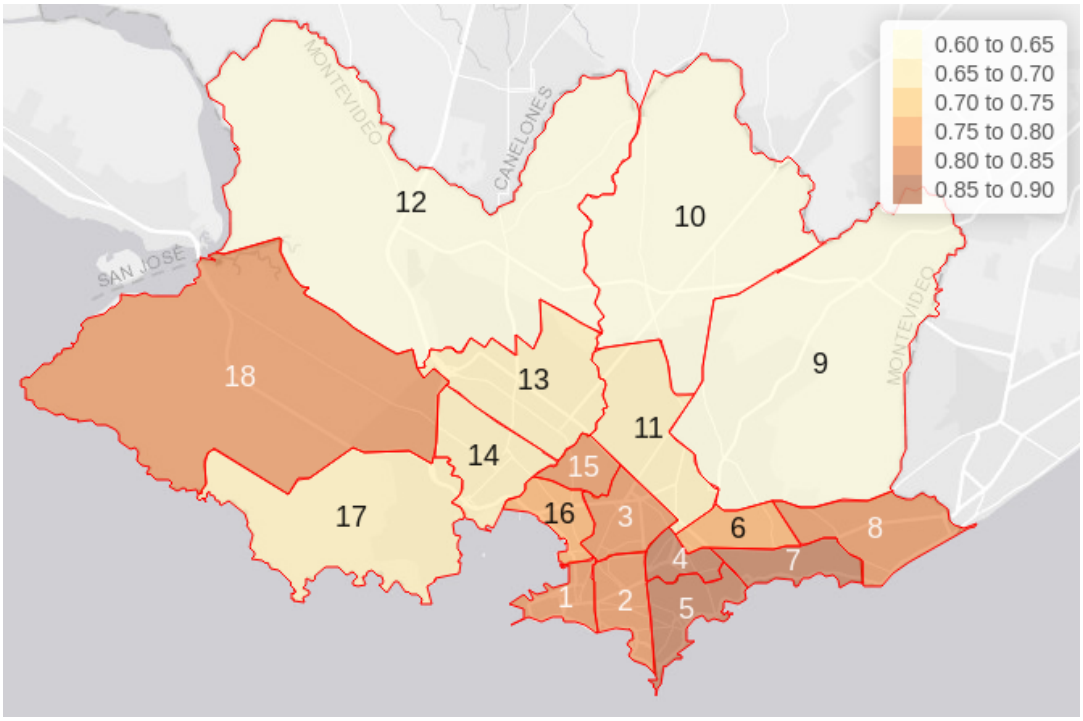

Figura 3.10: Promedio de pagos del último bimestre por CCZ.

Similar a lo que se ve en la figura 3.9, la figura 3.10 también presenta centros que tienden a comportarse de forma similar, centros comunales adyacentes tienden a tener proporciones de pago similares. Existen muchos motivos por lo cuales pueden suceder dichas tendencias, pero existe uno en especial, el cual se cree, y explica la probabilidad de pago en las Tarifas de Montevideo y es el cual apunta este trabajo de investigación, la *dependencia espacial*.

El cuadro 3.1 representa una descripción univariada de la base, sobre los datos tras su procesamiento. Los *CCZ* con mayor cantidad de observaciones son los de los centros 9, 11 y 3. Más del 50 % de las observaciones (Q1 y Q3) presentan una unidad ocupacional (*CANT_UN*) por TDS; existen tarifas con múltiples unidades sin embargo las TDS que predominan son las individuales. La *CATEGORIA* predominante es la categoría DOMICILIARIO. La variable *PROP_BIM_IMPAGOS* presenta en media un valor de 7,97 %, pero en mediana el valor es de 0 %, eso quiere decir que hay algunas observaciones con el porcentaje de bimestres impagos muy alto que provoca un corrimiento en la media hacia valores más elevados. En cuanto a la *ANTIGUEDAD* las TDS tienen en promedio 7,82 años, mientras que en mediana son de 21,59 años, esto se debe a que la mayoría de las TDS fueron dadas de alta en los inicios de la creación del impuesto, por tal razón la mediana es tan alta. Obsérvese la diferencia en la cantidad de observaciones en las categorías de la variable *BIM6*, el 75 % representan los casos de *pago* mientras que el

25 % restante, los casos de *impago*. Por otro lado, cabe destacar la baja cantidad de observaciones con categoría -1 en las variables *BIM1* a *BIM5* cuyo factor es tomado en cuenta en la sección 4.1.

Cuadro 3.1: Estadística descriptiva de los datos procesados.

Variable	Mínimo	Media/Q1	Mediana/Q3	Máximo
CANT_UNIDADES	0[1]	1	1	1134[2]
PROP_BIM_IMPAGOS	0.00	7.76	0.00	100.00
ANTIGUEDAD	0.14	7.82	21.59	21.91

Variable	Categoría	Frecuencia
CCZ	9	30326
	11	29277
	3	25574
	13	23705
	5	20547
	2	20546
	Otros	123256
CATEGORIA	COMERCIAL/INDUSTRIAL	29269
	DOMICILIARIO	243962
BIM6	0 (impago)	68748
	1 (pago)	204483
BIM5	0 (impago)	51812
	1 (pago)	220496
	-1 (sin generar)	923
BIM4	0 (impago)	44824
	1 (pago)	226642
	-1 (sin generar)	1765
BIM3	0 (impago)	43819
	1 (pago)	226797
	-1 (sin generar)	2615
BIM2	0 (impago)	42021
	1 (pago)	227846
	-1 (sin generar)	3364
BIM1	0 (impago)	40155
	1 (pago)	228832
	-1 (sin generar)	4244

[1]Las TDS con 0 unidades son casos particulares que corresponden a TDS en espacios públicos o espacios donde no existen tributos domiciliarios.

[2]Las TDS con gran cantidad de unidades corresponden en su mayoría a complejos habitacionales (CH) caracterizados por gran cantidad de unidades ocupacionales.

4. Resultados

En esta sección se presentan los modelos logísticos construidos con el objetivo de explicar la probabilidad de pago del próximo bimestre **BIM6**, en las TDS . Para evaluar su eficiencia, se construyen conjuntos de entrenamiento y de testeo; con los datos de entrenamiento se construyen los modelos y con los de testeo se evalúa su predicción mediante el uso de matrices de confusión. Adicionalmente, se usa el WAIC para comparar entre modelos que incorporen un enfoque bayesiano.

En esta primera etapa (y a modo de simplificación) se dividen las observaciones por CCZ. De esta forma, se construye un modelo por cada CCZ. Esto permite trabajar con un número de observaciones mas reducida y, principalmente, con observaciones mas homogéneas. Obsérvese la cantidad de observaciones por CCZ en la página 35 figura 3.3. Se dispone de un total de 292.427 observaciones que tras el procesamiento de datos mencionado en la sección 3.2, terminan siendo 292.231.

Las observaciones para los modelos que incluyen un enfoque frecuentista, son divididas en 18 conjuntos de entrenamiento con respecto a cada CCZ; se constituirán el 90 % de las observaciones para los conjuntos de entrenamiento, mientras que el 10 % restante constituye el conjunto de testeo. Para evaluar la matriz de confusión, se estiman múltiples modelos bajo distintos conjuntos donde se calculan el error global, la sensibilidad, la especificidad y la precisión (2.3.1) dentro de cada conjunto de testeo y se promedian.

Se construye una secuencia de modelos con distintos paradigmas, empezando desde un enfoque frecuentista hacia un enfoque bayesiano, buscando ir agregando complejidad a los modelos, desde los mas simples hacia los mas complejos. El motivo por el que se realiza esta transición entre el enfoque frecuentista hacia el bayesiano responde a la conveniencia computacional de este último para incorporar efectos aleatorios con estructura espacial. Los modelos con enfoque bayesiano utilizan todas las observaciones para su estimación, las métricas mencionadas son evaluadas sobre las mismas observaciones sobre las que se construyen los modelos.

4.1. Modelo Logístico I

En la ecuación 4.1 se presenta el modelo que incluye todas las variables para explicar la variable dependiente **BIM6**:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) =& \beta_0 + \beta_1 \text{CANT\_UN}_i + \beta_2 \text{CATEG.DOM}_i + \beta_3 \text{BIM5.-1}_i + \beta_4 \text{BIM5.1}_i + \beta_5 \text{BIM4.-1}_i \\ &+ \beta_6 \text{BIM4.1}_i + \beta_7 \text{BIM3.-1}_i + \beta_8 \text{BIM3.1}_i + \beta_9 \text{BIM2.-1}_i + \beta_{10} \text{BIM2.1}_i \\ &+ \beta_{11} \text{BIM1.-1}_i + \beta_{12} \text{BIM1.1}_i + \beta_{13} \text{ANTIGUEDAD}_i + \beta_{14} \text{PROP\_BIM\_IMP}_i \end{aligned} \tag{4.1}$$

La variable dependiente **BIM6**, representa el pago de la TDS del bimestre 6, cuyos valores son *0* (bimestre impago) y *1* (bimestre pagado), siendo *0* la categoría de referencia. La variable *CATEGOR* representa la categoría de la TDS cuyos valores son *DOM* (domiciliario) o *COM* (comercial), siendo este último el de referencia. Las variables *BIM5*, ... *BIM1* pueden tomar valores *0* (bimestre impago), *1* (bimestre pagado) y *-1* (sin generar deuda). Al igual que para *BIM6*, la categoría de referencia para estas variables es *0*. En las salidas de los modelos, las variables y sus categorías son representadas separadas por un punto, de la siguiente forma `<variable>.<categoría>`, ejemplo *BIM5.-1* refiere a la variable *BIM5* y la categoría *-1*.

De esta forma, la función `glm()` de `R`, se encarga de ajustar el modelo logístico. En la tabla 4.1 se presentan los resultados obtenidos para el modelo 4.1 en los CCZ 01, 05 y 17.

Cuadro 4.1: Resumen Comparativo de los Modelos 4.1 para los CCZ 01, CCZ 05 y CCZ 17.

	CCZ 01				CCZ 05				CCZ 17			
	Estim.	Desvío	Estad.	p-valor	Estim.	Desvío	Estad.	p-valor	Estim.	Desvío	Estad.	p-valor
(Intercept)	-4.39	0.34	-13.02	<0.001	-3.97	0.26	-14.99	<0.001	-3.15	0.33	-9.57	<0.001
CANT_UN	0.04	0.01	4.09	<0.001	0.08	0.01	6.15	<0.001	0.10	0.05	2.03	0.043
CATEG.DOM	0.35	0.11	3.12	0.002	0.33	0.09	3.60	<0.001	-0.25	0.24	-1.04	0.298
BIM5.-1	4.32	0.97	4.45	<0.001	4.09	0.77	5.31	<0.001	3.37	0.67	5.03	<0.001
BIM5.1	5.61	0.32	17.49	<0.001	5.89	0.24	24.20	<0.001	5.39	0.24	22.10	<0.001
BIM4.-1	-1.18	1.28	-0.92	0.358	-0.36	0.93	-0.39	0.699	-0.19	1.03	-0.18	0.857
BIM4.1	-0.21	0.43	-0.49	0.626	-0.92	0.31	-2.97	0.003	-0.33	0.34	-0.96	0.336
BIM3.-1	-0.83	1.15	-0.73	0.468	-1.42	0.90	-1.58	0.115	-0.80	0.90	-0.88	0.377
BIM3.1	-1.34	0.38	-3.50	<0.001	-0.98	0.26	-3.78	<0.001	-1.12	0.32	-3.54	<0.001
BIM2.-1	1.86	1.18	1.58	0.114	0.72	1.24	0.57	0.565	2.59	0.75	3.48	<0.001
BIM2.1	2.60	0.47	5.51	<0.001	2.33	0.33	7.06	<0.001	2.07	0.36	5.78	<0.001
BIM1.-1	0.39	0.97	0.40	0.691	1.06	1.02	1.05	0.296	-1.55	0.66	-2.36	0.019
BIM1.1	-0.31	0.45	-0.70	0.483	-0.07	0.31	-0.21	0.832	-0.00	0.33	-0.01	0.990
ANTIGUEDAD	0.03	0.01	2.41	0.016	0.01	0.01	1.23	0.219	-0.01	0.01	-1.19	0.232
PROP_BIM_IMP	-0.02	0.01	-3.39	<0.001	-0.02	0.01	-5.91	<0.001	-0.02	0.01	-5.50	<0.001

Los errores globales de los modelos en los conjuntos de testeo en estos 3 CCZ, son aproximadamente del 6,6 %, 5,7 % y 5,9 % respectivamente. En promedio, el error global entre los 18 modelos, alcanza el 7,2 % aproximadamente, lo cual destaca una gran efectividad de la predicción.

Por otro lado, la última columna de las tablas anteriores, representa el p-valor obtenido en las pruebas de Wald, las pruebas de significación individual. De esta forma, es inmediato observar que no todas las variables ni todas las categorías de las mismas son significativas para explicar la variable dependiente de los modelos presentes en las tablas anteriores.

A un nivel de significación del 5 % se observa que los siguientes predictores *BIM4.-1* , *BIM4.1* , *BIM3.-1* , *BIM1.-1* , *BIM1.1* no tienen un aporte significativo a los modelos y podrían ser descartados para los 3 modelos. El predictor *BIM2.-1* en el modelo del CCZ 17 y el predictor *ANTIGUEDAD* en el modelo CCZ 01 son significativos, en cambio en los otros 2 modelos, no lo son. Por otro lado, los predictores *Intercept, CATEG.DOM, BIM5.-1 , BIM5.1 , BIM3.1 , BIM2.1 , CANT_UN y PROP_BIM_IMP* son significativos en los 3 modelos.

Las figuras 4.1 y 4.2 representan los resultados obtenidos en los 18 modelos para los coeficientes *BIM5.1* y *BIM1.1* . Se observan 2 gráficos, a la izquierda en la escala log-odds y la derecha en la escala odds-ratio, las estimaciones puntuales junto a sus intervalos de confianza al 95 %. Si bien los resultados al efecto de *BIM5.1* son más claros en todos los modelos, la situación no es la misma en el caso de *BIM1.1*.

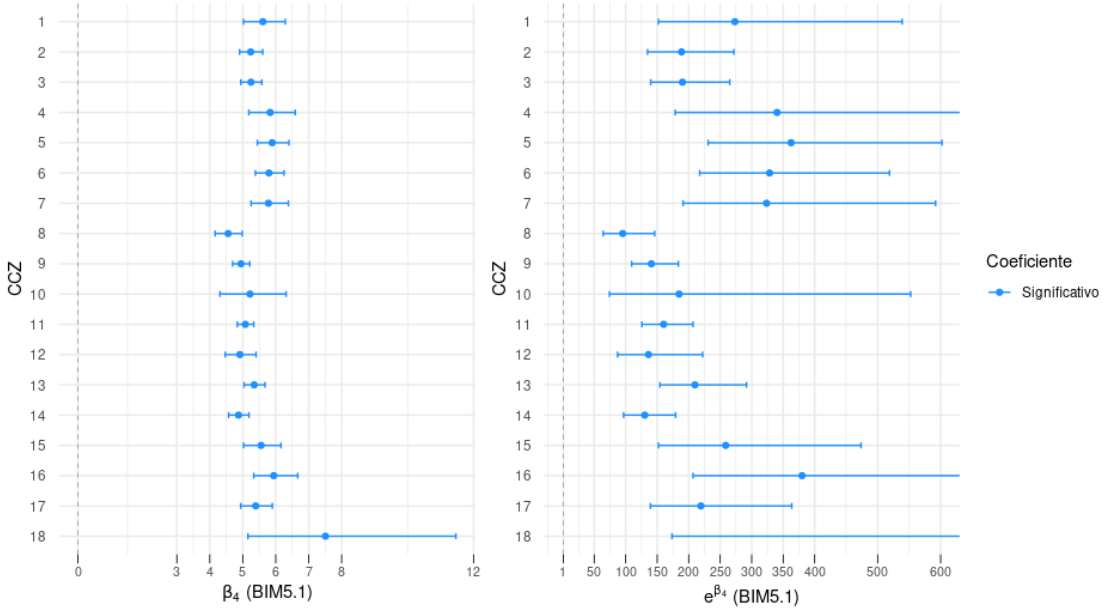

Figura 4.1: Coeficientes estimados asociados a la variable *BIM5.1* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha).

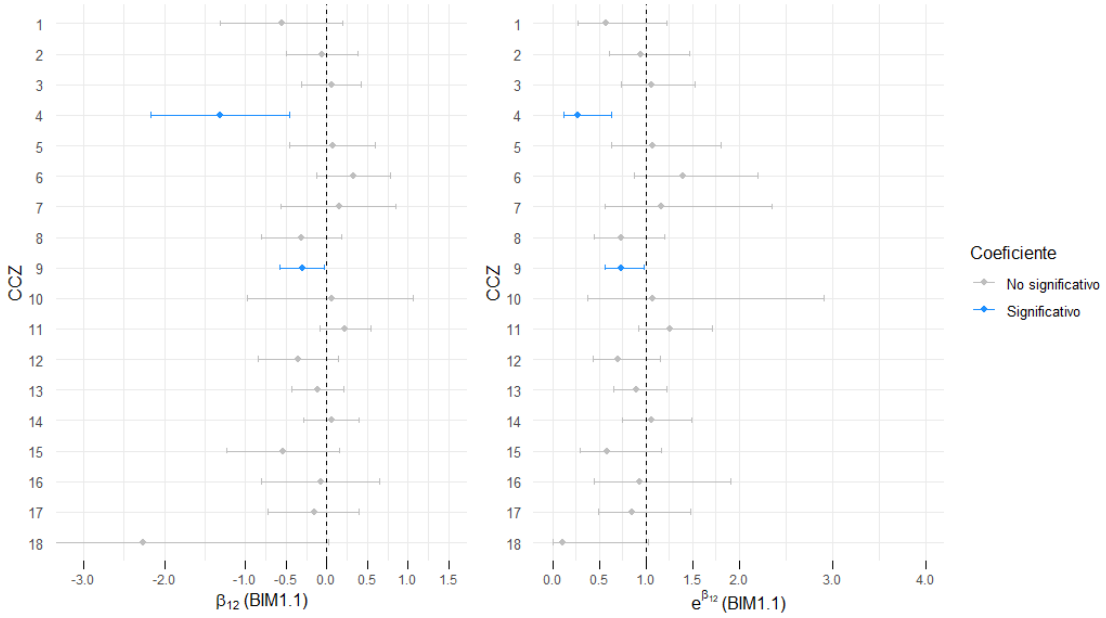

Figura 4.2: Coeficientes estimados asociados a la variable *BIM1.1* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha).

En este contexto, un coeficiente es significativo si su intervalo de confianza no incluye al 0 en la escala log-odds (o de forma equivalente, no incluye al 1, en la escala odds-ratio).

De forma similar ocurre en los predictores restantes. El apéndice A.1 contiene mas resultados, incluyendo los coeficientes *CANT_UN*, *CATEG.DOM* y *PROP_BIM_IMP*.

Siguiendo la misma lógica, se descartan los predictores que no son significativos en la mayoría de los modelos y se retienen aquellos que si, pero si se observa, en los últimos 3 modelos, hay un patrón que se repite, la variable que representa el estado de pago del quinto bimestre tiene un efecto tan notorio, que multiplica en aproximadamente 200 veces la chance de ocurrencia de *BIM6*. De por sí sola, prácticamente determina la ocurrencia o no de la variable de respuesta.

Esto quiere decir que si la información de la variable *BIM5* es igual a 1 (bimestre anterior pagado), la chance en promedio entre modelos, aumenta alrededor de 20.000 % dejando el resto de variables constantes, obsérvese la escala odds-ratios de la figura 4.1.

Los efectos parciales de un modelo logístico representan el cambio en el logaritmo de las odds (log-odds) para un cambio unitario de la variable predictora. Para el modelo del CCZ 01 en 4.1, el coeficiente de *BIM5.1* es 5.61, esto significa que el efecto parcial del logaritmo de las odds de *BIM6*, dejando el resto de variables constante, es 5.61 unidades más alto cuando *BIM5* es 1 en comparación a la categoría de referencia (0, deuda impaga).

En términos de odds-ratios (OR), el OR para *BIM5.1* es:

$$OR_{BIM5.1} = e^{5.61} \approx 272.96$$

Esto indica que las odds de pagar en el bimestre 6, son aproximadamente 273 veces más altas cuando el bimestre 5 ha sido pagado en comparación a cuando no ha sido pagado, dejando el resto de variables constantes, en el CCZ 01.

De forma análoga, la probabilidad de que el evento de interés ocurra, dado que *BIM5* es 1, manteniendo todas las demás variables constantes para el modelo del CCZ 01, es:

$$Prob = \frac{272.96}{1 + 272.96} = 0.9963498 \approx 1$$

El modelo es tan bueno, que solo incluyendo la variable explicativa *BIM5* sería posible predecir la variable de respuesta con 92,6 % de efectividad (tabla 4.2).

A raíz de esta observación se decide realizar una tabla de doble entrada donde se cruzan los valores de las variables *BIM6* y *BIM5* filtrando las TDS con *BIM5* igual a -1 (sin generar deuda). Se obtiene una matriz que permite observar la cantidad de observaciones

que pagan el *BIM5* y también pagan el *BIM6* como también aquellas que pagan el *BIM5* pero no el *BIM6*, obsérvese el cuadro 4.2 a continuación:

Cuadro 4.2: Tabla de doble entrada de *BIM6* respecto a *BIM5*.

		BIM5	
		Impago	Pago
BIM6	Impago	49.972 (96,4 %)	18.333 (8,3 %)
	Pago	1.840 (3,5 %)	202.163 (91,7 %)

El cuadro 4.2, refleja las frecuencias absolutas y las frecuencias relativas del cruce de las variables *BIM6* y *BIM5* (estos bimestres corresponden al período Agosto-Setiembre y Octubre-Noviembre 2022). Obsérvese como los casos donde BIM6 y BIM5 no coinciden, impago-pago y viceversa, son pocos en proporción, 8,3 % y 3,5 % respectivamente. A partir de esto, se puede destacar la alta asociación entre las variables *BIM5* y *BIM6*.

Esta relación se sostiene si se considera el historial de cada usuario, es decir, si se filtran todas las TDS que no pagaron en los últimos 5 bimestres para observar su comportamiento de pago (refiriéndose a las variables *BIM1* a *BIM5* inclusive con valor 0), se obtienen 31.780 observaciones de las cuáles el 98,1 % tampoco paga el bimestre próximo (*BIM6*). Por otro lado, de las TDS que sí pagaron en los últimos 5 bimestres (es decir variable *BIM1* a *BIM5* con valores 1), de esas 210.461 observaciones, el 91,7 % continúa pagando.

De forma análoga, y teniendo en cuenta únicamente la información de pago del bimestre anterior, se puede predecir con una tasa de error de solo un 7,38 %, el pago o no del próximo bimestre, dado que el 96,4 % de los no pagadores, no pagan el próximo bimestre, y dado que el 91,7 % de los pagadores, si pagan el bimestre próximo.

Por lo tanto, encontrar un modelo que supere estos hechos, tendría escasa utilidad práctica. Por este motivo se decide predecir la probabilidad de pago sobre aquellas observaciones en las que **no** se tiene historial de pagos realizados del último bimestre o anterior. Estas observaciones son las que contienen valor -1 en las variables *BIM1* a *BIM5*. La cantidad resultante son 923 observaciones, una reducción significativa respecto a la cantidad inicial, pero estas observaciones seleccionadas ofrecen una oportunidad para obtener mayores ganancias en términos predictivos.

Cabe aclarar, que de estas 923 TDS, 702 no cuentan con historial de pago en ninguno

de los 5 últimos bimestres, es decir, que se trata de cuentas que llevan tiempo inactivas y este bimestre próximo, dejan de estarlo, y vuelven a estar activas y generar deuda. La nueva pregunta de investigación apunta a estas TDS, ¿que probabilidad de pago tienen las TDS que estaban cortadas y vuelven a generar deuda? .

De esta forma, las 923 TDS son obtenidas a partir de filtrar las observaciones con $BIM5 = -1$, las variables *BIM1* a *BIM5* no son incluidas en los modelos, dado que fueron filtradas. A modo de comentario, de no incluirse *BIM5*, e incluirse *BIM4* a *BIM1*, la variable que cobra mas significancia para predecir pasa a ser *BIM4*, lo mismo sucede si se opta por prescindir de *BIM4* y probar con el resto, por dicha razón se deciden sacar las variables de información de pago.

4.2. Modelo Logístico II

El nuevo conjunto de observaciones presenta las siguientes estadísticas descriptivas:

Cuadro 4.3: Estadística descriptiva de los datos.

Variable	Mínimo	Media/Q1	Mediana/Q2	Máximo
CANT_UNIDADES	1	1	1	32
PROP_BIM_IMPAGOS	0.00	19.60	0.38	96.70
ANTIGUEDAD	0.14	8.98	2.56	21.64

Variable	Categoría	Frecuencia
CCZ	9	199
	11	105
	17	87
	3	73
	14	68
	2	57
	Otros	334
CATEGORIA	COMERCIAL/INDUSTRIAL	145
	DOMICILIARIO	778
BIM6	0 (impago)	443
	1 (pago)	480

A diferencia de los datos iniciales, la cantidad de observaciones con éxito (pago) y fracaso

(no pago) en la variable de respuesta tras el filtrado es mas equitativa (véase el cuadro 3.1). La *PROP_BIM_IMPAGOS* es en media 19,6 %, mientras que en mediana es 0,38 %, esta diferencia se debe a la misma razón que se explica en el conjunto de datos anterior y se debe a que presenta mucha asimetría. Con la *ANTIGÜEDAD* sucede algo similar, la media de la variable ronda en 8,98 años mientras que la mediana es de 2,56 años, existen algunas cuentas que son muy antiguas lo cual eleva el valor de la media.

Continuando con este enfoque, se construye el modelo logístico incluyendo todas las variables disponibles. En esta ocasión no se genera un modelo por cada CCZ, sino que se opta por incluirla en el modelo:

$$\begin{aligned}\log\left(\frac{p_i}{1-p_i}\right) =& \beta_0 + \beta_1 \text{CCZ.1}_i + \cdots + \beta_{18} \text{CCZ.18}_i \\ &+ \beta_{19} \text{CANT\_UN}_i + \beta_{20} \text{CATEG.DOM}_i + \beta_{21} \text{ANTIGUEDAD}_i \\ &+ \beta_{22} \text{PROP\_BIM\_IMP}_i\end{aligned} \tag{4.2}$$

El cuadro 4.4 presenta los resultados del modelo 4.2. Las categorías de referencia son *CCZ.1*, para la variable *CCZ* y *CATEG.COM* para la variable *CATEGORIA*.

Cuadro 4.4: Resumen del Modelo 4.2.

	Estimación	Desvío	Estadístico	p-valor
(Intercept)	3.04	0.56	5.40	<0.001
CCZ.2	0.56	0.61	0.92	0.360
CCZ.3	0.31	0.57	0.55	0.586
CCZ.4	0.94	0.88	1.07	0.285
CCZ.5	0.30	0.81	0.37	0.709
CCZ.6	-0.90	0.67	-1.34	0.181
CCZ.7	0.39	0.74	0.52	0.600
CCZ.8	1.20	1.24	0.97	0.334
CCZ.9	-0.88	0.54	-1.62	0.106
CCZ.10	0.12	0.80	0.15	0.878
CCZ.11	-0.99	0.57	-1.74	0.083
CCZ.12	-0.14	0.80	-0.18	0.857
CCZ.13	-1.52	0.68	-2.22	0.026
CCZ.14	-0.27	0.65	-0.41	0.681
CCZ.15	-1.50	0.89	-1.68	0.093
CCZ.16	0.31	0.91	0.34	0.735
CCZ.17	-0.69	0.65	-1.06	0.288
CCZ.18	1.61	5.09	0.32	0.752
CANT_UN	0.08	0.15	0.55	0.584
CATEG.DOM	-0.11	0.32	-0.35	0.724
ANTIGUEDAD	-0.12	0.02	-7.56	<0.001
PROP_BIM_IMP	-0.11	0.01	-10.2	<0.001

El error global del modelo, en los conjuntos de testeo es en promedio del 19,09 %, la sensibilidad de 77,95 %, la especificidad es de 85,89 % y la precisión alcanza un 88,86 %.

El error del modelo aumenta considerablemente en comparación al modelo 4.1 debido a que no considera las variables de los bimestres anteriores, sin embargo el potencial de predicción sigue siendo alto.

En cuanto al aporte de cada variable explicativa, considerando un nivel de significación del 5 %, los p-valores obtenidos en las pruebas de signficación individual, muestran que los predictores *CANT_UN* y *CATEG* no son significativos, mientras que *Intercept*, *ANTIGUEDAD* y *PROP_BI* si lo son. El caso de *CCZ* no es claro en tanto que en algunos casos el coeficiente de algún CCZ es significativo pero otros no.

Se realiza una prueba de significación conjunta para las categorías de la variable *CCZ*,

véanse los resultados en el cuadro 4.5:

Cuadro 4.5: Test de Significación Conjunta del Modelo 4.2

	LR Chisq	Df	Pr(>Chisq)
CCZ	25.01	17	0.094

Manteniendo el mismo nivel de significación, la variable CCZ no es significativa para el modelo, por lo cual se decide descartarla, así como la variable *CANT_UN* y *CATEG* cuyo p-valores en el cuadro 4.4 son 0,584 y 0,724 respectivamente, siendo no significativas.

En términos prácticos, descartar las variables anteriores lleva a las siguientes conclusiones. Para las TDS que tienen deuda sin generar en el bimestre anterior ($BIM5 = -1$), el modelo resultante, que predice la probabilidad de pago en el próximo bimestre, no incluye ni *CANT_UN* ni *CATEG* como variables explicativas. Esto implica que, independientemente de si la tarifa es comercial o domiciliaria, o si corresponde a una unidad o a varias, estas variables no influyen significativamente en la probabilidad de pago, resultando irrelevantes en el modelo.

4.3. Modelo Logístico III

El modelo resultante de la etapa anterior solo incluye las variables explicativas *ANTIGUEDAD* y *PROP_BIM_IMPAGOS*:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{ANTIGUEDAD}_i + \beta_2 \text{PROP\_BIM\_IMP}_i \tag{4.3}$$

El cuadro 4.6 representa la salida de resúmen del modelo 4.3:

Cuadro 4.6: Resúmen del Modelo 4.3.

	Estimación	Desvío	Estadístico	p-valor
(Intercept)	2.24	0.20	11.34	<0.001
ANTIGUEDAD	-0.09	0.01	-7.85	<0.001
PROP_BIM_IMP	-0.11	0.01	-10.66	<0.001

En este modelo, el error global promedia en el entorno, del 18,28 %, la sensibilidad 75,06 %, la especificidad 95,39 % y la precisión alcanza un 97,10 %. En comparación al modelo 4.2, el modelo propuesto, presenta una mejora en la predicción de la variable de respuesta, minimizando en apenas un punto porcentual el error del modelo.

La sensibilidad se mantiene menor al 80 % en ambos modelos, dando lugar a una mayor dificultad para identificar correctamente los casos positivos entre los casos verdaderamente positivos.

Por otro lado, la especificidad es mas alta en el modelo 4.3 que en el 4.2, reflejando una mejora en la identificación de los casos negativos entre los casos verdaderamente negativos.

Finalmente, la precisión, tiene una mejora sustancial pasando de 88,86 % a 97,10 %, lo cual implica que el modelo propuesto clasifica mejor la predicción de los casos positivos que son correctamente casos positivos.

Los coeficientes obtenidos son coherentes, para la variable *ANTIGUEDAD*, a la que le corresponde un odds-ratio de 0,913 indicando que, por cada año adicional, los odds de la variable dependiente disminuye en un factor de 0,913 , es decir disminuye un 8,7 %. Cuanto mas antigua la TDS, menor es la probabilidad de pago. Vale la pena recordar que se esta trabajando sobre un conjunto de 923 TDS las cuáles en media presentan 8,98 años de antigüedad, mientras que en mediana presentan 2,56 años. Esto indica que existen cuentas muy antiguas que provocan que la media de la variable aumente, pero en mayoría las cuentas suelen ser mas recientes (menores a tres años).

4.4. Modelo Logístico IV

Este modelo, introduce el enfoque bayesiano del modelo 4.3 a efectos de ser una base de comparación con los modelos siguientes. La matriz de confusión continúa siendo el indicador de desempeño primario, y adicionalmente, se introduce el WAIC (véase capítulo 2.3.2). De esta forma se construye y se estima el modelo 4.4 con las distribuciones previas indicadas en 4.5:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \text{ANTIGUEDAD}_i + \beta_2 \text{PROP\_BIM\_IMP}_i \tag{4.4}$$

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(0, 10^2) \\ \beta_1 &\sim \mathcal{N}(0, 10^2) \\ \beta_2 &\sim \mathcal{N}(0, 10^2) \end{aligned} \tag{4.5}$$

El cuadro 4.7 refleja el resultado de las estimaciones del modelo 4.4:

Cuadro 4.7: Resumen del Modelo 4.4.

	Estimación	Desvío	IP inf. 95 %	IP sup. 95 %	Rhat	Bulk ESS
Intercept	2.267	0.190	1.912	2.650	1.002	2410
ANTIGUEDAD	-0.087	0.010	-0.107	-0.066	1.001	2050
PROP_BI	-0.111	0.011	-0.134	-0.092	1.001	1482

Los resultados obtenidos son prácticamente iguales a los obtenidos en el modelo desde un enfoque frecuentista (cuadro 4.6). El $\hat{R}$ y el *Bulk ESS*, son indicadores de diagnóstico de convergencia y eficiencia del muestreo, los cuáles, en este caso no dan indicios de problema en el muestreo de la distribución posterior. El apéndice A.1 contiene los traceplots y posterior plots que forman parte del diagnóstico del modelo 4.4.

El error global alcanza el 18,31 %, la sensibilidad 97,08 %, la especificidad 65,01 % y la precisión 75,04 %. En términos de error del modelo, en comparación al modelo 4.3 el error es el mismo, en cuanto a la sensibilidad el modelo actual tiene una mejora sustancial y una desmejora en la especificidad y la precisión, en conclusión el modelo actual mejora la clasificación de los casos verdaderos positivos, pero empeora en la correcta clasificación de los verdaderos negativos y en la precisión general de las predicciones clasificadas positivas. El *WAIC* del modelo es de 719,9 y un desvío de 37,4 .

4.5. Estructura espacial

Crear pesos espaciales es un paso necesario al usar datos con estructura espacial, nos permiten verificar que no queda ningún patrón espacial en los residuos del modelo 4.4.

De esta forma, es necesario definir la estructura de vecindad y las ponderación espaciales para crear la matriz de pesos espaciales. En primer lugar se define una relación de vecindad entre las observaciones (TDS), y de esta forma la asignación de pesos no nulos a cada unidad de área que le corresponda. Para ello se debe seleccionar un criterio y luego asignar los pesos a los vecinos según dicho criterio.

A partir de las observaciones a nivel individual, se forman grupos a nivel agregado de CCZ las cual conforman las unidades de área. La relación entre CCZ se define según contigüidad, es decir, un vecino se determina si comparte un borde o no, formando una matriz de vecinos binaria (`B`). Esta matriz se caracteriza por ser simétrica con dimensión 18×18, con ceros en su diagonal principal y unos si un CCZ comparte un borde con otro

CCZ. La figura 4.3, ilustra la estructura de vecindad de las unidades de área utilizadas.

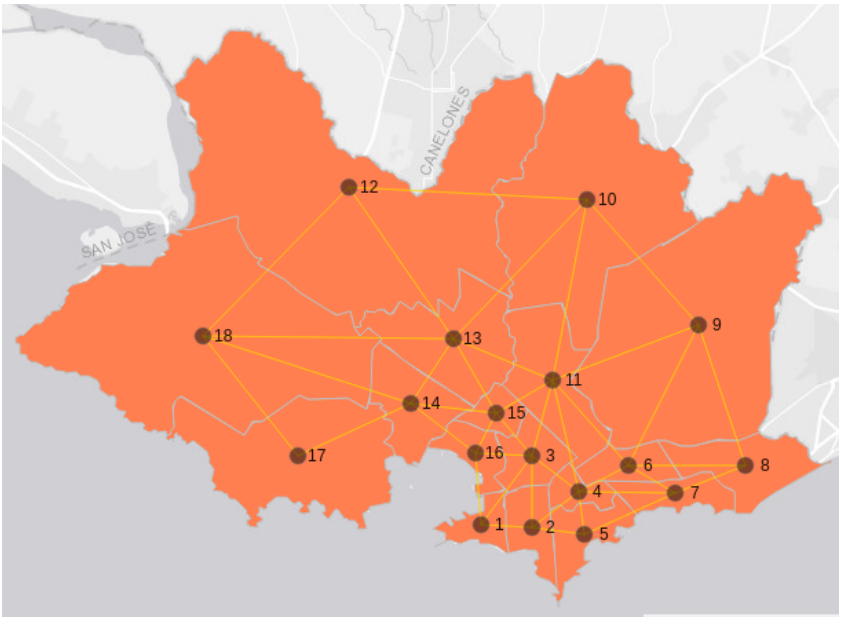

Figura 4.3: Mapa de Montevideo de los CCZ y su estructura de vecindad según el criterio de contigüidad.

Una vez que se establece la lista de conjuntos de vecinos para la región de estudio, procedemos a asignar pesos espaciales a cada relación. La matriz de vecinos con unos y ceros es transformada asignando pesos para cada unidad, de forma que sumen uno por fila; esto también se conoce como estandarización por filas (`W`). Los pesos varían entre uno dividido por el mayor y el menor número de vecinos, y las sumas de los pesos para cada unidad es igual a uno por fila. Los pesos para los enlaces que se originan en áreas con pocos vecinos son mayores que los que se originan en áreas con muchos vecinos.

De acuerdo a las estructuras de vecinos definidas en las dos matrices anteriores, podemos comenzar a utilizarlos para explorar la presencia de autocorrelación espacial.

A partir del modelo resultante (4.4), se extraen los residuos de devianza y junto con las matrices de pesos espaciales de estilo binaria y estandarizada por fila (`B` y `W`) se realiza el test de Moran (capítulo 2.4.4). El Índice de Moran permite evaluar la presencia de autocorrelación en la variable de interés, dada la estructuras de vecindad presente en la figura 4.3. Vale la pena aclarar que a efectos de emplear este indicador fue necesario agregar los residuos a nivel de CCZ. En este estudio se optó por utilizar la media y la mediana de los mismos en cada CCZ. Recordemos que el test de Moran consiste en una prueba de hipótesis, donde la hipótesis nula considera la ausencia de autocorrelación

espacial, y en la alternativa la existencia, por lo tanto, un rechazo de la hipótesis nula frente a un $p - valor$ menor a cierto nivel de significación implica presencia de autocorrelación espacial en los residuos del modelo, lo cual refleja que no se estaría captando en su totalidad la estructura espacial en la variable de respuesta.

Intentar detectar patrones en los mapas de residuos de forma visual no es una opción totalmente fiable, obsérvese las figuras 4.4 y 4.5, la presencia o ausencia de un patrón espacial presente en los residuos del modelo, no es tan evidente visualmente para la media de los residuos pero si tal vez para la mediana. Estas figuras representan las medias y las medianas de los residuos de devianza del modelo 4.4 a nivel agregado por CCZ.

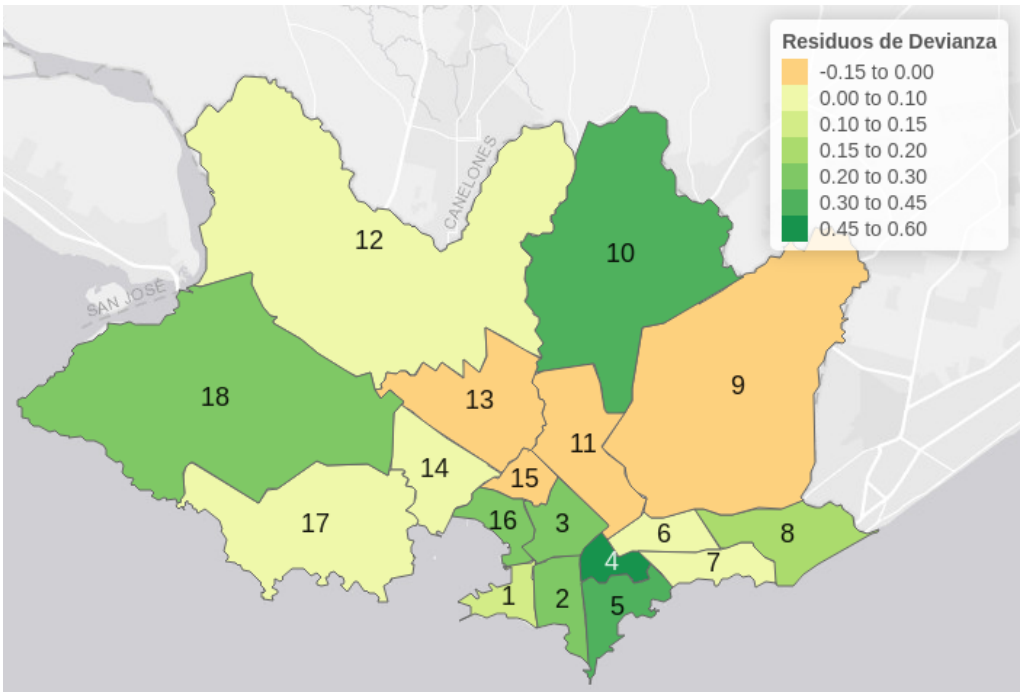

Figura 4.4: Mapa de Montevideo según la *media* de los residuos de devianza del modelo 4.4 por CCZ.

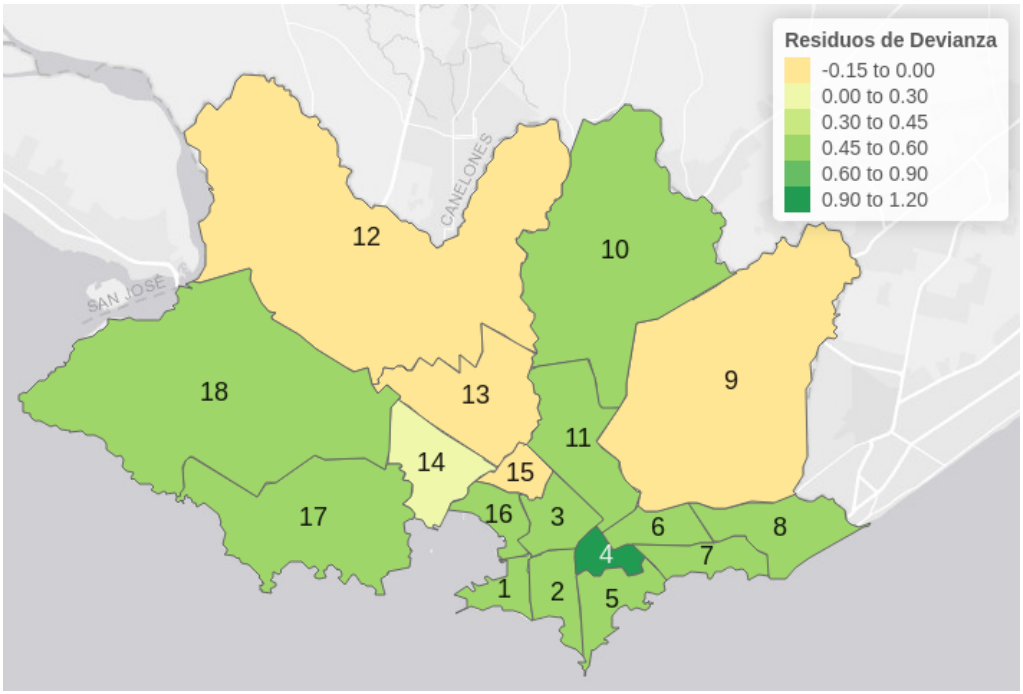

Figura 4.5: Mapa de Montevideo según la *mediana* de los residuos de devianza del modelo 4.4 por CCZ

Bajo esta estructura de vecindad se realiza el test de Moran para evaluar la autocorrelación espacial sobre la media y la mediana de los residuos de devianza del modelo utilizando una estructura de vecindad a nivel de CCZ, y los estilos `B` y `W` para las ponderaciones de las matrices. La tabla 4.8 refleja los resultados de los tests.

Cuadro 4.8: Tests de Moran para la media y la mediana de los residuos de devianza del modelo 4.4 a nivel agregado de CCZ.

Media					*Mediana*				
Style	I	$\mathbb{E}(I)$	$\mathbb{V}(I)$	$p-valor$	*Style*	I	$\mathbb{E}(I)$	$\mathbb{V}(I)$	$p-valor$
`B`	0.0518	-0.0588	0.0180	0.205	`B`	0.173	-0.0588	0.0173	0.0385
`W`	0.0360	-0.0588	0.0202	0.252	`W`	0.0354	-0.0588	0.0193	0.0694

Para el estilo `B`, el valor del índice de Moran (I) para la media es 0.0518, con una expectativa $\mathbb{E}(I)$ de -0.0588 y una varianza $\mathbb{V}(I)$ de 0.0180, resultando en un p-valor de 0.205, lo que indica que no hay evidencia significativa de autocorrelación espacial. En contraste. En el caso de la mediana es 0.173, con la misma media y una varianza de 0.0173, resultando en un p-valor de 0.0385, sugiriendo que existe evidencia significativa de autocorrelación espacial en la mediana de los residuos.

Para el estilo `W`, el valor del índice de Moran para la media es 0.0360, con una media de -0.0588 y una varianza de 0.0202, resultando en un p-valor de 0.252, lo que nuevamente indica que no hay evidencia significativa de autocorrelación espacial. Sin embargo, para la mediana, el índice de Moran es 0.0354, con una varianza de 0.0193 y un p-valor de 0.0694, lo que sugiere cierta evidencia de autocorrelación espacial, aunque no significativa al nivel del 5 %.

En resumen, mientras que los resultados del test de Moran para la media de los residuos no indican autocorrelación espacial significativa para ambos estilos, los resultados para la mediana de los residuos sí muestran evidencia de autocorrelación espacial significativa para el estilo `B` y una ligera evidencia para el estilo `W`.

De forma complementaria se realiza el test de Moran sobre el modelo 4.3, con el objetivo de observar si existen diferencias en la presencia de la autocorrelación espacial sobre la media y la mediana de los residuos de devianza entre modelos desde enfoques bayesiano o frecuentista; los resultados presentes en el cuadro 4.9 reflejan resultados similares que para el modelo 4.3 en el cuadro 4.8. Las conclusiones son las mismas, para la media de los residuos no hay indicio de autocorrelación espacial significativa para ambos estilos, mientras que para la mediana de los residuos hay evidencia de autocorrelación espacial significativa para el estilo `B` y una ligera evidencia para el estilo `W`.

Cuadro 4.9: Tests de Moran para la media y la mediana de los residuos de devianza del modelo 4.3 a nivel agregado de CCZ.

Media					*Mediana*				
Style	I	$\mathbb{E}(I)$	$\mathbb{V}(I)$	$p-valor$	*Style*	I	$\mathbb{E}(I)$	$\mathbb{V}(I)$	$p-valor$
`B`	0.0499	-0.0588	0.0180	0.209	`B`	0.176	-0.0588	0.0172	0.0368
`W`	0.0342	-0.0588	0.0202	0.257	`W`	0.149	-0.0588	0.0193	0.0676

Según el índice de Moran, se verifica la presencia de autocorrelación espacial en los residuos del modelo, por lo que se opta por incluir un componente que capture dicha variabilidad en el modelo. En el capítulo 2.5.1 se introduce una clase de modelos mixtos llamados autorregresivos condicionales que incorporan un componente aleatorio que varía según el espacio. Esta clase de modelos asume que la variable dependiente esta condicionalmente correlacionado a los valores de la misma variable en ubicaciones vecinas. Este supuesto coincide con los resultados del modelo 4.4, en donde se observa presencia de autocorrelación espacial presente en los residuos de devianza. El test de Moran sobre la media de la variable dependiente *BIM6* por CCZ para las mismas observaciones del

modelo 4.4, sugiere una ligera evidencia de autocorrelación espacial para los estilos `B` y `W`. Véase el cuadro 4.10.

Cuadro 4.10: Tests de Moran para la media de *BIM6* por CCZ.

Style	I	$\mathbb{E}(I)$	$\mathbb{V}(I)$	$p - valor$
`B`	0.155	-0.0588	0.0182	0.0564
`W`	0.135	-0.0588	0.0205	0.0876

La figura 4.6 representa la media de la variable *BIM6* por CCZ. Unidades vecinas en los CCZ al sur del departamento poseen un comportamiento similar con altos promedios de pago, mientras que los CCZ al norte de Montevideo presentan el comportamiento contrario que aparenta extenderse en unidades vecinas, este comportamiento puede ser visto como un patrón espacial.

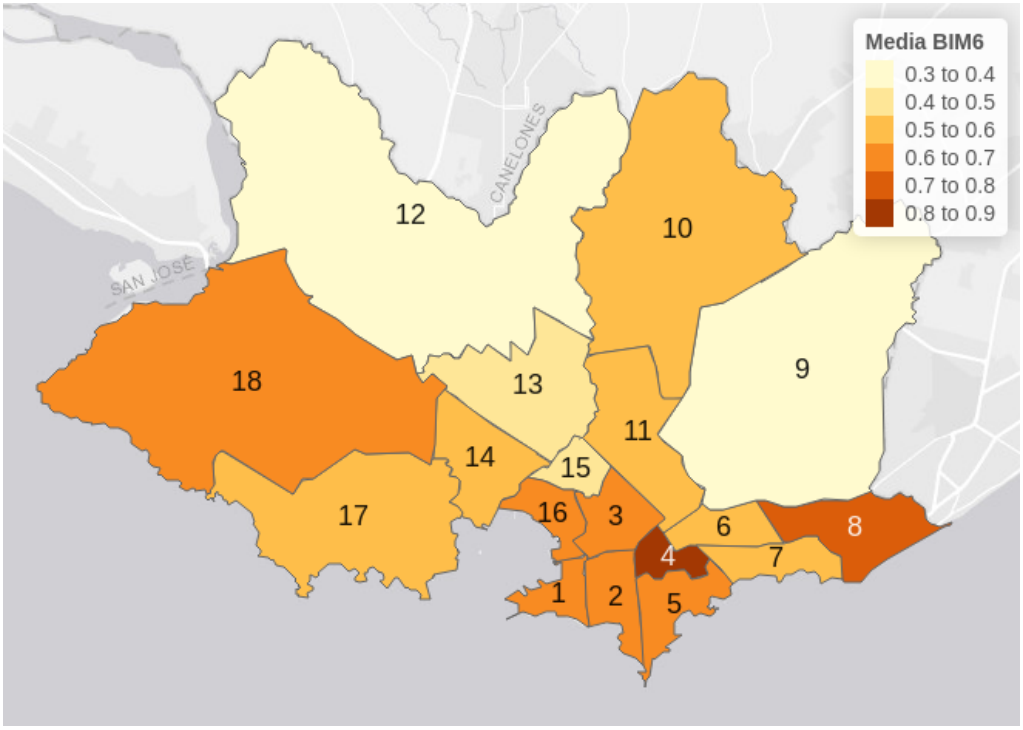

Figura 4.6: Mapa de Montevideo de los CCZ según la media de la variable dependiente *BIM6*.

4.6. Modelo Logístico - CAR

En un contexto de datos jerárquicos, los modelos CAR introducen efectos aleatorios con el objetivo de capturar la estructura espacial presente. Esta clase de modelos pertenece a la familia de modelos mixtos. En el capítulo previo, se detectó mediante el uso del índice de Moran, la presencia de autocorrelación. Esto motiva el uso de modelos que capturen dicha variabilidad espacial, incorporando así el modelo CAR para la distribución de efectos aleatorios a nivel de CCZ.

Bajo un enfoque bayesiano, se especifica y se estima el modelo 4.6, dadas las distribuciones previas seleccionadas en 4.7:

$$\begin{aligned} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_0 + \beta_1 \text{ANTIGUEDAD}_{ij} + \beta_2 \text{PROP\_BIM\_IMP}_{ij} + \phi_i \\ & i = 1, \ldots, M. \quad j = 1, \ldots, n. \end{aligned} \tag{4.6}$$

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(0, 10^2) \\ \beta_1 &\sim \mathcal{N}(0, 10^2) \\ \beta_2 &\sim \mathcal{N}(0, 10^2) \\ \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathbf{0}, \left(\frac{1}{\sigma}\boldsymbol{D}(\mathbb{I} - \alpha\boldsymbol{B})\right)^{-1}\right) \\ \alpha &\sim U(0,1) \\ \sigma &\sim t_3(0, 2.5) \cdot \mathbb{I}(\sigma > 0) \end{aligned} \tag{4.7}$$

donde p_{ij} es la probabilidad de éxito para la observación j en el grupo i, los grupos son compuestos por los 18 CCZ, β_0 representa una constante entre los i CCZ, $ANTIGUEDAD_{ij}$ y $PROP\_BIM\_IMP_{ij}$ son las variables explicativas asociadas a la observación j en el CCZ i, β_1 y β_2 son los coeficientes asociados a dichas variables explicativas y ϕ_i es el vector aleatorio espacialmente correlacionado que captura la variabilidad espacial para el conjunto de CCZ.

El cuadro 4.11 presenta los resultados para el modelo 4.6:

Con respecto al diagnóstico del modelo (véase el apéndice A.1 para ver los trace y posterior plots), los resultados de los coeficientes son consistentes y no presentan problemas de convergencia en general, sin embargo la estimación del coeficiente α_{CAR} no es muy eficiente, se sabe que α_{CAR} controla el grado de correlación espacial, tomando valores entre $[0, 1]$, dado que su intervalo de probabilidad al 95 % lo sitúa en el intervalo $[0.042, 0.986]$, se concluye que la precisión al estimar este parámetro es muy baja. La figura 4.7 repre-

Cuadro 4.11: Resumen del Modelo logístico CAR.

	Estimación	Desvío	IP inf. 95 %	IP sup. 95 %	Rhat	Bulk ESS
Intercept	2.467	0.374	1.849	3.207	1.003	1084
ANTIGUEDAD	-0.095	0.012	-0.120	-0.073	1.000	3711
PROP_BI	-0.111	0.011	-0.135	-0.092	1.001	5090
α_{CAR}	0.593	0.284	0.042	0.986	1.002	1665
σ_{CAR}	0.790	0.346	0.195	1.573	1.003	751

senta la distribución posterior del coeficiente α_{CAR}.

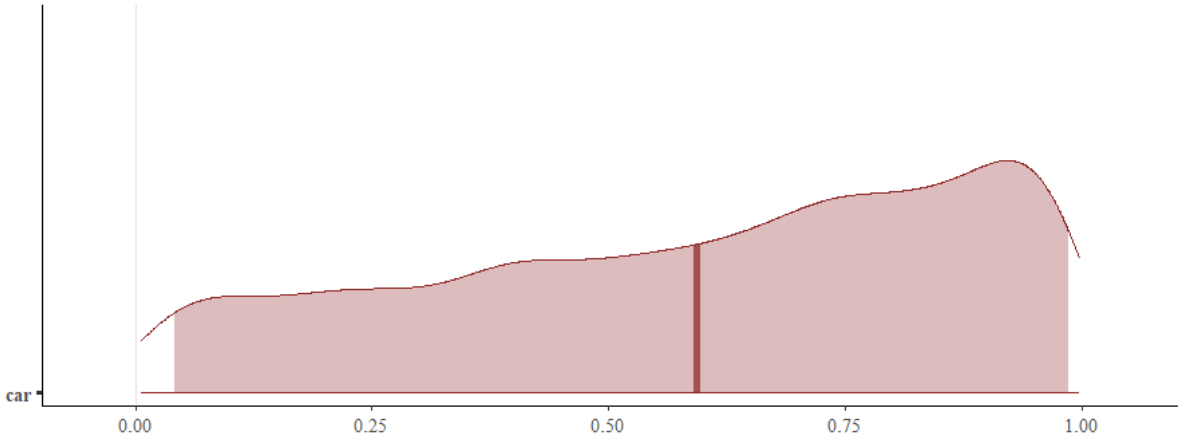

Figura 4.7: Posterior plot del coeficiente α del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad.

La estimación recorre todo el recorrido de la previa uniforme del coeficiente sin identificar un valor con la suficiente precisión tal y como se anticipó en la lectura del intervalo de probabilidad al 95 %.

Por otro lado, el error global del modelo es de 15,82 %, la sensibilidad es de un 96,46 %, la especificidad de un 70,88 % y la precisión de un 78,21 %. Si se compara con el modelo anterior (modelo 4.4), el modelo CAR mejora en casi todos los aspectos, mejorando en 2,49 puntos el error global, disminuyendo en 0,62 puntos la sensibilidad y aumentando en 5,87 y 3,17 puntos la especificidad y la precisión respectivamente. El modelo CAR mejora la predicción de las observaciones mediante la incorporación de los efectos aleatorios espacialmente correlacionados $\mathbf{\Phi}$, captando la variabilidad presente en los datos. La estimación de estos efectos aleatorios se ven reflejados en la figura 4.8, donde se ve el valor de ϕ_i para cada CCZ.

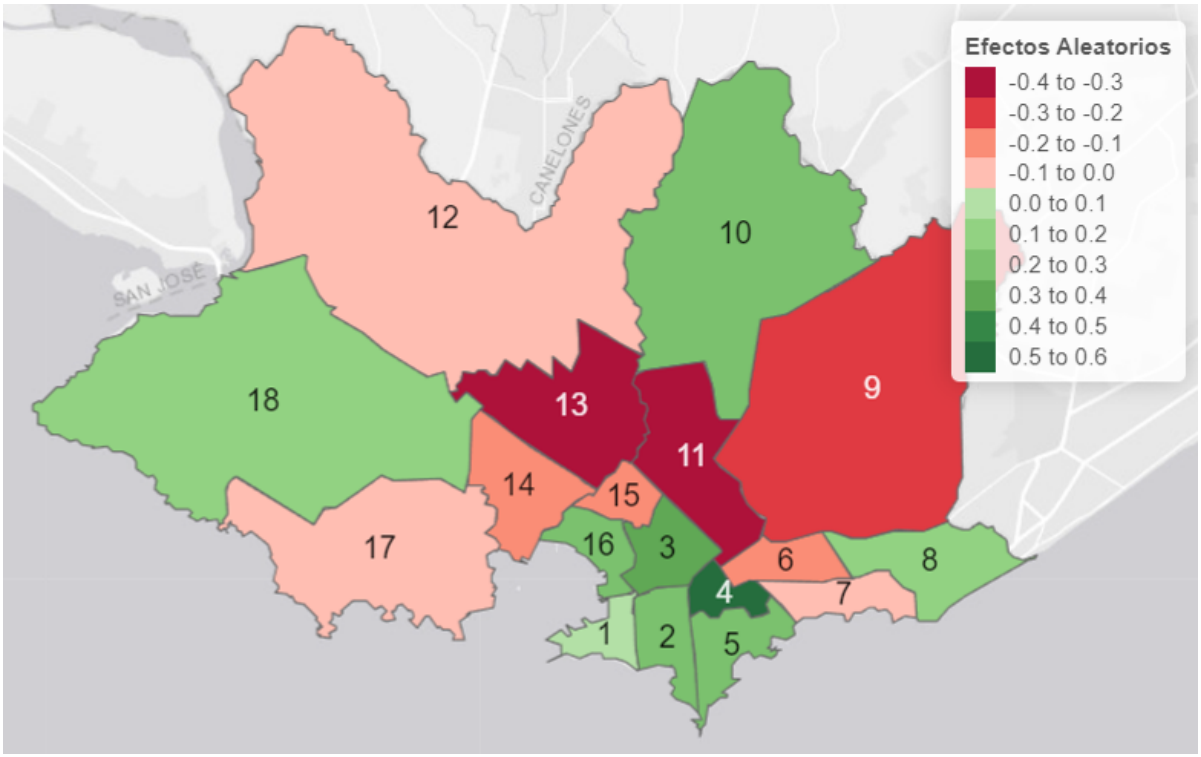

Figura 4.8: $e^{\phi} - 1$ de los efectos aleatorios, del modelo 4.6.

Los CCZ 6, 7, 9, 11, 12, 13, 14, 15 y 17 en distintas tonalidades de rojo, presentan efectos negativos, mientras que los CCZ 1, 2, 3, 4, 5, 8, 10, 16 y 18 en tonalidades de verde presentan efectos positivos sobre la probabilidad de pago de BIM6. Es decir, las observaciones presentes en los CCZ 13 y 11, disminuyen en 35,7 % y 32,6 % respectivamente, las odds de BIM6, mientras que observaciones pertenecientes a los CCZ 4 y 3, aumentan las odds de BIM6 en 56,7 % y 36,1 % respectivamente. En el cuadro 4.12, se pueden observar los coeficientes de los efectos aleatorios para cada CCZ.

Cuadro 4.12: Efectos aleatorios ϕ y odds-ratio por CCZ, modelo CAR 4.6.

CCZ	ϕ	$e^{\phi} - 1$
1	0.0911	0.0954
2	0.2591	0.2957
3	0.3083	0.3611
4	0.4494	0.5674
5	0.2179	0.2435
6	-0.1629	-0.1503
7	-0.0968	-0.0923
8	0.1141	0.1209
9	-0.3422	-0.2898
10	0.2601	0.2971
11	-0.3910	-0.3236
12	-0.0809	-0.0777
13	-0.4382	-0.3548
14	-0.1426	-0.1329
15	-0.2086	-0.1883
16	0.2588	0.2953
17	-0.0514	-0.0501
18	0.1090	0.1151

El *WAIC* del modelo CAR es 714,6 con un desvío de 37,4 mejorando 5,3 puntos con respecto al modelo sin efectos aleatorios (modelo 4.4), a menor WAIC, mejor grado de ajuste dado que este indicador está en la misma escala que AIC y DIC.

4.7. Modelo Logístico - ICAR

Los modelos ICAR son un caso particular del CAR, donde el coeficiente α que mide el grado de correlación espacial, se asume que es igual a 1. Los autores Banerjee, Carlin, y Gelfand (2014), recomiendan trabajar con esta especificación intrínseca CAR, e ignorar la impropiedad de la especificación estándar CAR, al fin y al cabo, sólo se utiliza el modelo CAR como previa, y por lo general, la posterior seguirá emergiendo como adecuada, por lo que la inferencia bayesiana será adecuada (Banerjee, cols, 2014, pág. 155).

Siguiendo la misma lógica utilizada para la construcción de los modelos, el modelo ICAR

presenta las variables explicativas *ANTIGUEDAD* y *PROP_BIM_IMP* con efectos fijos para todas las TDS en los i CCZ, mientras que ϕ_i es el efecto aleatorio del i-esimo CCZ. En la ecuación 4.8 se presenta el modelo ICAR y en 4.9 las previas especificadas:

$$\begin{aligned} \log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_0 + \beta_1 \text{ANTIGUEDAD}_{ij} + \beta_2 \text{PROP\_BIM\_IMP}_{ij} + \phi_i \\ i &= 1,\ldots,M. \quad j = 1,\ldots,n. \end{aligned} \tag{4.8}$$

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(0, 10^2) \\ \beta_1 &\sim \mathcal{N}(0, 10^2) \\ \beta_2 &\sim \mathcal{N}(0, 10^2) \\ \boldsymbol{\Phi} &\sim \mathcal{N}\left(\mathbf{0}, \left(\frac{1}{\sigma}\boldsymbol{D} - \boldsymbol{W}\right)^{-1}\right) \\ \sigma &\sim t_3(0, 2.5) \cdot \mathbb{I}(\sigma > 0) \end{aligned} \tag{4.9}$$

A diferencia del modelo CAR en 4.6, la especificación intrínseca, asume autocorrelación espacial total entre los efectos aleatorios ($\alpha = 1$). Esto simplifica la matriz de varianzas y covarianzas de la distribución previa de $\boldsymbol{\Phi}$ a $(\boldsymbol{D} - \boldsymbol{W})^{-1}$. Una de las ventajas de asumir $\alpha = 1$ es que la interpretación es exactamente lo que queremos, mientras que con la inclusión de α producía un suavizado sobre los efectos aleatorios espaciales de modo que $\mathbb{E}(\phi_i) = \alpha \sum_{j\sim i} w_{ij}\phi_j$. Por otro lado, cuando $\alpha = 1$ la esperanza de ϕ_i pasa a ser el promedio del total de vecinos para la observación i. La distribución previa de CAR (ecuación 2.5.1, generalmente no ofrece suficiente similitud espacial a menos que α esté bastante cerca de 1, y de hecho, en el ajuste del modelo, α suele ser cercano a 1 (Banerjee y cols., 2014); esto incluso se puede notar levemente en la figura 4.7.

Los resultados de la etapa inferencia del modelo 4.8 se presentan en el cuadro 4.13.

Cuadro 4.13: Resumen del Modelo logístico ICAR

	Estimación	Desvío	IP inf. 95 %	IP sup. 95 %	Rhat	Bulk ESS
Intercept	2.476	0.226	2.056	2.932	1.010	634
ANTIGUEDAD	-0.095	0.012	-0.119	-0.071	1.015	303
PROP_BI	-0.111	0.011	-0.134	-0.092	1.003	4385
σ_{ICAR}	0.702	0.313	0.148	1.411	1.027	198

En el diagnóstico del modelo, la estimación de los coeficientes son consistentes y no presentan problemas de convergencia, el $\hat{R}$ y *Bulk ESS* son adecuados tras 7000 iteraciones donde la mitad fueron utilizadas como calentamiento. En el apéndice A.1 se pueden visualizar los trace y posterior plots para los 4 coeficientes. Por otro lado, la distribución

posterior de σ_{ICAR}, presente en la figura 4.9 muestra una región estimada con precisión, la media representada por la línea vertical en la figura, tiene un valor 0,702 con un desvío de 0,3 es una estimación alta en relación a los valores obtenidos para los ϕ_i presentes en el cuadro 4.14 y en la figura 4.10, dado que el promedio de los efectos aleatorios en valor absoluto ronda en 0,221 .

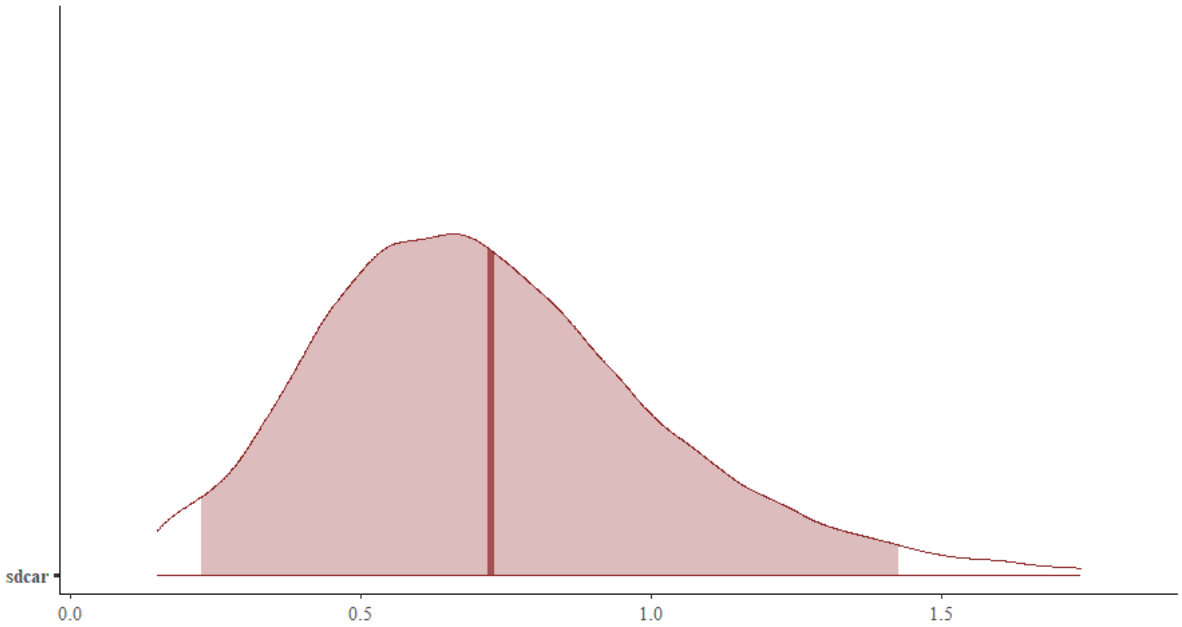

Figura 4.9: Posterior plot de la estimación del desvío σ_{CAR} del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad.

Cuadro 4.14: Efectos aleatorios ϕ y odds-ratio por CCZ, modelo ICAR 4.8.

CCZ	ϕ	$e^{\phi} - 1$
1	0.0911	0.0954
2	0.2591	0.2957
3	0.3083	0.3611
4	0.4494	0.5674
5	0.2179	0.2435
6	-0.1629	-0.1503
7	-0.0968	-0.0923
8	0.1141	0.1209
9	-0.3422	-0.2898
10	0.2601	0.2971
11	-0.3910	-0.3236
12	-0.0809	-0.0777
13	-0.4382	-0.3548
14	-0.1426	-0.1329
15	-0.2086	-0.1883
16	0.2588	0.2953
17	-0.0514	-0.0501
18	0.1090	0.1151

Las observaciones pertenecientes a los CCZ 13, 11 presentan una disminución en 35,5 % y 32,4 % respectivamente las odds de BIM6, un efecto ligeramente menor que en el modelo CAR, mientras que observaciones pertenecientes a CCZ como el 4 y 3 aumentan en 56,7 % y 36,1 % las odds de BIM6 respectivamente.

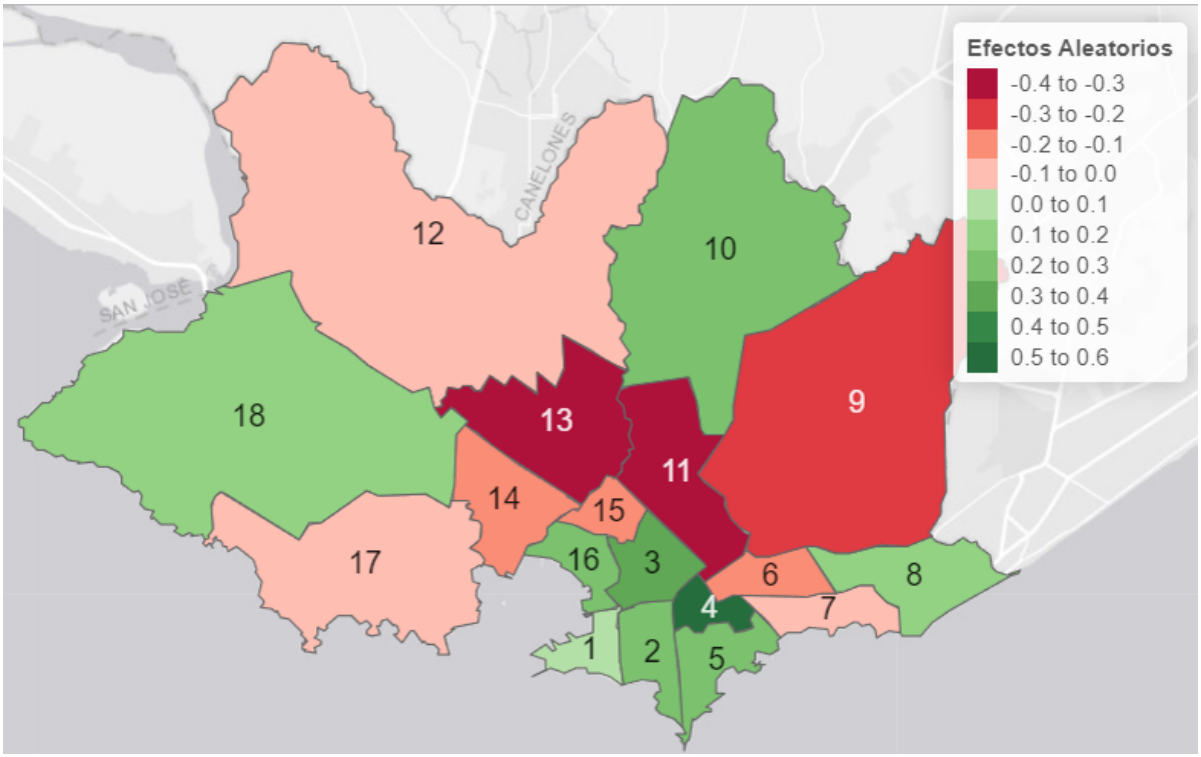

Figura 4.10: $e^{\phi} - 1$ de los efectos aleatorios, del modelo 4.8.

El error global del modelo alcanza un 16,68 %, la sensibilidad 96,25 %, la especificidad 69,3 % y la precisión 77,26 %. En términos de ajuste de predicción, el modelo CAR supera al modelo actual por 0,86 puntos, sin embargo, frente al modelo 4.4 que no incluye efectos aleatorios, mejora en 1,63 puntos.

El *WAIC* del modelo es de 713,0 con un desvío de 37,4 superando en 1,6 puntos al del modelo CAR. En términos de ajuste dentro de la muestra, el modelo ICAR no supera al modelo CAR según el computo del error global, sin embargo dado el indicador WAIC, nos dice que el valor esperado para observaciones futuras es mejor en el ICAR que en el CAR.

5. Conclusiones

Con el objetivo de encontrar el modelo estadístico que mejor prediga la probabilidad de pago del bimestre próximo de los usuarios de TDS de Montevideo, se recurrió a una serie de modelos logísticos. Luego de llevar a cabo una selección de observaciones a predecir y una selección de variables explicativas, se optó por incluir efectos aleatorios cuya correlación dependen de la posición geográfica del CCZ. Se implementaron los modelos CAR e ICAR sobre el modelo obtenido en 4.4, para capturar la variabilidad espacial presentes entre las unidades de análisis anidadas en los CCZ de Montevideo.

Se decidió limitar el análisis a la información recopilada desde la oficina de Saneamiento de la Intendencia de Montevideo. Es decir, se decidió no utilizar ninguna variable socio-económica de la población de estudio, esta omisión de variables podría cambiar los resultados considerablemente, e incluso capturar la correlación espacial presente en la variable de respuesta y en los residuos del modelo que se demostró. Entre las variables relevantes omitidas se encuentra el valor catastral a nivel de padrón, como también variables a nivel de CCZ que puedan representar la capacidad de pago del usuario, como el ingreso promedio, desempleo, nivel educativo entre otros.

Con respecto a la predicción, se concluye que hay dos subconjuntos de TDS, aquellas con las que se cuenta con información de pago de al menos el último bimestre y aquellas que no, las cuáles representan el 99,68 % y el 0,32 % respectivamente de la muestra total. Para la mayor proporción de TDS, se concluye que el mejor modelo es aquel que incluye la variable BIM5 como variable explicativa de la variable de respuesta BIM6. La tabla de doble entrada presente en el cuadro 4.2 verifica que si la TDS abonó el bimestre anterior, en el 96,4 % de los casos se paga el bimestre próximo, en cambio si la TDS no abona el bimestre anterior, en el 91,7 % de los casos, tampoco abona el bimestre siguiente. En resumen, se predice con un 92,6 % de efectividad, el pago del bimestre próximo.

Por otro lado, el subconjunto menor de observaciones, corresponden a TDS que no se tiene información de pago del bimestre anterior, por tratarse de nuevas o inactivas TDS. Para estas observaciones, se concluye que los mejores modelos son aquellos que incluye las variables explicativas proporción de bimestres impagos, antigüedad, y los efectos aleatorios a nivel de CCZ que capturan la variabilidad espacial presente en el modelo.

Los resultados asociados al menor subconjunto de observaciones, muestran una secuencia de modelos logísticos que mejoran en términos de ajuste de predicción dentro (error global) y fuera (WAIC) de la muestra. Los mejores modelos fueron los CAR e ICAR.

Estos modelos presentas tasas de efectividad del 84,2 % y 83,3 % respectivamente. Ambos tienen en común la especificación de efectos fijos (antigüedad y proporción de bimestres impagos) y efectos aleatorios relacionados a la geografía y definición de vecindad de las unidades de área.

Esta clase de modelos permite identificar a los usuarios con baja probabilidad de pagar el bimestre próximo de la TDS. Esto puede ser de gran de ayuda para identificar regiones, CCZ donde la intención de pago de los usuarios sea baja, y acudir con el objetivo de generar un cambio y encontrar una solución.

Una consideración a tener en cuenta, es que el tipo de observaciones adecuadas para este modelo, son aquellas sobre las que no se tiene información de pago sobre los últimos 6 bimestres de la TDS ya sea por tratarse de nuevas cuentas de TDS, o por ser cuentas viejas de TDS que reactivan sus consumos y por lo tanto sus vertimientos. En el caso de las observaciones con información completa o parcial en los 6 bimestres anteriores, la mejor predicción es aquella que es igual a su última realización. Esto indica que si pagó, volverá a pagar y si no pagó, continuará sin pagar. No hay variable explicativa que represente la intención de pago de los usuarios por lo que resulta muy difícil predecir un cambio en la intención, ya sea que ese cambio pueda venir por un descuido (olvidar de pagar), falta de interés, imposibilidad de pago o incluso desinformación en conocimiento de su obligación de pago.

Una de la trayectorias exploradas del trabajo, fue la definición de vecinos por distancia de las TDS con respecto al centroide de su padrón. Se detectó la presencia de autocorrelación espacial presente en los residuos de devianza de los modelos con una matriz de vecinos definida con respecto a determinada distancia. El desempeño de los modelos CAR e ICAR no fue el esperado a nivel de padrones. Por tal razón estos modelos fueron descartados.

Un aspecto pendiente a explorar fue el tipo de red de saneamiento con el que cada TDS cuenta, existen distintos tipos de colectores en la red de saneamiento distribuidos en el departamento, entre ellos colectores unitarios, mixtos, impulsión. Esto puede motivar preguntas como; ¿el tipo de colector influye sobre la probabilidad de pago? Por otro lado, estudiar la clasificación de los grupos de efectos aleatorios a nivel de barrios ¿permite capturar un efecto mas preciso? Existe otra clase de modelos asociados a los modelos Autorregresivos Condicionales (CAR), conocidos como Besag York Mollié (BYM) modelos que introducen nuevos conceptos y que podrían ajustar y predecir mejor. Aspectos como estos, quedan pendientes de contestar en posibles futuros trabajos.

A. Apéndices

A. Resultados Complementarios

A.1. Visualizaciones y diagnóstico

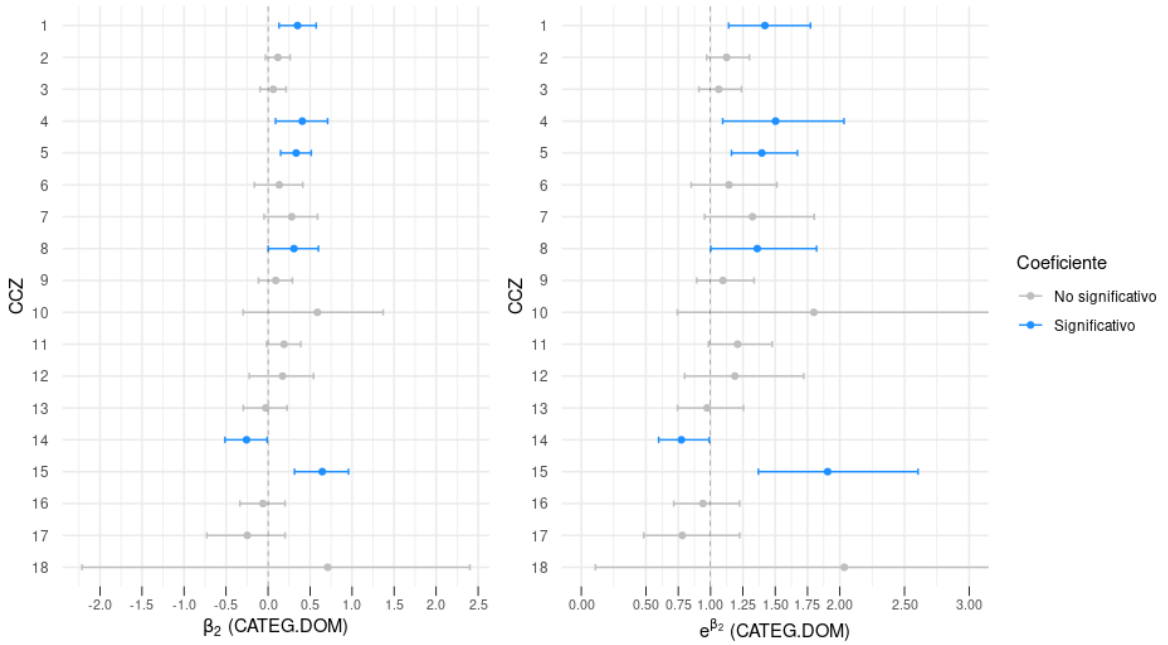

Figura A.1: Coeficientes estimados asociados a la variable *CATEG.DOM* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha).

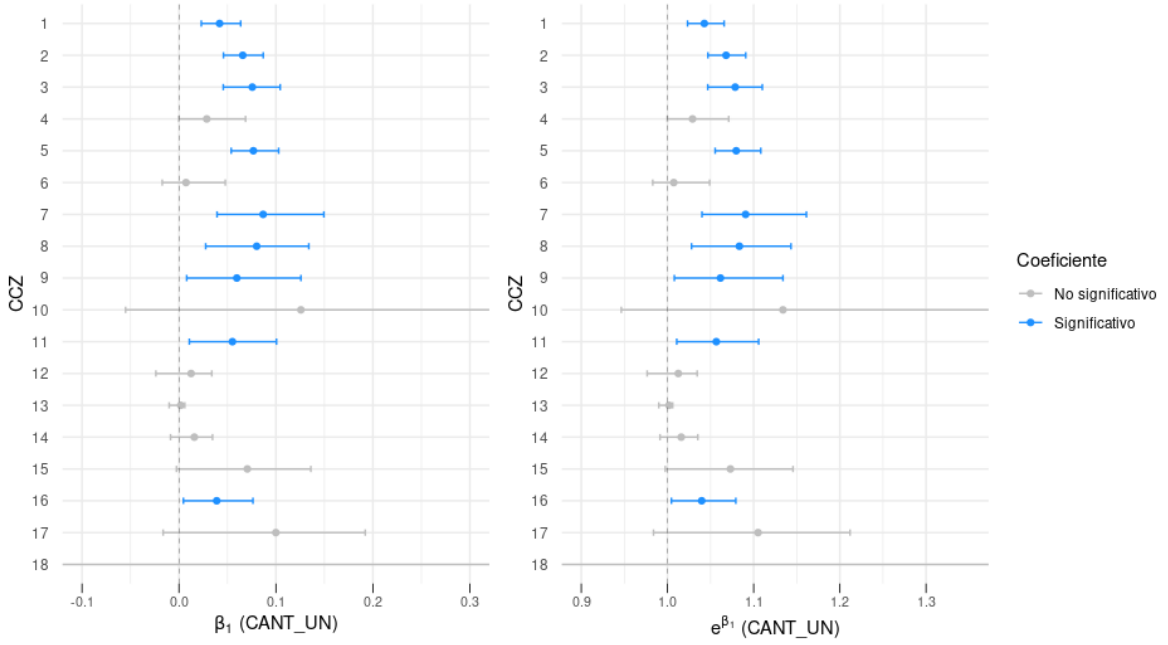

Figura A.2: Coeficientes estimados asociados a la variable *CANT_UN* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha).

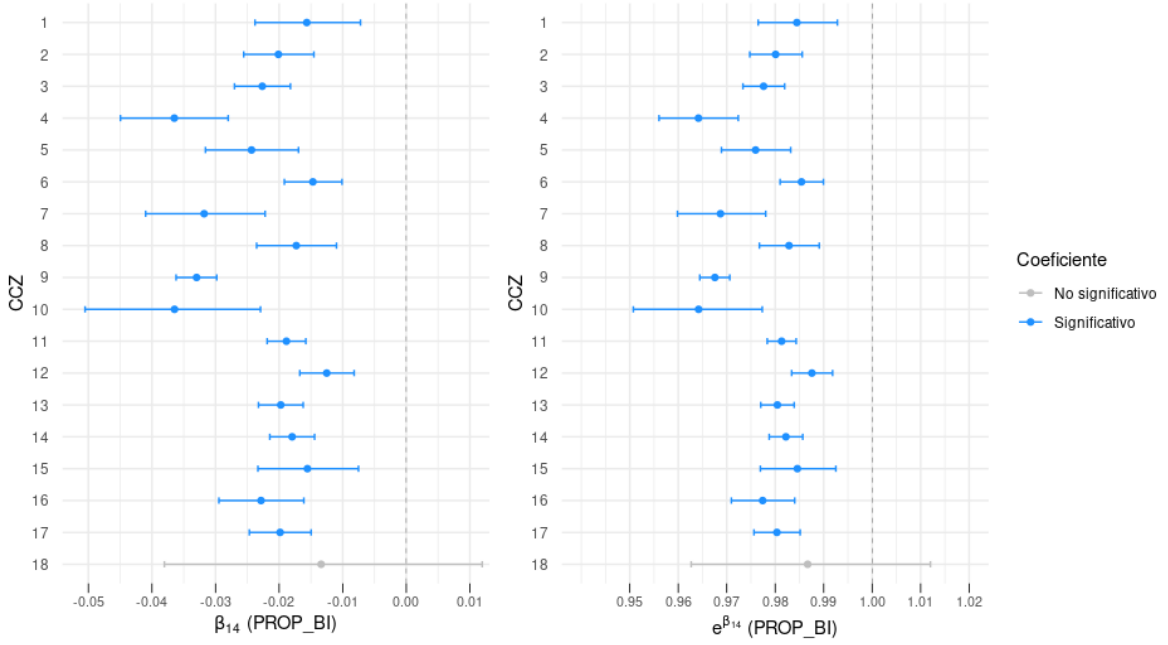

Figura A.3: Coeficientes estimados asociados a la variable *PROP_BIM_IMP* para los 18 modelos, en escalas log-odds (izquierda) y odds-ratio (derecha).

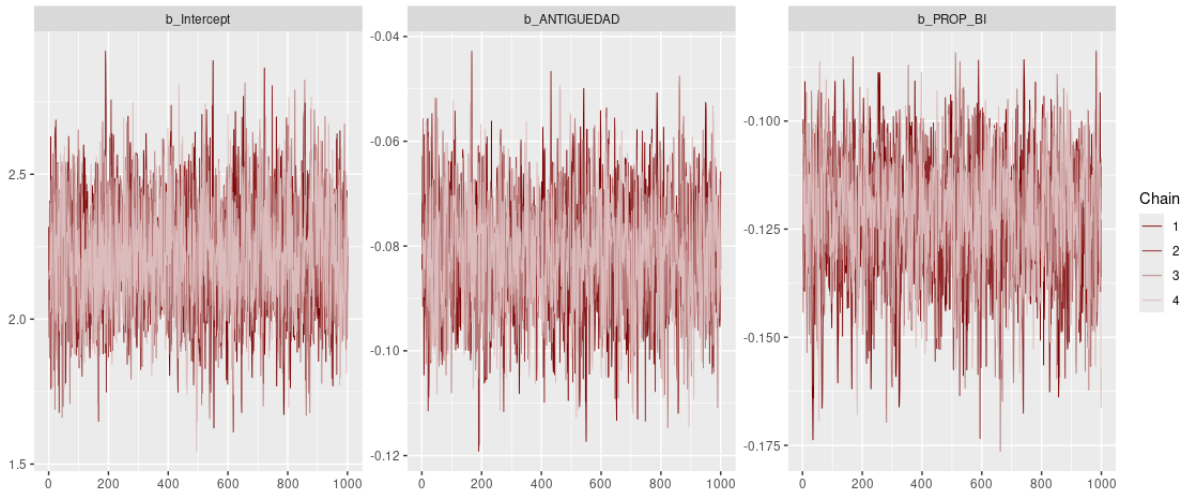

Figura A.4: Trace plots del modelo 4.4.

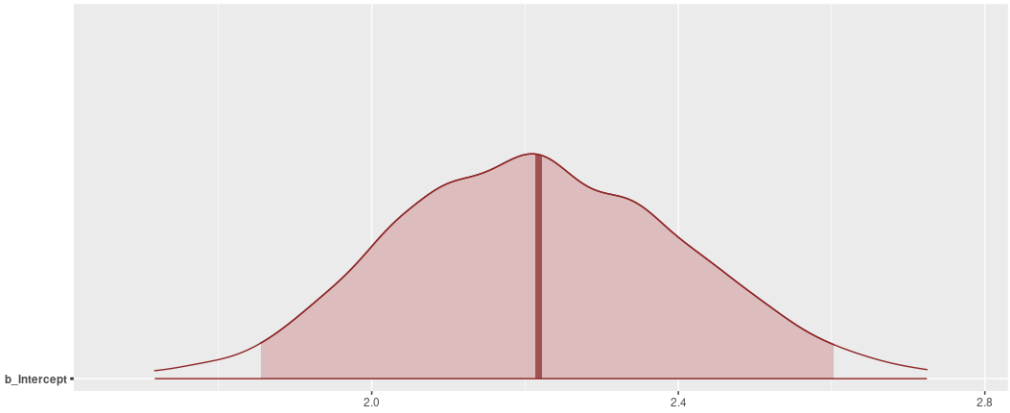

Figura A.5: Posterior plot del *Intercept* del modelo 4.4, el área sombreada es el intervalo al 95 % de probabilidad.

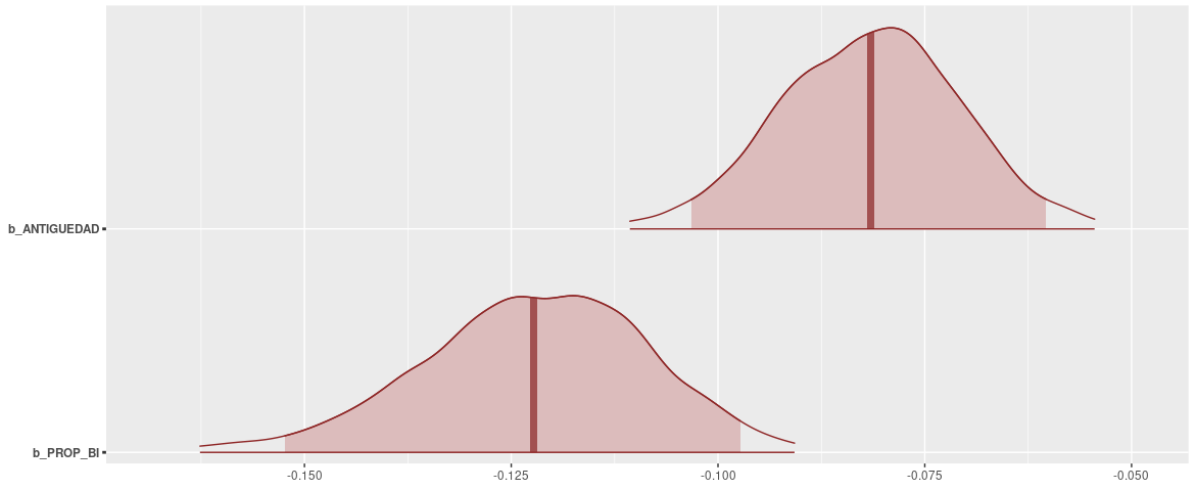

Figura A.6: Posterior plot del modelo 4.4, el área sombreada es el intervalo al 95 % de probabilidad.

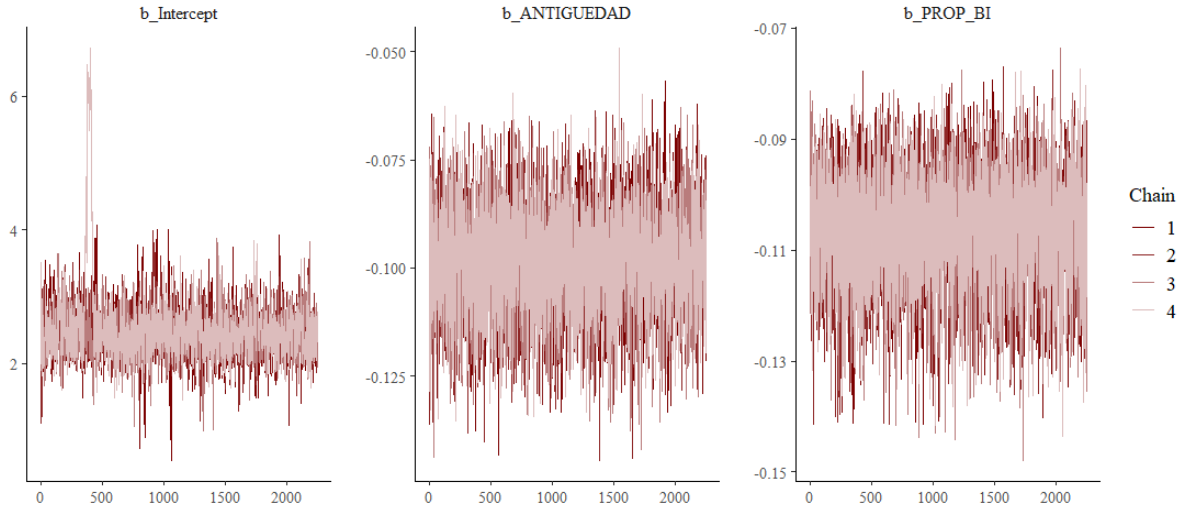

Figura A.7: Trace plots del modelo 4.6.

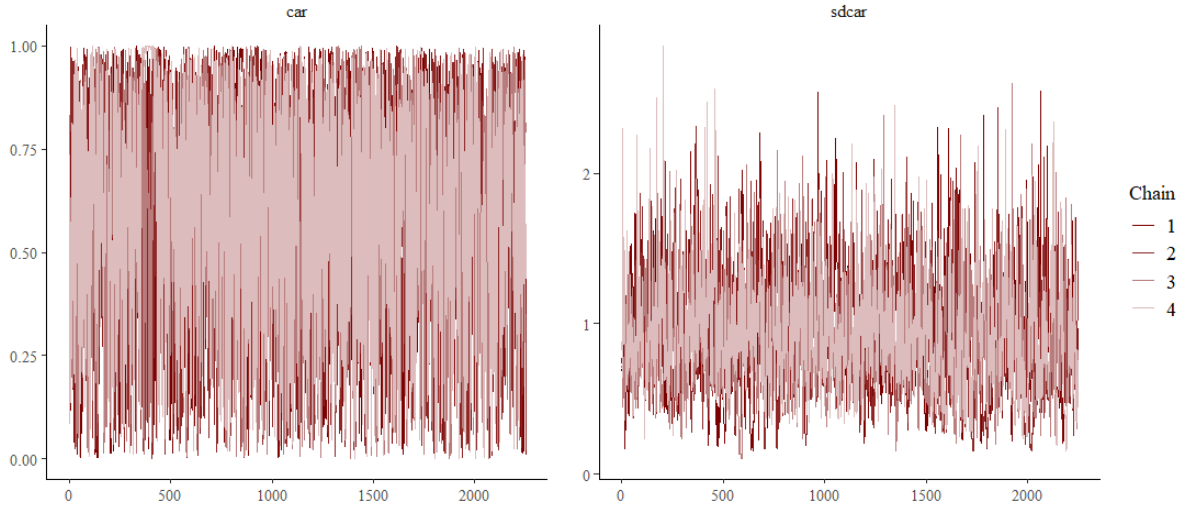

Figura A.8: Trace plots del modelo 4.6.

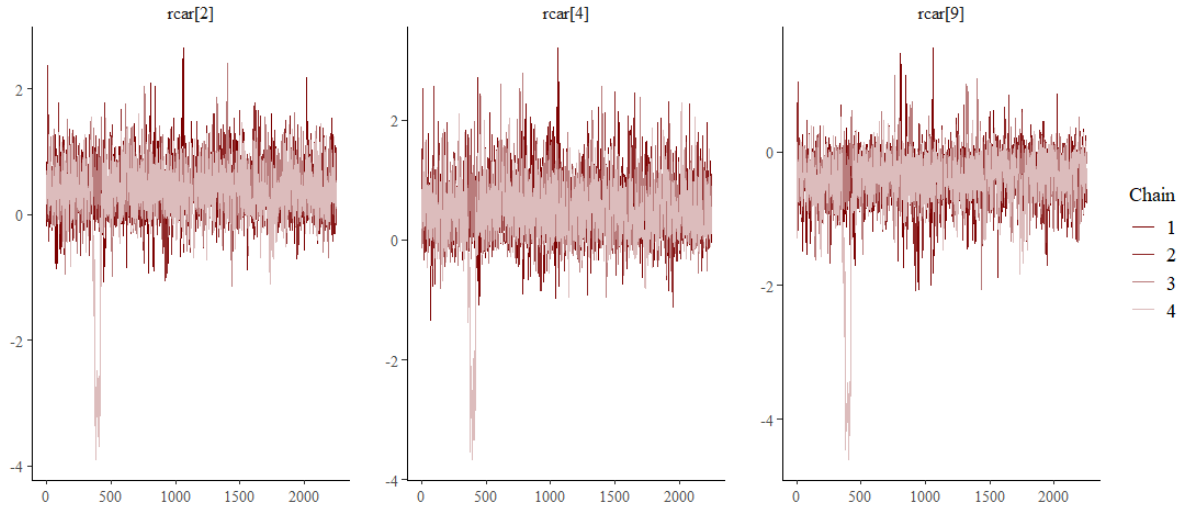

Figura A.9: Trace plots del modelo 4.6.

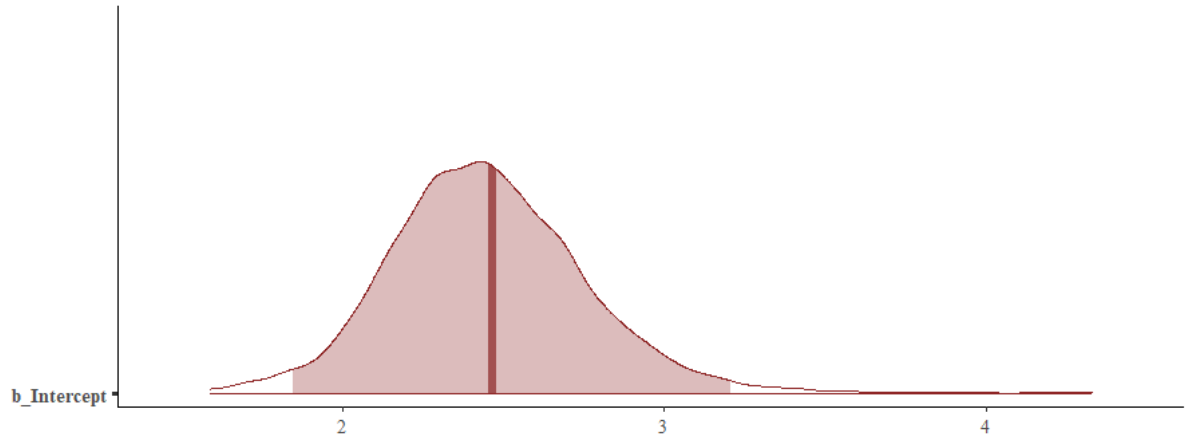

Figura A.10: Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad.

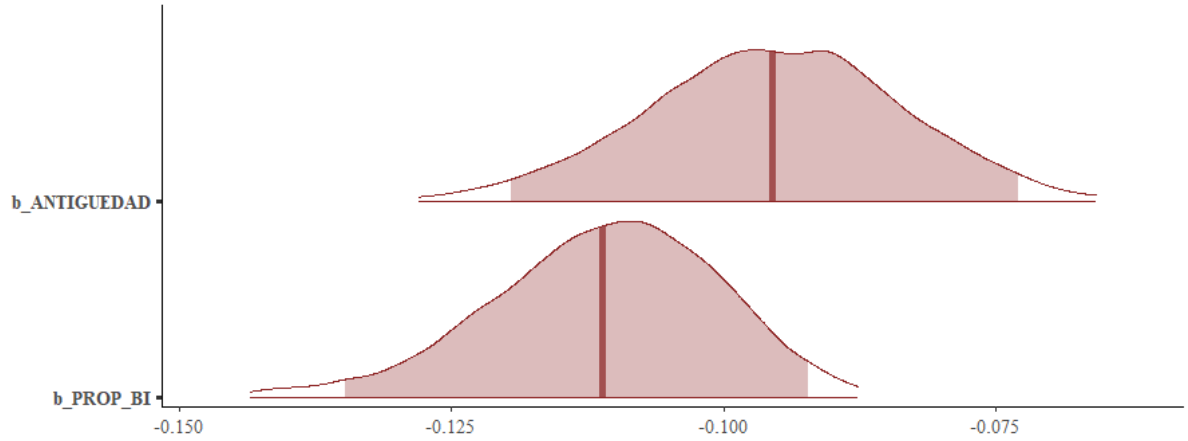

Figura A.11: Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad.

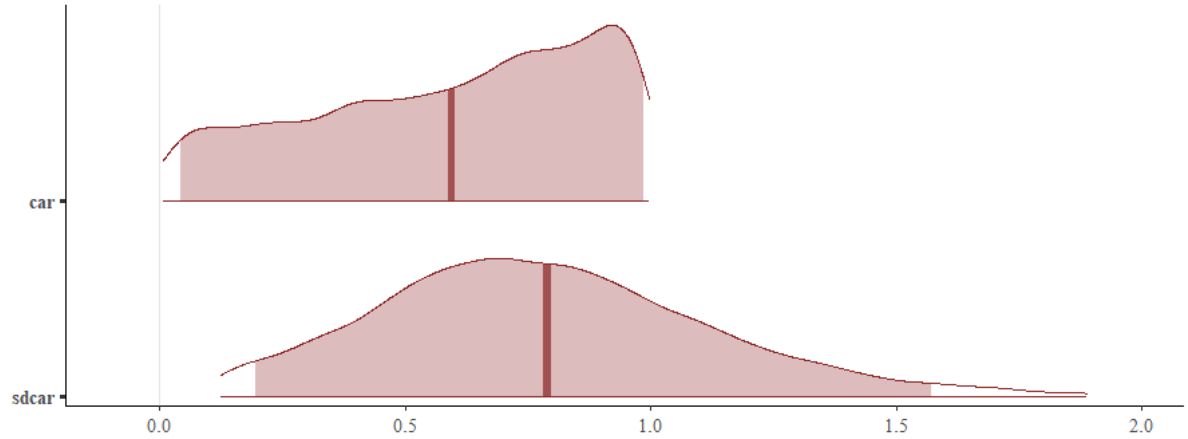

Figura A.12: Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad.

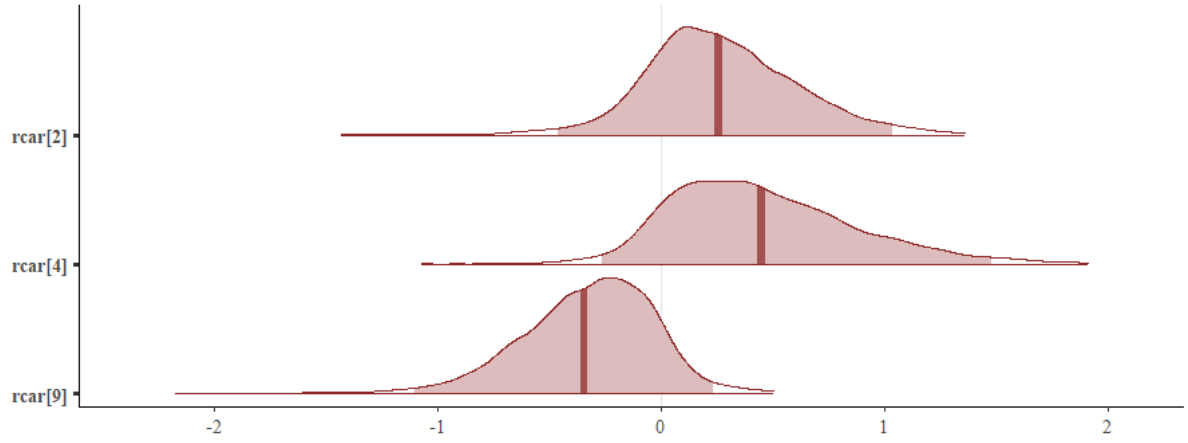

Figura A.13: Posterior plot del modelo 4.6, el área sombreada es el intervalo al 95 % de probabilidad.

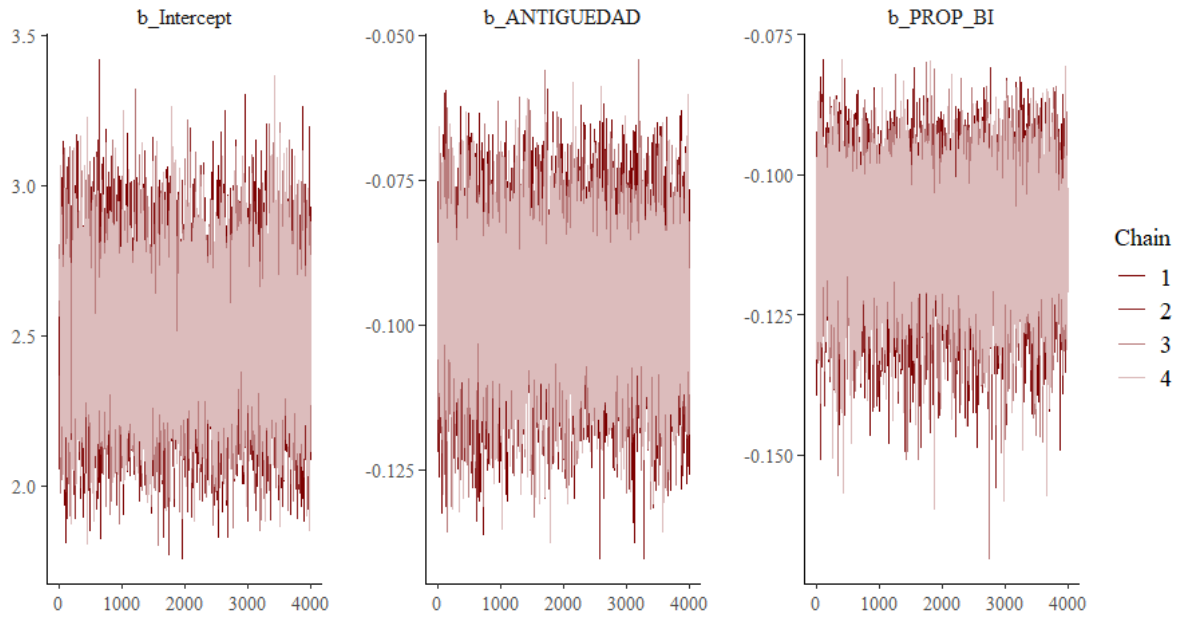

Figura A.14: Trace plots del modelo 4.8.

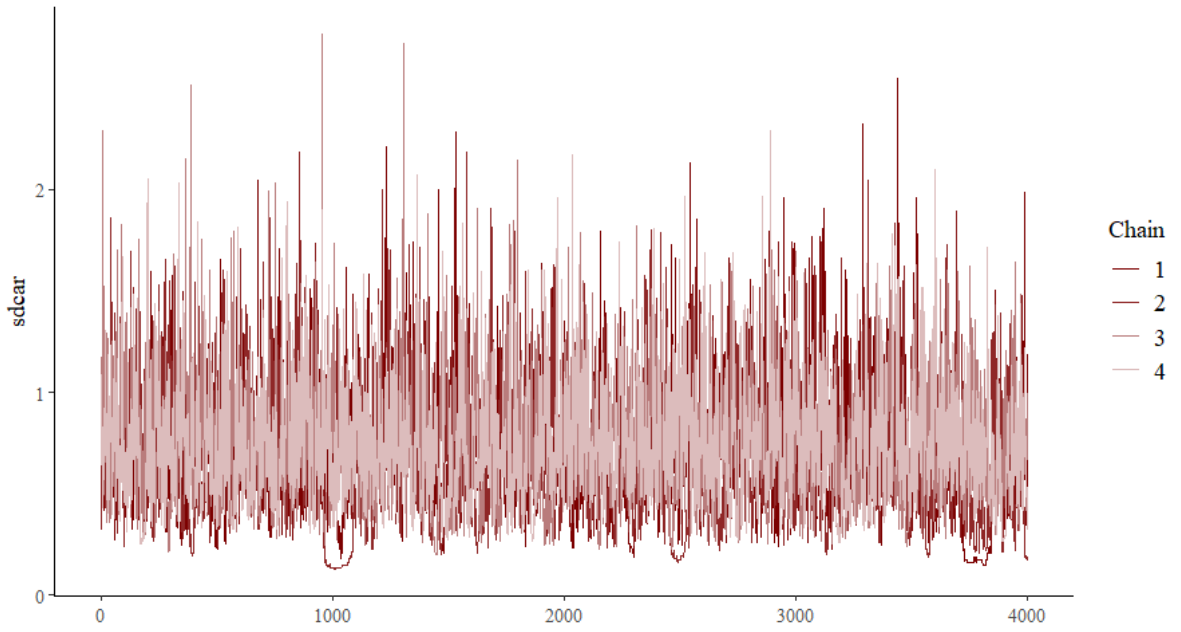

Figura A.15: Trace plots del modelo 4.8.

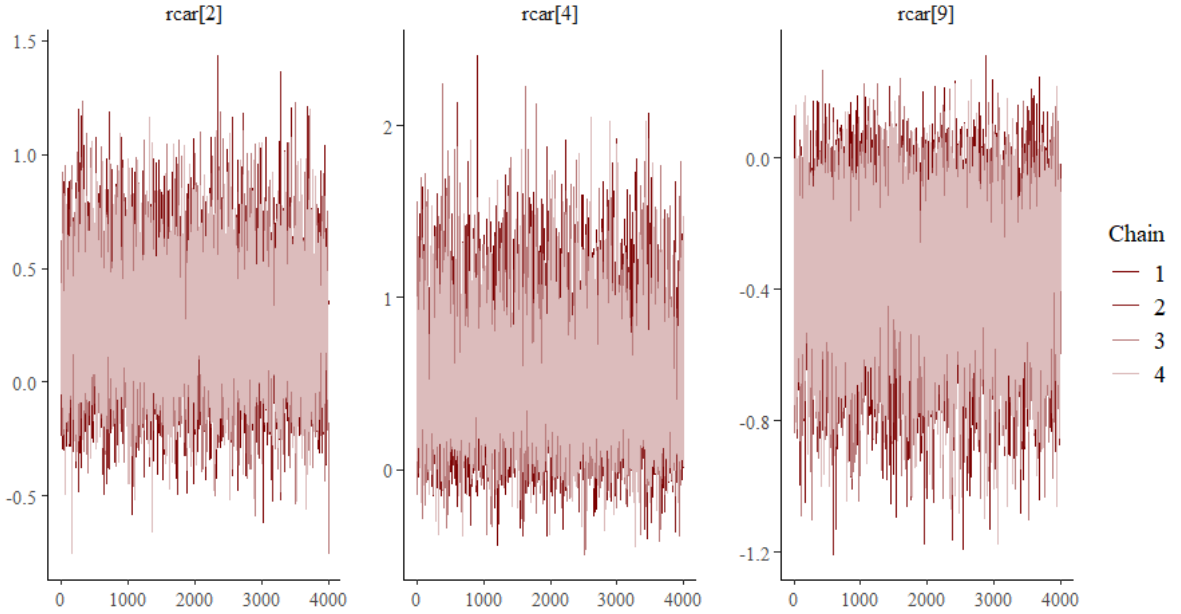

Figura A.16: Trace plots del modelo 4.8.

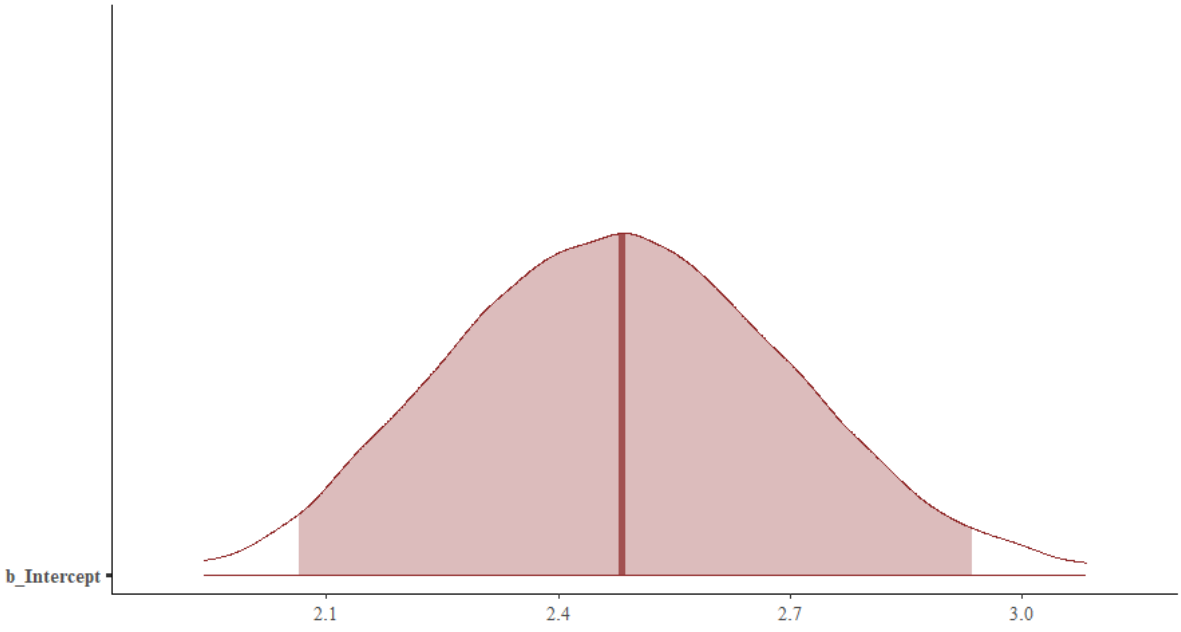

Figura A.17: Posterior plot del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad.

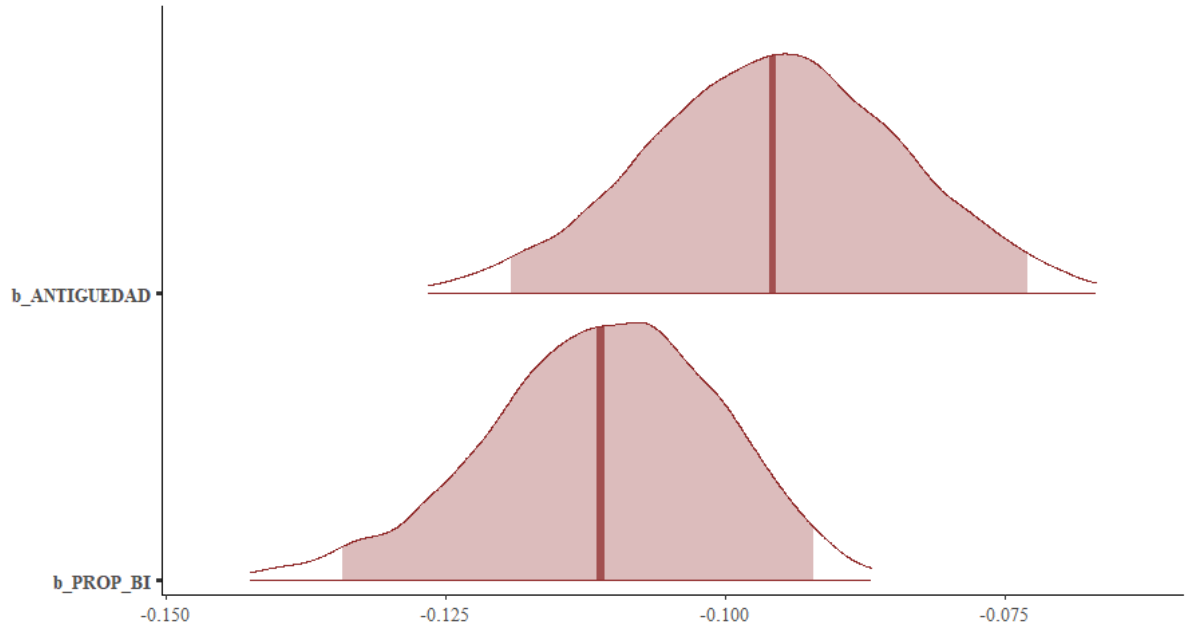

Figura A.18: Posterior plot del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad.

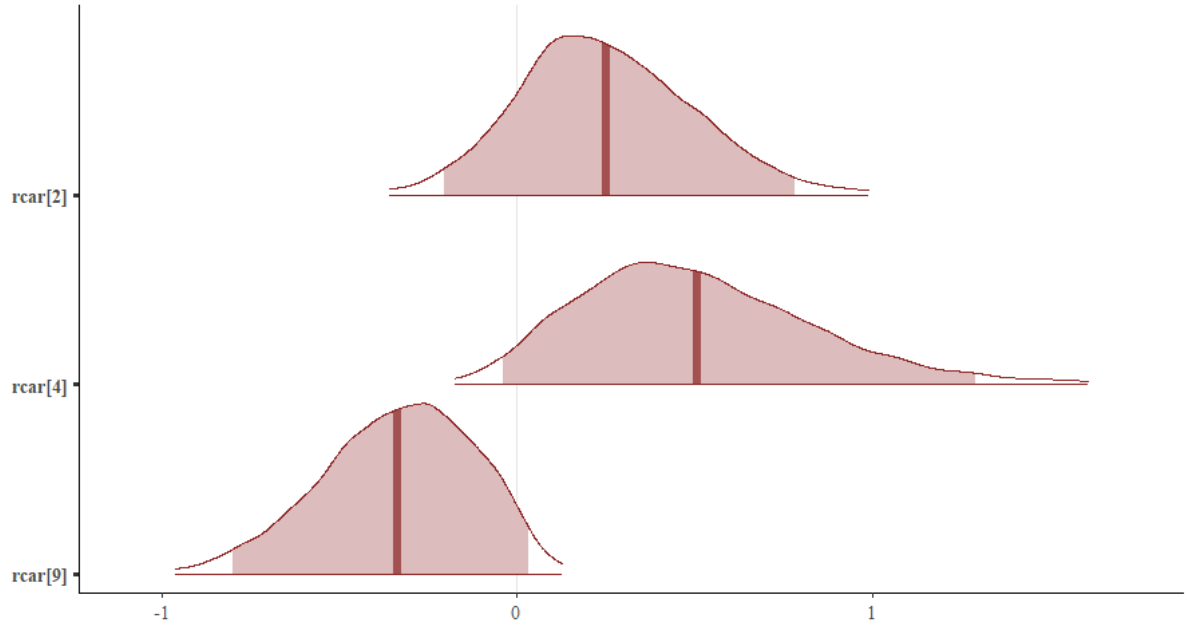

Figura A.19: Posterior plot del modelo 4.8, el área sombreada es el intervalo al 95 % de probabilidad.

Referencias

Agresti, A. (2015). Foundations of Linear and Generalized Linear Models.. Descargado de https://api.semanticscholar.org/CorpusID:153029174

Anselin, L. (1988). Spatial Econometrics: Methods and Models (Vol. 4). Dordrecht: Springer Netherlands. Descargado 2024-01-31, de http://link.springer.com/10.1007/978-94-015-7799-1 https://doi.org/10.1007/978-94-015-7799-1

Banerjee, S., Carlin, B. P., y Gelfand, A. E. (2014). Hierarchical Modeling and Analysis for Spatial Data (2nd Edition ed.). New York: Chapman and Hall/CRC. Descargado de https://doi.org/10.1201/b17115 https://doi.org/10.1201/b17115

Besag, J. (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), 192–236. Descargado 2023-11-01, de https://www.jstor.org/stable/2984812 (Publisher: [Royal Statistical Society, Wiley])

Bivand, R., y Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. TEST, 27(3), 716–748. (Version utilizada de la librería 1.2-8) https://doi.org/10.1007/s11749-018-0599-x

Bivand, R. S., Pebesma, E., y Gómez-Rubio, V. (2008). Applied Spatial Data Analysis with R. New York, NY: Springer. Descargado 2023-09-27, de https://link.springer.com/10.1007/978-1-4614-7618-4 https://doi.org/10.1007/978-1-4614-7618-4

Bürkner, P.-C. (2017). **brms** : An *R* Package for Bayesian Multilevel Models Using *Stan*. Journal of Statistical Software, 80(1). Descargado 2024-01-30, de http://www.jstatsoft.org/v80/i01/ https://doi.org/10.18637/jss.v080.i01

Cliff, A. D., y Ord, J. K. (1973). Spatial Autocorrelation. London: Pion.

Comisión Económica para América Latina y el Caribe. (2024). Estudio Económico de América Latina y el Caribe, 2024: trampa de bajo crecimiento, cambio climático y dinámica del empleo. Descargado de https://hdl.handle.net/11362/80595

Diniz-Filho, J. A. F., Bini, L. M., y Hawkins, B. A. (2003). Spatial autocorrelation and red herrings in geographical ecology. Global Ecology and Biogeography, 12(1), 53–64. Descargado 2024-01-31, de https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1466-822X.2003.00322.x (Publisher: John Wiley & Sons, Ltd)

Grossman, E., Marcus, L., Palmer, E., Pulham, K., y Rumments, A. (2021). An Introduction to Generalized Linear Models. Oregon State University. Descargado 2023-08-04, de https://generalizedregressionmodelingagency.github

.io/GRMA/

Guo, J., Gabry, J., Goodrich, B., Weber, S., Lee, D., Sakrejda, K., ... Steve, B. (2023, enero). rstan: R Interface to Stan. Descargado 2023-08-09, de https://cran.r-project.org/web/packages/rstan/

Hoff, P. D. (2009). A First Course in Bayesian Statistical Methods. New York, NY: Springer. Descargado 2023-08-04, de http://link.springer.com/10.1007/978-0-387-92407-6 https://doi.org/10.1007/978-0-387-92407-6

Junta Departamental de Montevideo. (2001). Decreto Nº 29.434 de la Junta Departamental de Montevideo (Presupuesto 2001 – 2005). Descargado de https://example.com/decreto_29434

McCullagh, P., y Nelder, J. A. (1989). Generalized Linear Models, Second Edition. CRC Press. (Google-Books-ID: 7tI_tQEACAAJ)

Morris, M., Wheeler-Martin, K., Simpson, D., Mooney, S. J., Gelman, A., y DiMaggio, C. (2019). Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan. Spatial and Spatio-temporal Epidemiology, 31, 100301. Descargado de https://www.sciencedirect.com/science/article/pii/S1877584518301175 https://doi.org/https://doi.org/10.1016/j.sste.2019.100301

Organización Mundial de la Salud. (2023). Saneamiento. Descargado 2024-01-29, de https://www.who.int/es/news-room/fact-sheets/detail/sanitation

Organización Naciones Unidas. (2023). Objetivos de Desarrollo Sostenible. Descargado 2024-01-29, de https://news.un.org/es/story/2023/02/1518287

Organización Panamericana de la Salud. (2023). Día Interamericano del Saneamiento. Descargado 2024-01-29, de https://www.paho.org/es/noticias/16-11-2023-dia-interamericano-saneamiento-508-america-latina-caribe-no-tiene-saneamiento

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, 10(1), 439–446. Descargado de https://doi.org/10.32614/RJ-2018-009 https://doi.org/10.32614/RJ-2018-009

Peña Barreto, J. A. (2016). Saneamiento Ambiental y Participación Ciudadana. Revista Scientific, 1(1), 53–71. Descargado de https://www.redalyc.org/journal/5636/563660226005/html/ https://doi.org/10.29394/scientific.issn.2542-2987.2016.1.1.4.53-71

Pinheiro, J., y Bates, D. (2006). Mixed-Effects Models in S and S-PLUS. Springer Science & Business Media. (Google-Books-ID: ZRnoBwAAQBAJ)

Powers, D. (2008, enero). Evaluation: From Precision, Recall and F-Factor to ROC,

Informedness, Markedness & Correlation. Mach. Learn. Technol., 2. Descargado de https://www.researchgate.net/publication/228529307_Evaluation_From_Precision_Recall_and_F-Factor_to_ROC_Informedness_Markedness_Correlation

Prindle, R. A., y Salas, A. C. (1967). Importancia del Saneamiento Ambiental para la Salud de la Comunidad. Boletín de la Oficina Sanitaria Panamericana, 63(4), 337–338. Descargado de https://iris.paho.org/bitstream/handle/10665.2/12638/v63n4p337.pdf?sequence=1

R Core Team. (2023). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Descargado de https://www.R-project.org/

Regueira, J. M. (2015). Saneamiento y Salud - SyS. Descargado de https://cicplata.org/wp-content/uploads/2019/08/Saneamiento_Jos-Mara-Regueira.pdf (Place: Buenos Aires)

Rencher, A. C., y Schaalje, G. B. (2008). Linear Models in Statistics. John Wiley & Sons. (Google-Books-ID: LHV_uiq6dXUC)

SIG. (2024). Descargado 2024-06-18, de https://sig.montevideo.gub.uy/

Stan Development Team. (2024). Stan Modeling Language Users Guide and Reference Manual. Descargado de https://mc-stan.org

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., y Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved \(\hatR\) for assessing convergence of MCMC. arXiv preprint arXiv:2103.08542.

Waller, L. A., y Gotway, C. A. (2004). Applied Spatial Statistics for Public Health Data. John Wiley & Sons. (Google-Books-ID: OuQwgShUdGAC)

Watanabe, S. (2010). Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory. Journal of Machine Learning Research, 11, 3571–3594. Descargado de http://jmlr.org/papers/v11/watanabe10a.html (Place: Yokohama, Japan Publisher: MIT Press)