Estimación de precios hedónicos en viviendas usadas en Bogotá, usando modelos espaciales de efectos mixtos.

Daniel Humberto Camargo Rincón

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2023

Estimación de precios hedónicos en viviendas usadas en Bogotá, usando modelos espaciales de efectos mixtos.

Daniel Humberto Camargo Rincón

Trabajo final presentado como requisito parcial para optar al título de:
Magister en Ciencias-Estadística

Director:
Ph.D. Oscar Orlando Melo Martínez

Universidad Nacional de Colombia
Facultad de Ciencias, Departamento de Estadística
Bogotá, Colombia
2023

Dedicatoria

Dedicado a todas las personas que, como el autor, saben que hay momentos en que las cosas parecen no avanzar. Sin embargo, lo más importante en esos momentos es tener la determinación para intentar hasta lograr avanzar.

No te rindas.

Agradecimientos

A mis padres, Jaime Camargo y Maria Rincón, que siempre me estuvieron apoyando en todos los sentidos de la palabra.

A mi hermana Maria Paula, quien siempre me supo animar a seguir adelante, y de quien puedo estar seguro que sin ella no sería la persona que soy hoy en día.

A mi tia Alicia Quiñones Ocampo, quien siempre me ha escuchado y apoyado cuando lo he necesitado a lo largo de los años, desde que era un bebe y ahora en mi vida adulta.

Al profesor Oscar Orlando Melo, de quien aprendí todo lo que sé de modelos de efectos mixtos, y que gracias a su ayuda, orientación, tiempo, correcciones y enseñanzas pude terminar este documento.

Resumen

El mercado inmobiliario de vivienda es fundamental en la economía. Por un lado, éste tiene un impacto en el sector de la construcción, el cual es uno de los mayores generadores de empleo del país, mientras que, por otro lado, está su impacto en la vida de una persona y su grupo familiar. Por esto, el propósito de este trabajo fue predecir el precio de venta inicial deseado por el vendedor, buscando entender cómo se valoran las características físicas de los inmuebles, siguiendo la teoría de precios hedónicos, con una muestra final de 7850 observaciones recolectadas mediante el uso de web scraping aplicado en portales especializados en la venta de inmuebles.

Dado el factor espacial asociado a este problema (el precio de una vivienda es influenciado por las demás viviendas cercanas), se estimaron diferentes modelos bajo la teoría de la estadística espacial para seleccionar el que mejor se ajusta a los datos recolectados. Se seleccionó el modelo de efectos mixtos espaciales, puesto que fue el de mejor desempeño en las medidas del criterio de información de bayesiano (BIC, por sus siglas en inglés), error cuadrático medio (ECM) y tiempo de cómputo.

Se resalta que, mediante el modelo de efectos mixtos espaciales, se encontró que el rango de dependencia espacial del valor de una vivienda, es de cerca de 924 metros, siendo significativo al igual que el parámetro de varianza asociado al efecto aleatorio del intercepto. También se observó que, si la vivienda es una casa, tiene una mayor valoración inicial en el precio de venta deseado (aproximadamente un 5 % más alto que los apartamentos, con un valor-p menor al 0.05). Igualmente, los estratos cuatro, cinco y seis tienen precios más elevados (19 %, 23 % y 30 %, respectivamente, todos con valor-p menor al 0.05) con respecto a los estratos uno, dos y tres.

Finalmente, los hallazgos encontrados mediante el modelo de efectos espaciales mixtos reflejan que existen efectos no incluidos en el modelo (ya sea por dificultad para medir estos, para hallar información al respecto o por errores en las mediciones de las variables), que pueden afectar de una u otra manera la valorización de una vivienda, por lo que deben ser tomados en cuenta de alguna forma.

Palabras clave: Precios hedónicos, modelo multinivel, modelos autorregresivos espaciales, efectos aleatorios

Abstract

Title: Estimation of hedonic prices in used homes in Bogota, using spatial mixed effects models.

The housing real state market is fundamental in economy. This has an impact in the construction sector which is one of the largest generators of employment in our country and in the life of a person and a family group. For this reason, the main purpose of this work is to make predictions on the initial sale price desired by seller, looking for to understand how people value physical features in estate, following hedonic pricing theory. To get it, sample of 7850 observations were recolleted using web scrapping in websites focused in saling real estate.

Given the spatial factor associated with this problem (the pricing of one housing is influenced by neighborhood estates), several models were estimated under spatial statistic theory, in order to select the model which fitted the best to the collected data. The mixed effects spatial model was selected because it had the best performance under bayesian information criterion (BIC), mean squared error (MSE) and computing time.

Additionally, using the mixed effects spatial model, the spatial dependence range is about 924 meters, Being significant like the variance parameter related with random effect of intercept. As well, if the real estate is a house, the desidered pricing sale is greater (around 5 % greater than apartments, with a p-value less than 0.05). In the same way, the socioeconomic strata fourth, fifth and sixth have greater pricing (19 %, 23 % y 30 % respectively, all with p-value less than 0.05) than socioeconomic strata first, second and third.

The findings found through the spatial model of mixed effects reflect the existence of effects not included in the model (either due to the difficulty to measure them or to find information about), which can affect in one or another way the value of a housing, so they must be taken into account in some way. Finally, it is confirmed that greater strata, also, the initial pricing sale will be greater; houses have greater value than apartments, this is possibly caused by the general vision that people have about socioeconomic strata and about the space of the houses.

Keywords: Hedonic prices, multilevel model, spatial autoregressive models, random effects

Declaración

Me permito afirmar que he realizado la presente tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados en la propia tesis. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Bogotá, D.C., 20.02.2024

Lista de Figuras

2-1. Regiones del ejemplo de matriz de contigüidad y W, donde DCN es el distrito central de negocios, tomado de LeSage & Pace (2009, pág. 3) 8
2-2. Diferentes tipos de contigüidad tomado de Toloza-Delgado (2020, pág. 5) . . 8

3-1. Histograma, densidad y diagrama de caja para el precio de venta deseado y el logaritmo de este precio. 25
3-2. Cantidad de viviendas por estrato y diferenciando entre casas y apartamentos. 26
3-3. Mapa de las localidades de las zonas del centro, este, oeste y norte de Bogotá con la ubicación de las viviendas de la base de datos filtrada. 27
3-4. Gráfico de Moran de los residuos versus sus rezagos, para una matriz de pesos estandarizada de contigüidad basada en vecinos de 0-1000 metros 30
3-5. Distribución estimada de los residuales estándar condicionados del modelo (azul) y función de densidad de una distribución normal estándar (rojo) . . . 35
3-6. Valor de los factores no considerados de valorización 51

A-1. Histograma y diagrama de caja para la variable metros cuadrados y para la base reducida debajo del percentil 99. 55

D-1. Distribución estimada de los residuales condicionados estándar del modelo (azul) después de transformar la variable Precio mediante Box-Cox con un $\lambda = 0{,}1414$. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 64
D-2. Distribución estimada de los residuales condicionados estándar del modelo (azul) después de transformar la variable logaritmo del Precio mediante Box-Cox con un $\lambda = 2$. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 65
D-3. Distribución estimada de los residuales condicionados estándar del modelo (azul) después de aplicar raíz cuadrada a la variable logaritmo del Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 65

E-1. Distribución estimada de los residuales condicionados estándar del modelo (azul) después de transformar la variable Precio mediante Box-Cox con un $\lambda = 0{,}1414$. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 66
E-2. Distribución estimada de los residuales condicionados estándar del modelo (azul) después de aplicar raíz cuadrada a la variable logaritmo del Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 67
E-3. Distribución estimada de los residuales condicionados estándar del modelo (azul) después de aplicar raíz cuadrada a la variable logaritmo del Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 67
E-4. Distribución estimada de los residuales condicionados estándar del modelo (azul) utilizando el logaritmo de la variable Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo) 68

Lista de Tablas

3-1. Coeficientes estimados mediante mínimos cuadrados ordinarios regresión lineal multiple . . . 29
3-2. Test de Moran I para dependencia espacial sobre los residuales de regresión lineal . . . 30
3-3. Tabla reporte efectos aletorios del modelo mixto antes de aplicar logaritmo a la variable metros cuadrados . . . 33
3-4. Tabla reporte efectos aletorios del modelo mixto después de aplicar logaritmo a la variable metros cuadrados . . . 34
3-5. Tabla reporte efectos fijos del modelo mixto . . . 34
3-6. Tabla reporte sobre el supuesto de homocedasticidad de los errores condicionados del modelo mixto . . . 35
3-7. Tabla reporte sobre los supuestos de los efectos aleatorios del modelo mixto . 36
3-8. Coeficientes estimados modelo de efectos mixtos basado en las coordenadas . 39
3-9. Prueba de verosimilitud para los efectos aleatorios del modelo mixto basado en coordenadas . . . 40
3-10. Tabla reporte sobre el supuesto de homocedasticidad de los errores condicionados del modelo mixto basado en coordenadas . . . 40
3-11. Coeficientes estimados modelo SAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros . . . 42
3-12. Efectos directos e indirectos estimados modelo SAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros . . . 42
3-13. Coeficientes estimados modelo SEM usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros . . . 44
3-14. Pruebas de Lagrange para determinar la pertinencia de un modelo SARAR . 44
3-15. Coeficientes estimados modelo SARAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros . . . 46
3-16. Efectos directos e indirectos estimados modelo SARAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros . . . 46

3-17.Coeficientes estimados de las variables usando filtrado espacial, seleccionando vectores propios mediante stepwise backward 47
3-18.Tabla de reporte de las métricas calculadas para los modelos utilizados . . . 49
3-19.Tabla comparativa entre modelos ponderando las métricas 50

1. Introducción

El mercado de las viviendas es especial por su complejidad social, económica y territorial (Serrano, 2019). En este sentido, para analizar el sector inmobiliario es necesario alejarse de la idea original de dividir y analizar productos (ya sean bienes o servicios) homogéneos, es decir, alejarse de la concepción de productos con las mismas características cuyo precio es determinado por la oferta y la demanda del mismo, a productos diferenciables cuyo precio es determinado por sus características.

Dado lo anterior y siguiendo la línea de otros trabajos enfocados en el mercado inmobiliario, se recurre a los trabajos de Rosen (1974) y Dubin (1992) como bases teóricas con los cuales se va a abordar éste desde la perspectiva de los precios hedónicos. El primero se enfoca en generar un marco teórico para analizar productos heterogéneos donde el precio es determinado por un vector de características asociado a éste, mientras el segundo realiza una aplicación basándose en lo expuesto por el primero, incluyendo el efecto del vecindario en el valor de la variable precio, y las limitaciones de este enfoque en cuanto a variables no observadas y con error de medición mediante el uso de técnicas geoestadísticas en la ciudad de Baltimore (Estados Unidos).

La importancia de analizar el mercado inmobiliario, en especial los determinantes del precio de una vivienda, reside en la relevancia de la adquisición o no de una unidad residencial (casa o apartamento) sobre la calidad de vida de los individuos y su grupo familiar, puesto que esta no es una decisión de corto plazo como en los bienes de consumo, sino que por el contrario, por su carácter fijo (no es posible mover de lugar una vivienda) y de larga duración, es una decisión que afecta aspectos emocionales, la riqueza y el patrimonio de estos (López-Garcia, 2019; Serrano, 2019).

En Colombia el mercado inmobiliario tiene un papel vital en la economía del país, presentando durante el 2021 registros históricos de compra y venta de viviendas usadas, consolidándose como un sector sólido con índices con variaciones positivas a pesar de mostrar una desaceleración de su crecimiento frente a los últimos diez años y de los efectos de la pandemia (Casas-Lugo, 2021; El nuevo siglo, 2021; León, 2021). Adicionalmente, la construcción (uno de los principales contribuyentes al producto interno bruto y a la generación de empleo del país), es un sector muy ligado a las dinámicas del mercado de las viviendas, por lo que se hace fundamental entender las dinámicas internas que lo caracterizan.

Dentro de la literatura sobre precios hedónicos en Colombia se encuentran las publicaciones de Melo-Martínez & Melo-Martínez (2003) quienes realizaron estimaciones del valor del metro cuadrado en la ciudad de Bogotá, Morales-Zurita & Arias-Arbeláez (2005) los cuales

estimaron el precio del arriendo en función de las características de los inmuebles, Mendieta-López & Perdomo-Calvo (2007) y Perdomo-Calvo (2010), ambos trabajos enfocados en la estimación del impacto de Transmilenio en el valor de las viviendas, Zorrilla-Pérez (2012) quien estimó la valoración de espacios ambientales. De similar manera, Araque-Solano et al. (2017) estimaron el precio implícito de la cercanía a vías, parques y centros urbanos, mientras Pabón-Estrella & Burbano-Ojeda (2017) estimaron la influencia de las partículas de menos de 10 micrómetros (PM10) y el ruido en el valor de las viviendas a través de las variables de distancias a vías principales y parques. Por su parte, Rey-Hernandez (2019) estimó la valoración que las personas dan al espacio público, Urrea-Zora & Cárdenas-Mazo (2019), quienes en su trabajo mediante los errores de un modelo de regresión espacial tratan de estimar la calidad de los diferentes barrios de Medellín, Toloza-Delgado (2020) estimó los precios hedónicos para las viviendas nuevas en la ciudad de Bogotá mediante modelos aditivos y Toloza-Delgado et al. (2021) quienes estimaron precios hedónicos para esta ciudad mediante modelos semiparamétricos.

Este trabajo se centra en el estudio de los precios de venta de viviendas usadas en Bogotá y su relación con diferentes variables indexadas en el espacio y la influencia de las observaciones cercanas. El foco del presente trabajo es modelar la dependencia espacial de las observaciones para mejorar las predicciones sobre el precio de venta deseado, teniendo en cuenta los problemas causados por el alto grado de correlación que existe entre las observaciones y los problemas de heterogeneidad de la varianza (la dispersión de los datos no es constante en el espacio) en este tipo de mercado, provocando problemas en los estimadores de los parámetros poblacionales y, por ende, errores en las predicciones que se realicen con éstos.

Para lo anterior, se utilizan técnicas de web scraping para obtener información georreferenciada de los portales web de mercado libre, metro cuadrado y finca raíz. De igual manera, se hace uso de información recolectada por la policía, la alcaldía de Bogotá y otras fuentes secundarias para la construcción de la base de datos final. La extracción de datos de las plataformas web anteriormente listadas se llevó a cabo desde el 17 de Marzo hasta 14 de Mayo del año 2022.

Este trabajo involucra la estimación de un modelo multinivel (Lee et al., 2017), con el cual se busca, mediante los términos aleatorios, involucrar la autocorrelación espacial y la heterogeneidad en la varianza generada por la dependencia espacial. Así mismo, se busca estimar los factores no considerados de valorización de los barrios en Bogotá basándose en la línea metodológica de Dubin (1992) mediante los errores del modelo. En este caso, según el citado autor, el término estocástico del error guarda relación con todas aquellas variables no observadas o que por su naturaleza son difíciles de medir y/o presentan errores en su medición, pero que afectan la valoración de los inmuebles y por ende, su precio deseado, siendo ejemplos de esto, la percepción de los niveles de ruido, de seguridad y de contaminación en la zona, al igual que, cercanía a zonas verdes, negocios, centros comerciales, especulación de precios en las cercanías, valorización del precio por razones sentimentales, entre otras.

Adicionalmente, se realiza la estimación de otros modelos: un primer modelo calculado mediante mínimos cuadrados ordinarios, tres modelos de autocorrelación espacial y un modelo de filtrado espacial. Es de aclarar que las estimaciones de estos modelos tienen como objetivo seleccionar aquel que presente un mejor desempeño y que cumpla con el principio de parsimonia en las estimaciones de los diferentes parámetros y estimación de los factores no considerados de valorización de los inmuebles, siendo estos últimos observados indirectamente a través de factores no controlados que afectan el precio de las viviendas.
La relevancia del presente trabajo recae en el uso de modelos de efectos mixtos para la estimación de los precios hedónicos del mercado de las viviendas, permitiendo modelar la heterogeneidad propia de este sector, así como la dependencia espacial que se da entre observaciones cercanas y que tienen algún grado de correlación entre sí; proporcionando una manera diferente de realizar esta estimación ya que dentro de los trabajos recopilados no se evidencia el uso de estos modelos para la estimación de este tipo de precios en el mercado de las viviendas usadas en Bogotá. De igual manera, mediante la estimación de los modelos se propone una métrica que permite observar y controlar los factores no considerados de valorización de los agregados sociales (ya sea manzanas, UPZ's, etc.), así como establecer la variabilidad de la misma, esto con el objetivo de brindar una herramienta a los entes territoriales para establecer prioridades en términos de obras, subsidios y demás.
El presente documento se divide en cuatro capítulos: En el siguiente capítulo se especificará el marco teórico, brindando definiciones y la teoría sobre la cual se trabaja. En el capítulo tres se describe la metodología de extracción de los datos, la construcción de los modelos y los resultados de las estimaciones. En el cuarto capítulo, se dan las conclusiones y limitaciones del estudio. Finalmente, se hace recomendaciones para trabajos futuros.

2. Marco conceptual

En esta sección se presentan conceptos y definiciones sobre los cuales se fundamenta este trabajo. Estos sirven para entender las características propias de la metodología de los precios hedónicos, datos espaciales y los métodos de estimación usados. Se ofrece al lector una guía teórica de soporte para entender por qué se debe tener en cuenta la dependencia espacial de los datos recolectados para el desarrollo de este trabajo.
La organización que se tiene para este capítulo es la siguiente: Inicialmente se exponen conceptos y definiciones de estadística espacial cruciales para entender el problema desde esta óptica. Posteriormente, se describe la idea central de los precios hedónicos para culminar con la exposición de los modelos usados en este trabajo y cómo éstos trabajan la dependencia espacial encontrada en los datos.

2.1. Estadística espacial

Los datos espaciales juegan un rol cada vez más importante, ya que estos provienen de amplias fuentes y existen dentro de la estadística espacial una gran variedad de metodologías con las cuales abordar problemas que involucran este tipo de datos. Dentro de las principales razones para involucrar la estadística espacial se encuentra la resolución de problemas dentro de procesos generados por los datos, ya que en ocasiones el objetivo de las investigaciones no se centra en la pregunta ¿cuánto? sino ¿cuánto y en dónde? (Bivand et al., 2008; Schabenberger & Gotway, 2004).

2.1.1. Proceso estocástico

Un proceso estocástico es una colección de variables aleatorias indexadas mediante alguna métrica; por ejemplo, una serie de tiempo $Y(t)$ esta indexada por los puntos $t_1, \ldots, t_n$, en los cuales ésta es observada. Adicional a lo anterior, cuando el índice es de una dimensión mayor a uno se denomina campo aleatorio (Schabenberger & Gotway, 2004).

2.1.2. Datos espaciales

Al igual que Schabenberger & Gotway (2004), se seguirá la clasificación utilizada por Cressie (1993) para caracterizar los datos según el dominio espacial. Para esto sea $\mathbf{s} \in R^d$ una

ubicación d-dimensional. Suponiendo que $Z(\mathbf{s})$ en la ubicación espacial $\mathbf{s}$ es una cantidad aleatoria y permitiendo variar este último sobre un conjunto de índices $D \subset R^d$, se tiene que el campo aleatorio generado es:

$$\{Z(\mathbf{s}) : \mathbf{s} \in D \subset R^d\},$$

el cual, dependiendo de las características de D, se puede agrupar como: datos geoespaciales, en los cuales el dominio de D es continuo, pero se considera que el espacio no varía de una realización a otra. Un segundo grupo denominado datos de área, en la cual el dominio de D es discreto, en este caso puede ser infinito, pero es indispensable que sea numerable y se considera que el espacio es fijo entre realizaciones de la variable de interés. Por último, se encuentran los patrones de puntos en el cual D es un conjunto aleatorio al igual que la variable $Z(\mathbf{s})$ (Cressie, 1993; Schabenberger & Gotway, 2004).

Definición 1. En general, un dato espacial es una realización de un proceso estocástico indexado en la ubicación d-dimensional $\mathbf{s}$, que pertenece a un conjunto de índices $D \subset R^d$, por ejemplo cuando los datos son indexados mediante latitud y longitud, el conjunto $D \subseteq R^2$, y si se incluye otra variable de ubicación como altura, se tendría que $D \subseteq R^3$.

2.1.3. Dependencia espacial

Si bien el supuesto de independencia es deseable (ya que cuando se trabaja con esta premisa tanto la teoría como los cálculos computacionales tienden a ser mucho más manejables), no siempre dentro de la naturaleza se encuentran datos donde esta suposición sea razonable, un claro ejemplo son las series de tiempo donde las observaciones sobre un mismo individuo no permiten pensar que estas no se vean afectadas entre sí.
Para ilustrar las consecuencias de la autocorrelación espacial se sigue la misma línea manejada por LeSage & Pace (2009) en su introducción a la dependencia espacial. Lo primero es suponer un proceso de datos de corte transversal no espacial de n observaciones independientes, de tal forma que para $i = 1, \ldots, n$ se tiene que

$$y_i = \mathbf{x_i}^t \boldsymbol{\beta} + \epsilon_i, \text{ con } \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2) \tag{2-1}$$

donde, $\mathbf{x_i}^t = (x_{i1}, x_{i2}, \ldots, x_{ik})$ es un vector de variables independientes observadas en el i-ésimo sujeto cuya dimensión es de $1 \times k$, $\boldsymbol{\beta}$ es el vector de parámetros poblacionales asociados a cada una de las variables independientes con dimensiones $k \times 1$ y y_i es el valor de la variable que se desea explicar mediante un modelo de regresión cuyo término estocástico se distribuye normal, idéntica e independiente, con media cero y varianza σ^2.
En este caso y como consecuencia de la independencia $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$, donde $\mathbf{Y} = (y_1, \ldots, y_n)$ es el vector de la variable dependiente de dimensión $n \times 1$, $\mathbf{X}$ es una matriz de dimensiones $n \times k$, compuesta por un vector de unos en su primera columna y con $k - 1$

variables explicativas, $\boldsymbol{\beta}$ es el vector de los parámetros de dimensiones $k \times 1$, e $\mathbf{I}_n$ es una matriz identidad de dimensiones $n \times n$.
Ahora, supóngase dos observaciones i y j vecinas, de tal manera que el proceso toma las siguientes características:

$$\begin{aligned} y_i &= \alpha_i y_j + \mathbf{x_i}^t \boldsymbol{\beta} + \epsilon_i, \text{ con } \epsilon_i \sim N(0, \sigma^2) \\ y_j &= \alpha_j y_i + \mathbf{x_j}^t \boldsymbol{\beta} + \epsilon_j, \text{ con } \epsilon_j \sim N(0, \sigma^2) \end{aligned} \tag{2-2}$$

Tras realizar el reemplazo de los valores de y_i y y_j en cada igualdad de la ecuación (2-2), y agrupando factores se tiene que

$$\begin{aligned} y_i(1 - \alpha_i \alpha_j) =& (\alpha_i \mathbf{x_j}^t + \mathbf{x_i}^t)\boldsymbol{\beta} + (\alpha_i \epsilon_j + \epsilon_i), \\ y_j(1 - \alpha_i \alpha_j) =& (\alpha_j \mathbf{x_i}^t + \mathbf{x_j}^t)\boldsymbol{\beta} + (\alpha_j \epsilon_i + \epsilon_j), \end{aligned}$$

Este sistema de ecuaciones se puede expresar de manera matricial como

$$\begin{pmatrix} y_i^* \\ y_j^* \end{pmatrix} = \begin{pmatrix} 1 & \alpha_i \\ \alpha_j & 1 \end{pmatrix} \begin{pmatrix} \mathbf{x_i}^t \\ \mathbf{x_j}^t \end{pmatrix} \otimes \boldsymbol{\beta} + \begin{pmatrix} 1 & \alpha_i \\ \alpha_j & 1 \end{pmatrix} \begin{pmatrix} \epsilon_\mathbf{i} \\ \epsilon_\mathbf{j} \end{pmatrix}$$

donde $y_i^* = y_i(1 - \alpha_i \alpha_j)$ y $y_j^* = y_j(1 - \alpha_i \alpha_j)$. Si se toma la matriz de los alfas como $\boldsymbol{\alpha}$, lo anterior se puede ver como $\mathbf{Y}^* = \boldsymbol{\alpha}\mathbf{X} \otimes \boldsymbol{\beta} + \boldsymbol{\alpha}\boldsymbol{\epsilon}$. En esta última forma, $\otimes$ simboliza el producto kronecker. En este caso, el valor esperado es $E[\mathbf{Y}^*|\mathbf{X}] = \boldsymbol{\alpha}\mathbf{X} \otimes \boldsymbol{\beta}$ y la varianza es $Var[\mathbf{Y}^*|\mathbf{X}] = \sigma^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\boldsymbol{t}$, siendo esta última una matriz no diagonal ya que

$$\boldsymbol{\alpha}\boldsymbol{\alpha}^\boldsymbol{t} = \begin{pmatrix} 1 + \alpha_i^2 & \alpha_i + \alpha_j \\ \alpha_i + \alpha_j & 1 + \alpha_j^2 \end{pmatrix}$$

$\mathbf{Y}^*$ se distribuye normal bivariada, y como consecuencia de esto, las observaciones presentan una covarianza que debe ser tenida en cuenta al momento de realizar la estimación de los parámetros. Adicionalmente, es de notar que existe homocedasticidad si y solo si $\alpha_i = \alpha_j$, de lo contrario la dependencia espacial genera también un problema de heterocedasticidad. Por lo tanto,

$$\begin{pmatrix} y_i^* \\ y_j^* \end{pmatrix} \sim N_2(\boldsymbol{\alpha}\mathbf{X} \otimes \boldsymbol{\beta}, \sigma^2 \boldsymbol{\alpha}\boldsymbol{\alpha}^\boldsymbol{t})$$

La generalización de este proceso para n observaciones se puede ver como

$$y_i = \sum_{j \neq i} \alpha_{ij} y_j + \mathbf{X_i}\boldsymbol{\beta} + \epsilon_i, \text{ con } \epsilon_i \sim N(0, \sigma^2) \tag{2-3}$$

en el cual α_{ij} cuantifica la influencia de las demás y_j sobre y_i. La representación matricial de la ecuación (2-3) es

$$\mathbf{Y} = \mathbf{AY} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \text{ con } \boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n), \tag{2-4}$$

donde $\mathbf{A}$ es la matriz de los elementos α_{ij} y todos los elementos de su diagonal son ceros, mientras que $\mathbf{I}_n$ es una matriz identidad de tamaño $n \times n$. Utilizando algebra lineal se llega a que la ecuación (2-4) se puede expresar como

$$\begin{aligned}(\mathbf{I}_n - \mathbf{A})\mathbf{Y} =& \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \mathbf{Y} =& (\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \mathbf{A})^{-1}\boldsymbol{\epsilon},\end{aligned}$$

con

$$\begin{aligned}E[\mathbf{Y}|\mathbf{X}] =& (\mathbf{I}_n - \mathbf{A})^{-1}\mathbf{X}\boldsymbol{\beta} \\ Var[\mathbf{Y}|\mathbf{X}] =& \sigma^2(\mathbf{I}_n - \mathbf{A})^{-1}((\mathbf{I}_n - \mathbf{A})^{-1})^t,\end{aligned}$$

y dado que $Var[\mathbf{Y}|\mathbf{X}]$ no es necesariamente una matriz diagonal, se puede afirmar que $\mathbf{Y}|\mathbf{X}$ se distribuye normal n-variada con la esperanza y varianza mostrados anteriormente. Es de mencionar que, en el caso bivariado, mediante esta generalización se llega exactamente a los mismos resultados mostrados para el modelo de dependencia espacial de la ecuación (2-2).

Definición 2. La dependencia espacial es una situación donde el valor observado de una ubicación i depende de los valores de las observaciones vecinas cercanas (LeSage & Pace, 2009).

2.1.4. Consecuencias de la dependencia espacial

La dependencia espacial genera una situación donde la matriz de varianza no necesariamente es diagonal (no se cumple el supuesto de independencia entre las observaciones) y donde ésta no siempre posee el mismo valor (no se cumple el supuesto de homocedasticidad). Por causa de lo anterior, los estimadores que se obtienen tras mantener estos supuestos en la estimación no son consistentes, es decir que a medida que aumenta el tamaño de la muestra, el estimador no necesariamente se acerca al verdadero valor poblacional. Adicionalmente, se generan problemas en la validez de las pruebas usadas para probar la significancia de los parámetros y en las predicciones realizadas (Schabenberger & Gotway, 2004).

2.1.5. Vecindad y matriz de pesos espaciales

En la estadística espacial se tienen en cuenta procesos con las características del sistema de ecuaciones (2-3), donde el valor de la i-ésima observación se ve influenciada por sus vecinos cercanos o que se encuentran alrededor. En este caso, hay diversas maneras de definir cuando dos datos espaciales son vecinos, aunque de manera formal dos puntos s_i y s_j son vecinos si

la distancia d_{ij} cumple con $0 \leq d_{ij} \leq d^*$, para una distancia máxima de vecindad d^* (Arbia, 2006).
Ahora bien, siguiendo lo propuesto por LeSage & Pace (2009), el primer paso para obtener la matriz de pesos espaciales es definir la matriz **C** de contigüidad, donde los elementos C_{ij} serán 1 si los pares son vecinos y cero en cualquier otro caso. Esta matriz se puede normalizar para que la suma por filas sea igual a uno, es decir, se puede normalizar para volverla una matriz estocástica denominada **W**. Por ejemplo, considere las siguientes matrices tomadas del ejemplo presentado por LeSage & Pace (2009), en donde se tiene que

$$\mathbf{C} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}_{7\times 7} \quad \text{y} \quad \mathbf{W} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}_{7\times 7}$$

Figura 2-1.: Regiones del ejemplo de matriz de contigüidad y W, donde DCN es el distrito central de negocios, tomado de LeSage & Pace (2009, pág. 3)

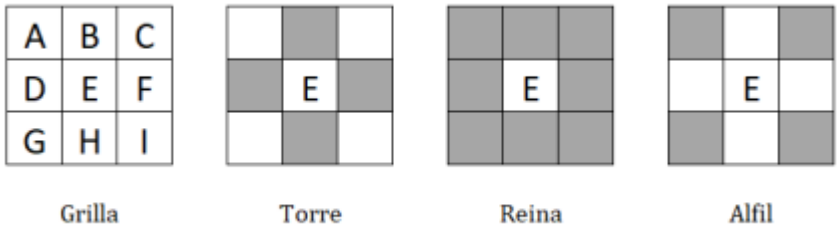

Figura 2-2.: Diferentes tipos de contigüidad tomado de Toloza-Delgado (2020, pág. 5)

En el ejemplo las regiones están organizadas en fila (ver figura **2-1**) por lo cual una región cuenta con máximo dos vecinos, pero es importante resaltar que la contigüidad puede definirse de diferentes maneras tal y como se observa en la figura **2-2**.
La matriz **W** se conoce como la matriz de pesos espaciales, estas son utilizadas en los modelos que se expondrán posteriormente para incluir la dependencia espacial de los datos, ya que esta sirve para ponderar los valores de las diferentes observaciones (LeSage & Pace, 2009). Igualmente es necesario aclarar que existen otros tipos de ponderaciones diferentes a la expuesta anteriormente, por lo que no existe una única manera de modelar la dependencia espacial. Por ejemplo,

$$w_{ij} = \frac{\beta_{ij}^b}{d_{ij}^a} \tag{2-5}$$

es una ponderación espacial donde d_{ij} es la distancia entre los puntos s_i y s_j, β_{ij} es la proporción de frontera que comparten las áreas, y los parámetros a y b sirven para ponderar las distancias y la cantidad de borde compartido, respectivamente (Cliff & Ord, 1970).

Definición 3. Una matriz de contigüidad se define como aquella cuyos elementos son 1 si los pares i y j son vecinos dentro del criterio que se haya seleccionado y cero en caso contrario.

Definición 4. Una matriz de pesos espaciales, **W**, es aquella cuyos elementos sirven para ponderar las diferentes observaciones e incluir la dependencia espacial del proceso estocástico.

Definición 5. Una matriz **W** de tamaño $n \times n$ es estocástica si se cumple que para cualquier fila i

$$\sum_{j=1}^{n} w_{ij} = 1, \tag{2-6}$$

es decir, si la suma de los elementos de una misma fila es igual a una unidad.

2.1.6. Índice de Moran

Al igual que otros supuestos en estadística, la dependencia espacial está sujeta a su validación, esto con el fin de comprobar si ésta debe ser tomada en cuenta o no dentro del contexto del problema que se esté trabajando. Para lo anterior, se utiliza normalmente el índice global de Moran (IM), con el cual se puede cuantificar la dependencia espacial relacionando la variable de interés con su rezago espacial, es decir, mide la correlación que existe entre el valor de una observación específica y el valor de sus vecinos (Pontarollo et al., 2019).
El IM se define como,

$$IM = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}, \tag{2-7}$$

donde n es igual al número de puntos o áreas, w_{ij} es el elemento en la posición (i, j) de la matriz de pesos espaciales W, y_i es el i-ésimo valor de la variable, $\bar{y}$ es el promedio del valor de los n puntos o áreas. Es de notar, que la forma expresada en la ecuación (2-7) se basa en la hipótesis de que la media es constante a través de los puntos, por lo que en general se podría tomar una variable dependiente y_i, con media dependiente de un vector de variables $\mathbf{X_i} = (x_{i1}, \ldots, x_{ik})$ y un vector de parámetros $\boldsymbol{\beta}$ asociadas a dichas variables de la siguiente forma:

$$IM = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \mathbf{x_i}\boldsymbol{\beta})(y_j - \mathbf{x_j}\boldsymbol{\beta})}{\sum_i (y_i - \mathbf{x_i}\boldsymbol{\beta})^2},$$

donde $\mathbf{x_i}$ y $\mathbf{x_j}$ son vectores filas asociados al i-ésima y j-ésima observación respectivamente. En la ecuación anterior, si se reemplaza el vector $\boldsymbol{\beta}$ por su estimación bajo el supuesto de independencia se puede expresar mediante los errores estimados de un modelo de regresión lineal $(\hat{\epsilon}_i)$, tal como se muestra a continuación.

$$\widehat{IM} = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(\hat{\epsilon}_i \hat{\epsilon}_j)}{\sum_i (\hat{\epsilon}_i)^2}, \tag{2-8}$$

Con la ecuación (2-8), se puede hacer inferencia sobre la hipótesis nula $IM = 0$, lo cual implica que el valor de la variable es totalmente aleatorio y no existe dependencia espacial. Para la obtención del valor-p de la prueba estadística se pueden utilizar diferentes métodos de estimación, ya sea mediante la distribución teórica suponiendo que cada observación de la variable dependiente se distribuye normal idéntica e independientemente, o mediante técnicas como las cadenas de Monte Carlo.

Adicional al índice global de Moran, se suele calcular el índice local de Moran para detectar estructuras de dependencias aglomeradas, reconociendo la existencia de estructuras focalizadas relativas a la estructura de dependencia general (Celemín, 2009; Hidalgo-Bucheli, 2020; Pontarollo et al., 2019). Otra manera de visualizar estructuras focalizadas es mediante el diagrama de dispersión de Moran, el cual facilita de manera inicial la observación de la desagregación de la asociación espacial en cuatro cuadrantes (Celemín, 2009).

2.2. Precios hedónicos

Esta teoría es expuesta por Rosen (1974), quien trabajó con aquellos productos cuyo precio es definido por un conjunto de características y la cantidad asociada a cada una de éstas. Suponga que un producto tiene asociado un vector de m características $\mathbf{Z}$=$(z_1, z_2, \ldots, z_m)$, donde z_i es la cantidad asociada a la i-ésima característica y su precio es $P(\mathbf{Z})$, es decir, es una función del vector de características.

Ahora bien, si se plantea una función de utilidad de una persona como $U(x, \mathbf{Z})$ estrictamente cóncava, donde x representa la cantidad de dinero invertido en otros bienes, y se maximiza esta con respecto a la restricción $w = x + P(\mathbf{Z})$, en la cual w simboliza el salario del

individuo, se obtiene que la función objetivo (utilizando multiplicadores de Lagrange para incluir la restricción) es

$$Q = U(x, \mathbf{Z}) - \lambda(w - x - P(\mathbf{Z})), \tag{2-9}$$

la cual, diferenciando, se puede encontrar que:

$$\begin{aligned} \frac{\partial Q}{\partial z_i} =& \frac{\partial U}{\partial z_i} + \lambda \frac{\partial P(\mathbf{Z})}{\partial z_i} \\ \frac{\partial Q}{\partial x} =& \frac{\partial U}{\partial x} + \lambda \end{aligned} \tag{2-10}$$

las cuales igualando a cero para encontrar los valores que optimicen la función Q, se obtiene

$$\frac{\frac{\partial U}{\partial z_i}}{\frac{\partial U}{\partial x}} = \frac{\partial P(\mathbf{Z})}{\partial z_i},$$

es decir, se halla que la tasa marginal de sustitución entre adquirir una unidad adicional de una característica z_i comparada con renunciar cierta cantidad de dinero x, es igual a la derivada de la función de precio con respecto a dicha característica.
De acuerdo a lo anterior, al derivar la función del precio con respecto a alguna característica, se obtiene el precio implícito que la gente está dispuesto a pagar por una unidad adicional de esa característica, es decir,

$$\frac{\partial P(\mathbf{Z})}{\partial z_i} = p_i, \tag{2-11}$$

cuantifica el valor de la i-ésima característica. Este valor es el precio hedónico de dicha característica.
Es de mencionar en este punto que para el mercado de las viviendas es común usar el logaritmo del precio más que el precio absoluto cuando se estiman precios hedónicos, esto con el fin de dar un significado económico a las estimaciones de los precios de las características, siendo los parámetros asociados a los z_i una semielasticidad, esto significa que expresa el cambio porcentual en el precio global ante un cambio absoluto en las cantidades de alguna característica manteniendo todo lo demás constante.

Definicion 6. Para la característica i dentro del vector de características $\mathbf{Z}$ de un producto específico, se define el precio hedónico de dicha característica como el valor implícito que el consumidor está dispuesto a pagar por una unidad adicional de ésta.

2.2.1. Estimación de la calidad de un vecindario y factores no considerados de valorización

Dubin (1992) estima el precio de las viviendas de la ciudad de Baltimore (Estados Unidos) y siguiendo la metodología de precios hedónicos establece el valor implícito de un conjunto de

variables seleccionadas previamente. Sin embargo, para este autor la calidad del vecindario es algo que se puede hallar mediante el termino estocástico del error del modelo, ya que en éste se agrupan variables que no se pueden observar o que son de difícil medición por su costo, así como agrupa errores en las mediciones de variables incluidas en el modelo, repercutiendo esto en el precio verdadero del inmueble vía calidad no controlable (Dubin, 1992; Urrea-Zora & Cárdenas-Mazo, 2019).
Lo anterior expresa la idea de que es posible obtener un estimador de la calidad mediante la estimación del error. Sin embargo, en este trabajo se propone utilizar el término factores no considerados de valorización, puesto que la definición de calidad no es clara ni en el trabajo de Dubin (1992) ni en el trabajo de Urrea-Zora & Cárdenas-Mazo (2019), aunque si se considera correcto que en la parte aleatoria del modelo, se encuentran agrupados estos factores que afectan el valor deseado por el propietario de la vivienda, ya sea porque dentro de estos se encuentran variables no observadas, de difícil medición o porque dentro de estos se agrupan razones emocionales y de especulación, que afectan a éste.

Definición 7. Los factores no considerados de valorización son aquellos que se relacionan con variables no observadas, de difícil medición, con error en su medición, emocionales, de especulación, entre otras, que no son incluidas en el modelo, pero que afectan el precio de venta inicial establecido o deseado por el vendedor. En este caso, la estimación se llevará a cabo mediante los residuos estimados de los modelos, a excepción del modelo mixto, en el cual la estimación del error del intercepto aleatorio se utilizará para la estimación del valor asociado a estos factores.

2.3. Modelos para tratar la dependencia espacial

Dentro de la estadística espacial existen diferentes metodologías y maneras de tratar la dependencia espacial como se expresó al inicio de este capítulo, siendo algunos de estos los modelos de filtrado espacial y los modelos autorregresivos espaciales, aclarando que estos son los que se van a utilizar en este trabajo para encontrar el mejor modelo al trabajar los datos de las viviendas usadas en Bogotá, no siendo los únicos que existen, pero si aquellos, que junto con el uso de modelos de efectos mixtos, se utilizaran para tratar la autocorrelación y la heterogeneidad espacial del precio inicial deseado por el vendedor.

2.3.1. Modelo jerárquico

Los modelos que incluyen efectos aleatorios son utilizados para resolver diversos problemas, entre los cuales se encuentran los asociados a estimaciones de áreas pequeñas, experimentos de medidas repetidas o cuando se cree que la variable respuesta depende de varios factores de los cuales no todos pueden ser medidos o conocidos (Lee et al., 2017).

En este caso, el ejercicio de estimación que se realiza toma en cuenta dos maneras de utilizar los modelos de efectos mixtos. Por un lado, al contar con varias observaciones de un mismo barrio (observaciones repetidas), es posible modelar el precio de venta deseado de una vivienda usada en la ciudad de Bogotá teniendo en cuenta factores asociados a los barrios que permitan modelar la heterogeneidad en este mercado, así como absorber la dependencia espacial que pueda haber en los datos.
Por otro lado, se puede tomar efectos aleatorios asociados a las coordenadas, con el fin de modelar la dependencia espacial como un efecto el cual no se puede medir o conocer concretamente. Es de resaltar que, aunque son dos enfoques diferentes, la manera en que se resuelve es idéntica, por lo cual, se procede a dar una explicación general, mostrando posteriormente la manera en que se plantea este modelo desde la visión de la teoría estadística espacial.
Lo primero es plantear un modelo como el que se muestra a continuación:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon}, \tag{2-12}$$

donde $\mathbf{X}$ es una matriz diseño asociada a los efectos fijos del modelo, $\boldsymbol{\beta}$ es un vector de parámetros asociados a estos mismos efectos, $\mathbf{Z}$ es una matriz diseño asociada a los efectos aleatorios del modelo y $\mathbf{U}$ es un vector aleatorio.
Adicionalmente,

$$\boldsymbol{\epsilon} \sim NMV(\mathbf{0}, \mathbf{D}) \;\text{ y }\; \mathbf{U} \sim NMV(\mathbf{0}, \boldsymbol{\Sigma})$$

por lo cual,

$$\begin{aligned} \mathbf{Y}|\mathbf{X} \sim & NMV(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}) \quad \text{donde,} \\ \mathbf{V} = & \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^t + \mathbf{D} \end{aligned}$$

Siguiendo a Lee et al. (2017), quienes plantean un primer paso para la estimación del modelo a través de la función de probabilidad conjunta de $\mathbf{Y}$ y $\mathbf{U}$, la cual se encuentra como

$$\begin{aligned} f_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{U}) =& f_{\boldsymbol{\theta}}(\mathbf{Y}|\mathbf{U}) f_{\boldsymbol{\theta}}(\mathbf{U}) \\ =& \left|2\pi\mathbf{D}\right|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{U})^t \mathbf{D}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{U})\right\} \\ & \left|2\pi\boldsymbol{\Sigma}\right|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}\mathbf{U}^t\boldsymbol{\Sigma}^{-1}\mathbf{U}\right\} \\ =& \left|2\pi\mathbf{D}\right|^{\frac{-1}{2}} \left|2\pi\boldsymbol{\Sigma}\right|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{U})^t \mathbf{D}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{U}) - \frac{1}{2}\mathbf{U}^t\boldsymbol{\Sigma}^{-1}\mathbf{U}\right\} \\ =& L(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y}), \end{aligned}$$

esto concuerda con la verosimilitud conjunta de los efectos aleatorios $\mathbf{U}$ y parámetros $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$, donde $\boldsymbol{\tau}$ es el vector de los parámetros de dispersión asociados a los elementos de las matrices $\mathbf{D}$ y $\boldsymbol{\Sigma}$.

Al trabajar con la log-verosimilitud ($\ln[L(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y})] = l(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y})$) y estimar la matriz de información observada de Fisher de los efectos aleatorios, se tiene que

$$-\frac{\partial^2 l(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y})}{\partial \mathbf{U}^t \partial \mathbf{U}} = \mathbf{Z}^t \mathbf{D}^{-1} \mathbf{Z} + \boldsymbol{\Sigma}^{-\mathbf{1}} = I_{\mathbf{U}}(\boldsymbol{\tau}),$$

lo que muestra que los efectos aleatorios no contienen información de los parámetros $\boldsymbol{\beta}$, y por ende, se puede hacer estimación simultánea de estos dos con la log-verosimilitud conjunta, ya que como muestran Lee et al. (2017), $\mathbf{U}$ es canónica para $\boldsymbol{\beta}$. Sin embargo, la matriz muestra dependencia de los parámetros de dispersión $\boldsymbol{\tau}$, por lo que para realizar la estimación sobre estos últimos es necesario usar un log-verosimilitud marginal $l(\boldsymbol{\tau})$.
En este caso, partiendo de la distribución conjunta $L(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y}, \mathbf{U})$, e integrando con respecto a $\mathbf{U}$ se obtiene $L(\boldsymbol{\beta}, \boldsymbol{\tau})$. Sin embargo, antes de evaluar la integral, se observa cuáles son las condiciones de primer orden para optimizar conjuntamente la log-verosimilitud de los efectos aleatorios $\mathbf{U}$ y los parámetros $\boldsymbol{\beta}$ para un $\boldsymbol{\tau}$ fijo. Esto implica que la cantidad a maximizar es:

$$\begin{aligned}
r =& -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{U})^t \mathbf{D}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\mathbf{U}) - \frac{1}{2}\mathbf{U}^t \boldsymbol{\Sigma}^{-1}\mathbf{U} \\
=& -\frac{1}{2}\Big(\mathbf{Y}^t\mathbf{D}^{-1}\mathbf{Y} - \mathbf{Y}^t\mathbf{D}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{Y}^t\mathbf{D}^{-1}\mathbf{Z}\mathbf{U} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{D}^{-1}\mathbf{Y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{D}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{D}^{-1}\mathbf{Z}\mathbf{U} \\
&- \mathbf{U}^t\mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Y} + \mathbf{U}^t\mathbf{Z}^t\mathbf{D}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{U}^t\mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Z}\mathbf{U} + \mathbf{U}^t\boldsymbol{\Sigma}^{-1}\mathbf{U}\Big) \\
=& -\frac{1}{2}\Big(\mathbf{Y}^t\mathbf{D}^{-1}\mathbf{Y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{D}^{-1}\mathbf{Y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{D}^{-1}\mathbf{X}\boldsymbol{\beta} - 2\mathbf{U}^t\mathbf{Z}^t\mathbf{D}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
&+ \mathbf{U}^t\Big(\mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Z} + \boldsymbol{\Sigma}^{-1}\Big)\mathbf{U}\Big)
\end{aligned}$$

la cual si se deriva parcialmente con respecto a $\mathbf{U}$ y $\boldsymbol{\beta}$ llegando a

$$\begin{aligned}
\frac{\partial r}{\partial \mathbf{U}} &= -\Big(\mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Z} + \boldsymbol{\Sigma}^{-1}\Big)\mathbf{U} + \mathbf{Z}^t\mathbf{D}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
\frac{\partial r}{\partial \boldsymbol{\beta}} &= \mathbf{X}^t\mathbf{D}^{-1}\mathbf{Y} - \mathbf{X}^t\mathbf{D}^{-1}\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^t\mathbf{D}^{-1}\mathbf{Z}\mathbf{U}
\end{aligned}$$

igualando a cero para encontrar las ecuaciones normales se tiene

$$\begin{aligned}
\Big(\mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Z} + \boldsymbol{\Sigma}^{-1}\Big)\mathbf{U} =& \mathbf{Z}^t\mathbf{D}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
(\mathbf{X}^t\mathbf{D}^{-1}\mathbf{X})\boldsymbol{\beta} =& \mathbf{X}^t\mathbf{D}^{-1}\mathbf{Y} - \mathbf{X}^t\mathbf{D}^{-1}\mathbf{Z}\mathbf{U}
\end{aligned} \tag{2-13}$$

Suponiendo que los valores de $\boldsymbol{\beta}$ al igual que los valores del vector $\boldsymbol{\tau}$ son fijos, se obtiene el vector $\hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}$ que máximiza la log-verosimilitud con lo obtenido en la primera igualdad de la ecuación (2-13). En este punto, se realiza una expansión de taylor de segundo orden alrededor de este vector, con lo cual:

$$
\begin{aligned}
l(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y}) \approx & l(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) + \frac{\partial l(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y})}{\partial \mathbf{U}}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}) + \frac{1}{2}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})^t \\
& \frac{\partial^2 l(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y})}{\partial \mathbf{U}^t \partial \mathbf{U}}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})
\end{aligned}
$$

Dado que $\hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}$ es el punto en que el gradiente de la log-verosimilitud es cero, el segundo término desaparece. De igual manera, el último término coincide con el negativo de la matriz de información de Fisher de los efectos aleatorios, por lo tanto:

$$
l(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y}) \approx l(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) - \frac{1}{2}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})^t I_{\mathbf{U}}(\boldsymbol{\tau})(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})
$$

Aplicando exponencial para obtener la función de verosimilitud, y por ende la distribución conjunta, se tiene que

$$
L(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y}) \approx L(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) \exp\Big\{ -\frac{1}{2}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})^t I_{\mathbf{U}}(\boldsymbol{\tau})(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})\Big\}
$$

Por lo cual, la integral a evaluar se torna en

$$
\begin{aligned}
\int_{-\infty}^{\infty} L(\boldsymbol{\theta}, \mathbf{U}; \mathbf{Y}) d\mathbf{U} \approx & L(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) \int_{-\infty}^{\infty} \exp\Big\{\frac{-1}{2}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})^t I_{\mathbf{U}}(\boldsymbol{\tau})(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})\Big\} d\mathbf{U} \\
\approx & L(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) \left|\frac{2\pi}{I_{\mathbf{U}}(\boldsymbol{\tau})}\right|^{\frac{1}{2}} \int_{-\infty}^{\infty} \left|\frac{I_{\mathbf{U}}(\boldsymbol{\tau})}{2\pi}\right|^{\frac{1}{2}} \exp\Big\{\frac{-1}{2}(\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})^t I_{\mathbf{U}}(\boldsymbol{\tau}) \\
& (\mathbf{U} - \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}})\Big\} d\mathbf{U} \\
\approx & L(\boldsymbol{\theta}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) \left|\frac{2\pi}{I_{\mathbf{U}}(\boldsymbol{\tau})}\right|^{\frac{1}{2}} \approx L(\boldsymbol{\beta}, \boldsymbol{\tau})
\end{aligned}
$$

Haciendo uso del mismo truco para la integral, teniendo en cuenta que $I_{\mathbf{U}}(\boldsymbol{\tau})$ no depende de $\boldsymbol{\beta}$, se tiene que la log-verosimilitud ajustada para los parámetros $\boldsymbol{\tau}$ usando la aproximación de Laplace para la integral es

$$
l_{\boldsymbol{\beta},\mathbf{U}}(\boldsymbol{\tau}) = l(\hat{\boldsymbol{\beta}}, \boldsymbol{\tau}, \hat{\mathbf{U}}_{\boldsymbol{\beta},\boldsymbol{\tau}}; \mathbf{Y}) - \frac{1}{2}\ln\left(\left|\frac{I_{\mathbf{U}}(\boldsymbol{\tau})}{2\pi}\right|\right) - \frac{1}{2}\ln\left(\left|\frac{I_{\boldsymbol{\beta}}(\boldsymbol{\tau})}{2\pi}\right|\right) \approx l(\boldsymbol{\tau}),
$$

con la cual se pueden obtener los parámetros de dispersión.
Ahora bien, para la estimación de todos los parámetros, Lee et al. (2017) proponen los siguientes pasos:

1. Tomar unos valores iniciales de los parámetros de dispersión $\hat{\boldsymbol{\tau}}$.

2. Con los valores anteriores, utilizar la log-verosimilitud conjunta para encontrar las estimaciones para $\hat{\boldsymbol{\beta}}$ y $\hat{\mathbf{U}}$.

3. Actualizar el valor de los parámetros de dispersión mediante la log-verosimilitud ajustada para los parámetros $\boldsymbol{\tau}$, obteniendo la estimación por máxima verosimilitud restringida (REML por sus siglas en inglés) de estos parámetros.

4. Iterar pasos 2 y 3 hasta converger a un resultado.

Una vez obtenidos los resultados, usando la ecuación (2-13) se puede obtener el sistema de ecuaciones de Henderson para un modelo mixto (Schabenberger & Gotway, 2004).

$$\begin{pmatrix} (\mathbf{X}^t\mathbf{D}^{-1}\mathbf{X}) & \mathbf{X}^t\mathbf{D}^{-1}\mathbf{Z} \\ \mathbf{Z}^t\mathbf{D}^{-1}\mathbf{X} & \mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Z}+\boldsymbol{\Sigma}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{U}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^t\mathbf{D}^{-1}\mathbf{Y} \\ \mathbf{Z}^t\mathbf{D}^{-1}\mathbf{Y} \end{pmatrix}, \tag{2-14}$$

Mediante la inversión de la matriz del lado izquierdo de (2-14) se tiene las estimaciones de la varianza de $\hat{\boldsymbol{\beta}}$ y del sesgo de $\hat{\mathbf{U}}-\mathbf{U}$. Por su parte, los errores estándar de los parámetros de dispersión se pueden hallar con la Hessiana de la log-verosimilitud ajustada.
Es de notar, que este proceso no se basa en una forma específica de las matrices $\mathbf{D}$ y $\boldsymbol{\Sigma}$; sin embargo, y siguiendo a Schabenberger & Gotway (2004), a menudo es conveniente describir un proceso espacial $\mathbf{Z}(\mathbf{s})$ de tal manera que cumpla con

$$\begin{aligned} E[\mathbf{Z}(\mathbf{s})] =& \mathbf{X}(\mathbf{s})\boldsymbol{\beta} \\ Var[\mathbf{Z}(\mathbf{s})] =& \sigma_\epsilon^2\mathbf{I}_n + \boldsymbol{\Sigma}(\boldsymbol{\theta}), \end{aligned} \tag{2-15}$$

en el cual σ_ϵ^2 representa la varianza del error estocástico (lo cual implica que $\mathbf{D}=\sigma_\epsilon^2\mathbf{I}_n$), y $\boldsymbol{\theta}$ incluye componentes de varianza así como de autocorrelación espacial, es decir, es un vector de parámetros que determinan las entradas de la matriz $\boldsymbol{\Sigma}$. Sin embargo, como hace notar el mismo autor, la dificultad de estimar simultáneamente estos parámetros lleva a proponer un modelo en el cual se incluye un proceso espacialmente variante, $\mathbf{S}(\mathbf{s})$, de tal manera que

$$\begin{aligned} E[\mathbf{Z}(\mathbf{s})|\mathbf{S}(\mathbf{s})] =& \mathbf{S}(\mathbf{s}) \\ Var[\mathbf{Z}(\mathbf{s})|\mathbf{S}(\mathbf{s})] =& \sigma_\epsilon^2\mathbf{I}_n \\ Cov[\mathbf{Z}(\mathbf{s_i}),\mathbf{Z}(\mathbf{s_j})|\mathbf{S}(\mathbf{s})] =& 0 \\ \mathbf{S}(\mathbf{s}) =& \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \mathbf{v}(\mathbf{s}), \end{aligned} \tag{2-16}$$

donde $\mathbf{v}(\mathbf{s})$ es un proceso estacionario de segundo orden con media cero y con matriz de varianza-covarianza, $\boldsymbol{\Sigma}(\boldsymbol{\theta_s})$, cuyas entradas dependen de unos parámetros asociados al proceso $\mathbf{v}(\mathbf{s})$ y de la distancia entre dos puntos. Con estas especificaciones, se tiene por ley de las esperanzas iteradas y varianza total que

$$\begin{aligned} E[\mathbf{Z}(\mathbf{s})] =& E[E[\mathbf{Z}(\mathbf{s})|\mathbf{S}]] = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} \\ Var[\mathbf{Z}(\mathbf{s})] =& E[Var[\mathbf{Z}(\mathbf{s})|\mathbf{S}]] + Var[E[\mathbf{Z}(\mathbf{s})|\mathbf{S}]] \\ =& \sigma_\epsilon^2\mathbf{I}_n + \boldsymbol{\Sigma}(\boldsymbol{\theta}_\mathbf{s}) = \mathbf{V}. \end{aligned} \tag{2-17}$$

Ahora bien, para la estimación se sigue que los efectos aleatorios $\mathbf{v}(\mathbf{s}) = \boldsymbol{\alpha}(\boldsymbol{s})\mathbf{U}$, donde $\boldsymbol{\alpha}(\boldsymbol{s})$, es una matriz diseño asociada a estos efectos y $\mathbf{U}$ es el vector de errores aleatorios con media cero y matriz de varianza-covarianza $\mathbf{G}(\boldsymbol{\theta}_s)$, con lo cual, $\boldsymbol{\Sigma}(\boldsymbol{\theta}_\mathbf{s}) = \boldsymbol{\alpha}(\boldsymbol{s})\mathbf{G}(\boldsymbol{\theta}_s)\boldsymbol{\alpha}(\boldsymbol{s})^t$. Por ello, se propone entonces el modelo,

$$\mathbf{Z}(\mathbf{s}) = \mathbf{X}(\mathbf{s})\boldsymbol{\beta} + \boldsymbol{\alpha}(\boldsymbol{s})\mathbf{U} + \boldsymbol{\epsilon},$$

lo cual lleva al sistema de ecuaciones de Henderson para un modelo mixto (Schabenberger & Gotway, 2004) y lleva a las mismas soluciones que las ecuaciones mostradas en (2-13).

2.3.2. Modelos autorregresivos

Dentro de la estadística espacial, se busca homologar la metodología aplicada a las series de tiempo, para lo cual se toma la idea de que para un momento t específico, su valor es determinado o está relacionado con una combinación lineal de sus valores pasados, siendo traducido esto para un punto específico, $\mathbf{s}$, el cual depende de una combinación lineal de sus "vecinos". Con lo anterior, los modelos autorregresivos introducen la estructura de dependencia espacial en los datos a través de la proximidad espacial que determina la vecindad de los datos (Schabenberger & Gotway, 2004).

Modelo autorregresivo espacial

El modelo autorregresivo espacial (SAR por sus siglas en inglés), es aquel que cuenta con un proceso generador de datos con las siguientes características, tal como lo expresan LeSage & Pace (2009):

$$\begin{aligned}\mathbf{Y} =& \rho\mathbf{W}\mathbf{Y} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\\ \mathbf{Y} =& (\mathbf{I}_n - \rho\mathbf{W})^{-1}\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}, \quad \text{con} \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n),\end{aligned} \tag{2-18}$$

donde $\mathbf{Y}$ es un vector de n observaciones indexadas en el espacio, las cuales se relacionan entre ellas por medio de la matriz de pesos espaciales $\mathbf{W}$ y un factor de correlación ρ, encargado de medir la asociación entre una observación y sus vecinos. Como se mostró anteriormente con la ecuación (2-4), reemplazando la matriz $\mathbf{A}$ por $\rho\mathbf{W}$ se llega a la forma que toma tanto el valor esperado como la varianza del proceso.
Siguiendo lo expuesto por LeSage & Pace (2009), la log-verosimilitud se puede expresar como:

$$l = C + \ln(|\mathbf{I_n} - \rho\mathbf{W}|) - \frac{n}{2}\ln(S(p)) \tag{2-19}$$

donde

$$S(p) = \mathbf{e}(p)^t\mathbf{e}(p) = \mathbf{e}_0^t\mathbf{e}_0 - 2p\mathbf{e}_0^t\mathbf{e}_d + p^2\mathbf{e}_d^t\mathbf{e}_d$$

con

$$\begin{aligned}
\mathbf{e}(p) =& \mathbf{e}_0 - p\mathbf{e}_d \\
\mathbf{e}_0 =& \mathbf{WY} - \mathbf{Z}\boldsymbol{\delta}_0 \\
\mathbf{e}_d =& \mathbf{WY} - \mathbf{Z}\boldsymbol{\delta}_d \\
\boldsymbol{\delta}_0 =& (\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{Y} \\
\boldsymbol{\delta}_d =& (\mathbf{Z}^t\mathbf{Z})^{-1}\mathbf{Z}^t\mathbf{WY},
\end{aligned}$$

donde C es una constante que no influye en la estimación de los parámetros, $|\ \ |$ es el determinante de la matriz, $\mathbf{Z} = [\mathbf{1}_n, \mathbf{X}_r]$, es decir, una matriz que presenta de forma particionada la matriz diseño original $\mathbf{X}$, compuesta por un vector de unos de tamaño n y las demás variables usadas en la regresión contenidas en la matriz $\mathbf{X}$ diferentes al intercepto, y $\boldsymbol{\delta} = [\alpha, \boldsymbol{\beta}]$ un vector que contiene la estimación del parámetro asociado al intercepto y las variables $\mathbf{X}_r$. En este caso, LeSage & Pace (2009) propone:

1. Crear una grilla con valores para ρ que varían entre $[\text{mín}(\boldsymbol{\omega})^{-1}, \text{máx}(\boldsymbol{\omega})^{-1}]$, donde $\boldsymbol{\omega}$ es un vector de los eigen-valores de $\mathbf{W}$.

2. Evaluar los valores de la grilla en la función de log-verosimilitud para hallar el valor óptimo $\hat{\rho}$.

3. Con el valor de $\hat{\rho}$, LeSage & Pace (2009) proponen los valores estimados para los parámetros del modelo δ, así como para los parámetros de varianza y covarianzas como:

$$\begin{aligned}
\hat{\boldsymbol{\delta}} =& \boldsymbol{\delta}_0 - \hat{\rho}\boldsymbol{\delta}_d \\
\hat{\sigma}^2 =& n^{-1}S(\hat{p}) \\
\hat{\boldsymbol{\Sigma}} =& \hat{\sigma}^2[(\mathbf{I}_n - \hat{\rho}\mathbf{W})^t(\mathbf{I}_n - \hat{\rho}\mathbf{W})]^{-1}
\end{aligned}$$

A la par que se realizan las estimaciones, también se estiman los efectos directos e indirectos de las observaciones, siendo el primero cuánto influye un cambio en la r-ésima variable sobre la observación i, mientras el segundo captura el impacto sobre la observación i al cambiar la r-ésima variable sobre las otras j observaciones (LeSage & Pace, 2009).
También se resalta que existen formas no lineales de este modelo no abarcadas en este trabajo, pero que llegan a ser una generalización bastante útil cuando se quiere comprobar la existencia de relaciones no lineales entre variables independientes con la variable dependiente, siendo su principal limitante la falta de interpretación económica de los términos (Toloza-Delgado et al., 2021).

Modelo autorregresivo espacial del error

El modelo autorregresivo espacial del error (SEM por sus siglas en inglés), es aquel que como señala LeSage & Pace (2009), tiene un proceso generador de datos caracterizado por:

$$\begin{aligned}\mathbf{Y} =& \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ \mathbf{u} =& \lambda \mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}\end{aligned} \tag{2-20}$$

que se puede expresar como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + (\mathbf{I}_n - \lambda\mathbf{W})^{-1}\epsilon, \quad \text{con} \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n)$$

En este caso, el término de rezago espacial, λ, cuantifica la relación entre el valor de la i-ésima observación, con una combinación lineal de un término estocástico que depende de la matriz de pesos espaciales, $\mathbf{W}$. Al igual que en el modelo SAR, se parte de una matriz diseño, $\mathbf{X}$, que tiene asociado un vector de coeficientes $\boldsymbol{\beta}$. Es posible ver que el vector $\mathbf{Y}|\mathbf{X}$ es un vector aleatorio con

$$\begin{aligned}E[\mathbf{Y}|\mathbf{X}] =& \mathbf{X}\beta \\ Var[\mathbf{Y}|\mathbf{X}] =& \sigma^2(\mathbf{I}_n - \lambda\mathbf{W})^{-1}((\mathbf{I}_n - \lambda\mathbf{W})^{-1})^t,\end{aligned}$$

el cual, siguiendo nuevamente a LeSage & Pace (2009), tiene una log-verosimilitud concentrada con respecto a $\boldsymbol{\beta}$ y σ^2 como función de λ de tal manera que se puede expresar

$$l = C + \ln(|\mathbf{I}_n - \lambda\mathbf{W}|) - \frac{n}{2}\ln(S(\lambda)) \tag{2-21}$$

donde

$$\begin{aligned}S(\lambda) =& \mathbf{e}(\lambda)^t\mathbf{e}(\lambda) = [\mathbf{Y}(\lambda) - \mathbf{X}(\lambda)\boldsymbol{\beta}(\lambda)]^t[\mathbf{Y}(\lambda) - \mathbf{X}(\lambda)\boldsymbol{\beta}(\lambda)] \\ \boldsymbol{\beta}(\lambda) =& [\mathbf{X}(\lambda)^{\mathbf{t}}\mathbf{X}(\lambda)]^{-1}\mathbf{X}(\lambda)^{\mathbf{t}}\mathbf{Y}(\lambda)\end{aligned}$$

con

$$\begin{aligned}\mathbf{X}(\lambda) =& \mathbf{X} - \lambda\mathbf{W}\mathbf{X}, \\ \mathbf{Y}(\lambda) =& \mathbf{Y} - \lambda\mathbf{W}\mathbf{Y} \quad \text{y}, \\ \sigma^2(\lambda) =& \frac{\mathbf{e}(\lambda)^t\mathbf{e}(\lambda)}{n},\end{aligned}$$

siendo a diferencia de lo que sucede con el modelo SAR, una estimación complicada por medio de una grilla de valores para λ, puesto que requiere la manipulación de matrices

de tamaños de $n \times 1$ y $n \times k$ por cada λ elegido. Por lo anterior, LeSage & Pace (2009) propusieron trabajar con precomputación para calcular el término $S(\lambda)$ en torno a matrices de momentos con dimensiones $k \times k$ o menores, siendo estas matrices:

$$\begin{aligned}
\mathbf{A_{xx}}(\lambda) =& \mathbf{X}^t\mathbf{X} - \lambda\mathbf{X}^t\mathbf{W}\mathbf{X} - \lambda\mathbf{X}^t\mathbf{W}^t\mathbf{X} + \lambda^2\mathbf{X}^t\mathbf{W}^t\mathbf{W}\mathbf{X} \\
\mathbf{A_{xy}}(\lambda) =& \mathbf{X}^t\mathbf{Y} - \lambda\mathbf{X}^t\mathbf{W}\mathbf{Y} - \lambda\mathbf{X}^t\mathbf{W}^t\mathbf{Y} + \lambda^2\mathbf{X}^t\mathbf{W}^t\mathbf{W}\mathbf{Y} \\
\mathbf{A_{yy}}(\lambda) =& \mathbf{X}^t\mathbf{Y} - \lambda\mathbf{Y}^t\mathbf{W}\mathbf{Y} - \lambda\mathbf{Y}^t\mathbf{W}^t\mathbf{Y} + \lambda^2\mathbf{Y}^t\mathbf{W}^t\mathbf{W}\mathbf{Y} \\
\boldsymbol{\beta}(\lambda) =& [\mathbf{A_{xx}}(\lambda)]^{-1}\mathbf{A_{xy}}(\lambda) \\
S(\lambda) =& \mathbf{A_{yy}}(\lambda) - \boldsymbol{\beta}(\lambda)^t\mathbf{A_{xx}}(\lambda)\boldsymbol{\beta}(\lambda),
\end{aligned} \tag{2-22}$$

con las cuales se encuentra $\hat{\lambda}$. Al igual que en el modelo SAR, Tras encontrar este valor, las estimaciones de los parámetros se obtienen mediante:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} =& \boldsymbol{\beta}(\hat{\lambda}) \\
\hat{\sigma}^2 =& n^{-1}S(\hat{\lambda}) \\
\hat{\boldsymbol{\Sigma}} =& \hat{\sigma}^2[(\mathbf{I}_n - \lambda\mathbf{W})^t(\mathbf{I}_n - \lambda\mathbf{W})]^{-1}.
\end{aligned}$$

Modelo SARAR

Dentro de los modelos autorregresivos también es posible encontrar modelos donde se cuenta con rezago tanto en la variable dependiente, así como en los errores, este modelo se conoce como SARAR, aunque es posible encontrarlo como modelo SAC (LeSage & Pace, 2009). Este modelo se caracteriza por:

$$\begin{aligned}
\mathbf{Y} =& \rho\mathbf{W}_1 Y + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\
\mathbf{u} =& \lambda\mathbf{W}_2\mathbf{u} + \boldsymbol{\epsilon}, \text{ con } \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}_n),
\end{aligned} \tag{2-23}$$

donde $\mathbf{W}_1$ y $\mathbf{W}_2$ pueden ser iguales o distintas. Se puede observar que, en este caso, si $\rho = 0$ entonces el modelo se vuelve un modelo SEM, mientras que si $\lambda = 0$, el modelo se vuelve un modelo SAR, siendo un modelo lineal normal si $\rho = 0$ y $\lambda = 0$. En este caso la log-verosimilitud del modelo, siguiendo a LeSage & Pace (2009) queda definida como:

$$l = -\frac{n}{2}\ln(\pi\sigma^2) + \ln(|\mathbf{A}|) + \ln(|\mathbf{B}|) - \frac{\mathbf{e}^t\mathbf{e}}{2\sigma^2} \tag{2-24}$$

donde

$$\begin{aligned}
\mathbf{e} =& \mathbf{B}(\mathbf{A}\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\
\mathbf{A} =& \mathbf{I}_n - \rho\mathbf{W}_1 \\
\mathbf{B} =& \mathbf{I}_n - \lambda\mathbf{W}_2,
\end{aligned}$$

el cual se puede concentrar con respecto a los parámetros $\boldsymbol{\beta}$ y σ^2 y se resuelve una optimización bivariada para los parámetros λ y ρ. En este caso, LeSage & Pace (2009) afirma que, dados los avances computacionales de la época, conseguir la optimización de este problema no es extremadamente difícil. Adicionalmente, al igual que en modelo SAR, se pueden calcular efectos de tipo directo e indirecto.

2.3.3. Modelo con filtrado espacial

En este trabajo se utilizará también la técnica de filtrado espacial. En esta metodología, se realiza una regresión en la cual se separa la componente sistemática, conformada por todas aquellas variables regresoras observadas y cuyo efecto sobre la variable dependiente se busca entender mediante los coeficientes estimados, y la componente espacial, la cual está asociada con una matriz de pesos espaciales, $\mathbf{W}$, que define la relación entre observaciones "vecinas", introduciendo estos efectos en la regresión mediante variables sintéticas no explicativas (Buendía-Azorín & Sanchez, 2013; Pontarollo et al., 2019).
Las variables sintéticas que se integran a la regresión se encuentran mediante los vectores propios de $\mathbf{MWM}$, donde $\mathbf{M} = (\mathbf{I}_n - \frac{1}{n}\mathbf{1_n 1_n^t})$, es una matriz utilizada para centrar los datos, mientras la expresión completa ($\mathbf{MWM}$) se relaciona con la sumatoria de la variable espacial, ponderada por los elementos de la matriz de pesos espaciales, la cual hace parte del denominador en el índice de Moran I. Por lo tanto, la regresión a estimar tiene la siguiente forma:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}\boldsymbol{\beta} + \sum_{i=1}^{g} \mathbf{E}_i + \boldsymbol{\epsilon}, \text{ con } \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2\mathbf{I}) \tag{2-25}$$

donde

$$(\mathbf{MWM})\mathbf{E}_i = a\mathbf{E}_i, \quad \text{con} \quad a \in R,$$

es decir, se toman g vectores propios, seleccionando aquellos cuyo valor propio a sea diferente de cero. La cantidad de vectores propios, $\mathbf{E}_i$, es igual al tamaño de la matriz ($\mathbf{MWM}$), la cual es de tamaño $n \times n$.

3. Datos y modelos

En esta sección se procede a describir la manera en que se realizó la extracción de los datos, especificando la manera en que se llevó a cabo el web scraping para la recolección de los mismos de los portales web mencionados en la introducción, con el fin de que pueda servir de guía para futuros trabajos. Adicionalmente, se procede a explicar la metodología para llegar a la muestra final que se utiliza en el presente trabajo, para posteriormente, mostrar los modelos estimados junto con los supuestos que conllevan cada uno de ellos.

3.1. Web scraping

La extracción de los datos de los anuncios de ventas de viviendas usadas se llevó a cabo mediante web scraping. En este caso, se desarrolló un programa en Python, haciendo uso de la librería "selenium" y la herramienta para automatización de tareas del navegador Firefox (lo que se conoce como el "web driver" del navegador), accediendo a la información a través de éste.

Para la extracción de los datos de los portales, primero se realiza una búsqueda manual de viviendas usadas en Bogotá dentro de cada uno de ellos. Adicionalmente, se filtraron para obtener resultados solo de casas o apartamentos usados, obteniendo un total de 42 páginas por portal, a excepción de metro cuadrado, debido a que los portales de mercado libre y finca raíz limitan los resultados a este número de páginas. Este link fue almacenado en una variable que se utilizó posteriormente para llevar la sesión del navegador web, por medio de "selenium", hacia la fuente de los datos.

Con la página web dentro del navegador iniciado con Python, se procedió mediante la librería "bs4", a extraer la información disponible en la previsualización de los anuncios, así como el link del anuncio. Para recorrer todos los resultados, se usó un bucle para recorrer todas las páginas. Posteriormente, a través de otro bucle, se recorrió y se extrajo información de todos los anuncios disponible en los portales.

Para obtener los datos geolocalizados, se utilizó la ubicación (dirección) extraída, con la cual y usando la librería "geocoder", y su vínculo con open street map, se obtuvieron los datos de longitud y latitud de cada una de las viviendas (ver anexo B).

3.2. Datos y variables

En total, entre los portales de mercado libre, finca raíz y metro cuadrado se recolectaron datos de 31447 viviendas. La base de datos está conformada por la siguiente lista de variables:

- Precio: Variable cuantitativa expresada en pesos colombianos que indica el valor solicitado por la vivienda.
- Tipo: Variable categórica nominal que expresa si la vivienda es casa o apartamento.
- Metros cuadrados: Variable cuantitativa que indica los metros cuadrados de la vivienda.
- Habitaciones: Variable cuantitativa que indica el número de habitaciones con las que cuenta la vivienda.
- Ubicación: Variable nominal en la que se almacena la dirección del inmueble.
- Link: Link de la publicación, se aclara que algunos de estos pueden haber sido eliminados ya sea por la venta del inmueble o porque el tiempo de publicación ya se cumplió.
- Baños: Variable cuantitativa que expresa el número de baños con los que cuenta el inmueble.
- Parqueaderos: Variable cuantitativa que indica el número de parqueaderos que tiene la vivienda.
- Estrato: Variable categórica ordinal que indica el estrato socioeconómico asignado al lugar en el que se encuentra la vivienda por entidades distritales. En Bogotá se encuentran estratos del 1 (uno) al 6 (seis).
- Latitud: Variable que almacena el valor correspondiente a la latitud del inmueble.
- Longitud: Variable que almacena el valor correspondiente a la longitud del inmueble.
- Antigüedad: Variable de tipo categórica ordinal que indica la antigüedad de la vivienda como un grupo determinado por el rango de edad del inmueble. Los grupos son entre 0 y 5 años, entre 5 y 10 años, entre 10 y 20 años, más de 20 años.
- Fuente: Variable categórica nominal en la cual se expresa el nombre del portal del que se obtuvo la información, siendo estos mercado libre, finca raíz y metro cuadrado.

3.3. Depuración y base de datos final

Los datos utilizados, fueron obtenidos usando librerías de Python especializadas en web scraping. En este caso, los portales seleccionados se eligieron con base a los trabajos recopilados durante la revisión bibliográfica, así como por conocimiento propio del autor de este documento.
Dado que cada página web tiene su propia estructura, fue necesario crear un archivo de datos por cada uno de los portales usados para la extracción de la información. Tras la obtención de la información, se procedió a observar la homogeneidad de las variables, obteniendo que la variable relacionada con la antigüedad del inmueble no estaba categorizada de igual manera a través de las diferentes plataformas, puesto que la escala de agrupación utilizada por el portal metro cuadrado y finca raíz son diferentes.
Para resolver la situación anterior, y debido a la cantidad de datos extraídos en cada portal y una preferencia por la escala manejada por metro cuadrado, se eligió trabajar con los datos de este portal, así como desde el portal de mercado libre, el cual maneja la información de antigüedad de la vivienda como una variable numérica, razón por la cual fue posible categorizar la variable de la misma manera que se encuentra en metro cuadrado. Adicionalmente, se llevó a cabo una eliminación de todas las viviendas cuya antigüedad fuese estrictamente mayor a 300 años, para excluir viviendas demasiado antiguas.
Posterior a la selección de las bases de datos, se procedió a contar los valores perdidos en cada base. En este punto y dada la manera en que se obtuvo la información, los valores perdidos de la variable parqueadero, que corresponde a aquellas viviendas que no poseen parqueadero, son reemplazados por el valor de cero, el cual indica la ausencia de parqueaderos en el inmueble. Tras lo anterior, se eliminaron las filas que contenían valores faltantes, conservando un total de 17852 viviendas.
Después de la eliminación de valores nulos, se tomaron las variables de precio, coordenada y tipo para filtrar los valores duplicados, siendo que, si dos viviendas contaban con las mismas coordenadas (lo cual podía ser dado por un error de aproximación de Python), eran del mismo tipo (ambas eran casas o ambas eran apartamentos) y tenían un precio igual, se consideraron como un anuncio duplicado. Después de aplicar este paso, quedaron 11769 inmuebles, de los cuales al unir las dos bases y repetir el procedimiento para los valores duplicados, no hubo cambio en el total de viviendas para analizar.
Tras la realización de los primeros filtros, se procedió a utilizar estadística descriptiva para observar ciertas características de las variables. En el caso de la variable precio, se observaron valores extremos muy alejados de sus medidas de tendencia central y de los cuartiles de los datos, siendo el valor máximo igual a cuatro billones de pesos colombianos (COP), bastante distanciado del tercer cuartil estimado, el cual se ubica en 1200 millones. Se revisó la base y los anuncios para verificar si se trataba de un error, lo cual fue cierto para un total de 20 anuncios.
En este caso, tras corregir los errores, se eliminaron datos por encima del percentil 95. Con

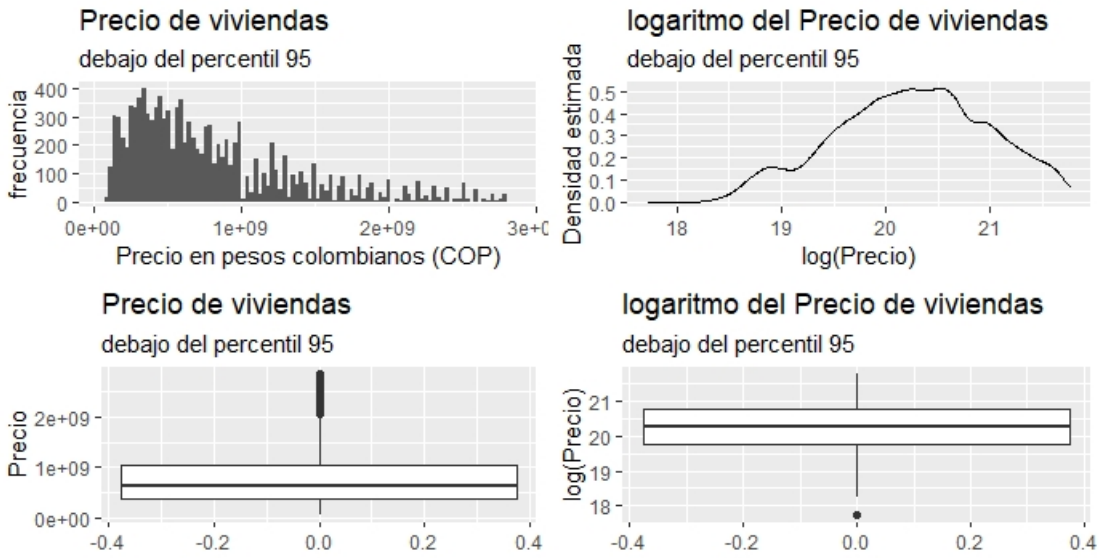

Figura 3-1.: Histograma, densidad y diagrama de caja para el precio de venta deseado y el logaritmo de este precio.

esta base reducida, se observó el comportamiento de la variable precio y se notó cierto sesgo a la derecha, por lo cual se decide aplicar el logaritmo natural a los datos, teniendo mediante esta transformación, un comportamiento bastante simétrico, siendo una razón por la cual se elige trabajar con ésta en lugar de la variable original, adicional al significado económico de los coeficientes al trabajar con logaritmo, convirtiéndose en semi-elasticidades, es decir, cambios porcentuales en la variable de interés ante cambios unitarios en las variables explicativas (ver figura **3-1**).

De manera similar, se evidenció que para la variable metros cuadrados, existían valores extremos que podían afectar los análisis, siendo el máximo valor reportado de 115000 metros cuadrados. Al realizar el mismo proceso para corregir información, se decidió tomar valores por debajo del percentil 99. Con esto, se obtuvieron los resultados mostrados en la figura **A-1**, la cual tiene un comportamiento más simétrico en los datos.

Posterior al análisis de las variables de precio y metros cuadrados de las viviendas, se observó la representación por estratos socioeconómicos. En este caso, los estratos 1 y 2 parecen estar poco representados en los datos, siendo menos del 5 % de los datos restantes (representan aproximadamente entre los dos estratos el 4.36 % de los datos). Con el fin de utilizar la mayor cantidad de datos, se juntaron los estratos 1, 2 y 3 en una sola categórica denominada base, contra la cual se van a realizar las comparaciones.

Cabe resaltar, que esta poca representación puede deberse a medios de comunicación diferentes usados en estos estratos, como lo pueden ser el voz a voz, y los anuncios físicos pegados en las puertas de las tiendas de víveres y papelerías de los barrios, o incluso en las ventanas de los mismos inmuebles. Adicionalmente, se observó una tendencia a aumentar la proporción de apartamentos frente a la proporción de casas, junto con el estrato socioeconómico,

es decir, a mayor estrato socioeconómico mayor cantidad de apartamentos con respecto a la cantidad de casas (ver figura **3-2**).

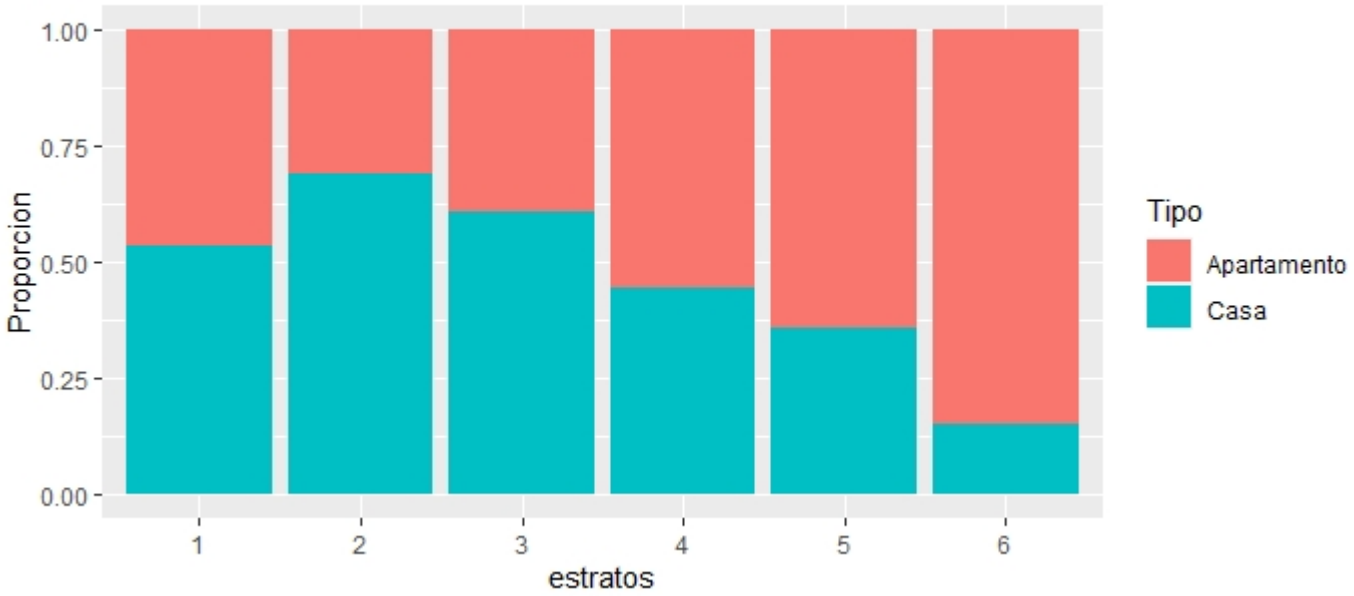

Figura 3-2.: Cantidad de viviendas por estrato y diferenciando entre casas y apartamentos.

Una vez realizado el proceso anterior, se obtuvo la base final mediante la unión de las bases de datos de las viviendas, delitos de alto impacto por barrios y los avalúos (tanto comerciales como catastrales) por manzanas de Bogotá, asociando a cada punto la información correspondiente del polígono al que pertenece. Este proceso se utilizó para evitar incluir registros que no se localizan en la ciudad de Bogotá, limitándose a usar las observaciones que pasaran estos filtros de ubicación. En este caso, la unión interna redujo la cantidad de unidades residenciales a 7852 viviendas.
Para finalizar, se eliminaron unos datos perdidos dentro de la variable antigüedad. Con la información de esta base, se graficaron las ubicaciones de las viviendas dentro del mapa de las localidades de Bogotá, en la figura **3-3**, se muestran las zonas del centro, este, oeste y norte de la ciudad, donde se concentran las observaciones. Dado lo anterior, cualquier análisis desprendido de los datos debe considerar que no se podrá extrapolar los resultados hacia afuera de estas zonas. Se excluyen los datos provenientes de la localidad de Sumapaz, por su distancia con el resto de datos y por la poca cantidad de viviendas reportadas dentro de esta (solo tres viviendas provenientes de esta localidad), dejando un total de 7850 observaciones.

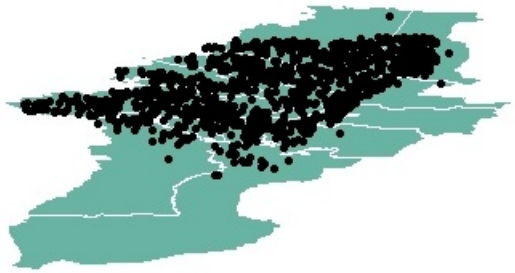

Figura 3-3.: Mapa de las localidades de las zonas del centro, este, oeste y norte de Bogotá con la ubicación de las viviendas de la base de datos filtrada.

3.4. Modelos

Es importante enfatizar la disparidad y heterogeneidad que existe en el precio de las viviendas dentro de algunas localidades de la ciudad de Bogotá, puesto que en algunos lugares se pueden apreciar barrios que, a pesar de tener costos asociados a la tenencia y residencia de un inmueble solo sostenibles por personas de ingresos altos, son rodeados o están cerca de barrios cuyas condiciones pueden carecer de servicios básicos y con costos significativamente más bajos, siendo económicamente accesibles para personas de más bajos recursos e ingresos. Siguiendo a Schabenberger & Gotway (2004) y a Urrea-Zora & Cárdenas-Mazo (2019), lo primero que se realiza es una regresión lineal mediante la técnica de mínimos cuadrados ordinarios (OLS, por sus siglas en inglés). A través de los residuales de este modelo, se realizan las pruebas I de Moran para detectar una posible autocorrelación espacial, así como pruebas para determinar heterocedasticidad.

3.4.1. Modelo lineal

$$\begin{aligned}\ln[Precio_{ijkl}] =& \mu + Tipo_j + E_k + A_l + m2_{ijkl} + habitaciones_{ijkl} + baños_{ijkl} \\ &+ parqueaderos_{ijkl} + \epsilon_{ijkl}, \quad \text{con} \quad \epsilon_{ijkl} \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned} \tag{3-1}$$

El modelo extendido que se observa en (3-1), se plantea para los tipos de vivienda $Tipo_j$, donde $j = casa, apartamento = 1, 2$; para los estratos socioeconómicos $E_k = 1, 2, 3, 4, 5, 6$; para la antigüedad A_l, donde $l = 0 - 5, 5 - 10, 10 - 20,$ mayor a 20 años $= 1, 2, 3, 4$; para las viviendas $i = 1, \ldots, n_{jkl}$, donde por cada combinación de tipo, estrato y antigüedad se puede tener una o más observaciones. En este caso, de cada vivienda (observación) se obtuvo información de los metros cuadrados ($m2_{ijkl}$), el número de baños ($baños_{ijkl}$), de habitaciones ($habitaciones_{ijkl}$), y de parqueaderos ($parqueaderos_{ijkl}$).

Para este modelo, los términos de error ϵ_{ijkl} se distribuyen idénticamente e independientes de manera normal con media cero y varianza σ^2. Además, se trabaja bajo el supuesto de que no existe covarianza entre los errores y las diferentes variables exógenas del modelo. Para la estimación de los parámetros, se realizó una reparametrización del modelo de tal manera que se estima:

$$\begin{aligned}\ln[Precio_i] =& \beta_0 + \beta_1 \text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 Baños_i + \beta_5 Parqueaderos_i \\ &+ \beta_6 \text{Estrato4}_i + \beta_7 \text{Estrato5}_i + \beta_8 \text{Estrato6}_i + \beta_9 \text{Antigüedad5-10años}_i \\ &+ \beta_{10} \text{Antigüedad10-20años}_i + \beta_{11} \text{AntigüedadMayor20años}_i \\ &+ \epsilon_i, \quad \text{con} \quad \epsilon_i \overset{\text{idd}}{\sim} N(0, \sigma^2),\end{aligned}$$

La regresión se realiza sobre el logaritmo del precio de las viviendas, por lo cual, los coeficientes estimados representan, en términos económicos, una semi-elasticidad, es decir, se interpretan como un cambio porcentual en el precio deseado de las viviendas ante un cambio en una unidad de una variable explicativa manteniendo el resto de valores constantes. Para este problema en particular, como se observa en la tabla **3-1**, $\hat{\beta}_1$ tiene signo negativo, indicando que las casas cuentan en promedio con una menor valoración frente a los apartamentos, relacionándose este coeficiente estimado con la variable de TipoCasa$_i$, la cual toma el valor de uno si la i-ésima vivienda es una casa y cero en caso contrario.
La valoración que obtiene un inmueble manteniendo todas las variables en un valor constantes y solo incrementando la variable $m2_i$ en una unidad para una i-ésima vivienda sería alrededor del 0.3 % dado el coeficiente estimado $\hat{\beta}_2$; $\hat{\beta}_3$ indica que manteniendo constante todo lo demás, una vivienda con una habitación adicional perdería su valoración en aproximadamente 6 %, esto indica que aumentos unitarios en la variable Hab_i disminuyen el valor de la vivienda; $\hat{\beta}_4$ muestra que dos inmuebles con características iguales, si uno de estos llegará a tener un baño adicional, éste adquiere un 17.82 % más de valor, frente al otro inmueble;$\hat{\beta}_5$ proporciona el impacto de un parqueadero adicional en la vivienda, el cual es un incremento alrededor del 6 %, este coeficiente estimado está relacionado con la variable $parqueaderos_i$, que índica el número de parqueaderos de la i-ésima vivienda.
Los coeficientes estimados $\hat{\beta}_6$, $\hat{\beta}_7$ y $\hat{\beta}_8$ asociadas a las variables Estrato4$_i$, Estrato5$_i$ y Estrato6$_i$ respectivamente, muestran que a medida que el estrato aumenta, existe una mayor valoración de las viviendas con respecto a la categoría base, la cual agrupa viviendas de los estratos 1, 2 y 3, siendo las variables mencionadas de tipo binarias, tomando el valor uno si el i-ésimo inmueble pertenece al estrato especificado y cero en caso contrario. Las variables antigüedad 5-10 años, antigüedad 10-20 años y antigüedad Mayor 20 años (variables binarias con valor de uno si el inmueble está en ese grupo de antigüedad y cero en caso opuesto), con coeficientes estimados $\hat{\beta}_9$, $\hat{\beta}_{10}$ y $\hat{\beta}_{11}$ respectivamente, muestran que a mayor antigüedad, existe una depreciación en el valor con respecto a viviendas que se encuentran entre los 0 y 5 años de antigüedad.

Tabla 3-1.: Coeficientes estimados mediante mínimos cuadrados ordinarios regresión lineal multiple

Coeficientes	**Estimación**	**EE**	**Estadístico t**	**Valor-p**
Intercepto	19.0200	0.01711	1111.76	<0.0000
TipoCasa	-0.0746	0.01137	-6.56	<0.0000
$m2$	0.0037	0.00006	56.60	<0.0000
Hab	-0.0676	0.00450	-15.02	<0.0000
Baños	0.1782	0.00527	33.82	<0.0000
Parqueaderos	0.0634	0.00488	12.98	<0.0000
Estrato4	0.3593	0.01204	29.85	<0.0000
Estrato5	0.4298	0.01401	30.68	<0.0000
Estrato6	0.5943	0.01534	38.75	<0.0000
Antigüedad5-10años	-0.0840	0.01552	-5.41	<0.0000
Antigüedad10-20años	-0.1250	0.01350	-9.26	<0.0000
AntigüedadMayor20años	-0.1127	0.01350	-5.41	<0.0000

En este caso, pese a que el interés de la regresión se centra más en la obtención de las estimaciones de los errores que de los parámetros en sí, se resalta que todos éstos fueron significativos a un nivel de confianza del 5 %. Igualmente, se observa en la tabla **3-1** (se presenta por cada coeficiente, su estimación, su error estándar, su estadístico de prueba y su valor-p), que los signos de los efectos asociados a las variables referentes al tipo de inmueble (valen más los apartamentos frente a las casas) y número de habitaciones parecen contrarios a lo esperado. Sin embargo, no deja de tener explicación en la percepción de seguridad de una torre de apartamentos frente a una casa en medio de un barrio, la antigüedad de las casas frente a los apartamentos y el efecto ancla en el precio de torres nuevas de apartamentos en barrios que originalmente consistían en solo casas, es decir, las nuevas viviendas entran en el mercado con un precio muy similar o incluso por encima del precio de las viviendas alrededor, así como el cambio en preferencias de los consumidores ante familias más pequeñas en número de integrantes que viven en un mismo hogar.

Posterior a la regresión, se obtuvieron los errores estimados, con los cuales se calculó el índice I de Moran estimado para evaluar la hipótesis nula de que los precios de venta deseados por los vendedores son ubicados espacialmente de manera aleatoria, en contraste con la hipótesis alternativa de que existe una dependencia espacial de los datos. Para este proceso se realizaron dos pruebas utilizando la misma matriz de pesos espaciales.

Las pruebas se distinguen por el uso o no del supuesto de normalidad de los residuos, en la cual una prueba se basa en este supuesto y la otra se basa en una simulación del valor-p y el estadístico de prueba a partir de simulaciones de cadenas de Monte Carlo. Por su parte, la matriz usada fue una matriz estandarizada de contigüidad, donde un vecino es definido por estar en dentro de un radio de 1000 metros de otra observación. En ambos casos se

rechazó la hipótesis nula a un nivel de significancia del 5 %, con lo cual no se rechaza que el precio de venta inicial deseado por los vendedores es influenciado por sus vecinos y, por ende, es necesario tener presente la influencia que pueden tener observaciones cercanas en una observación especifica i (ver tabla **3-2**).

Tabla 3-2.: Test de Moran I para dependencia espacial sobre los residuales de regresión lineal

Test de Moran I		
Matriz	Estadístico	P-value
Matriz de vecinos estandarizada (normalidad)	63.21	<0.0000
Matriz de vecinos estandarizada (MC)	0.13	0.001

El signo positivo del estadístico en el test I de Moran indica que los datos se agrupan de manera que viviendas de altos precios están rodeadas de viviendas de altos precios, al igual que viviendas de bajos precios se encontrarían rodeadas de viviendas de bajos precios. Una visualización de este efecto se observa en la figura **3-4**, en la cual para la matriz de pesos estandarizada existe una tendencia positiva entre los residuales y sus vecinos.

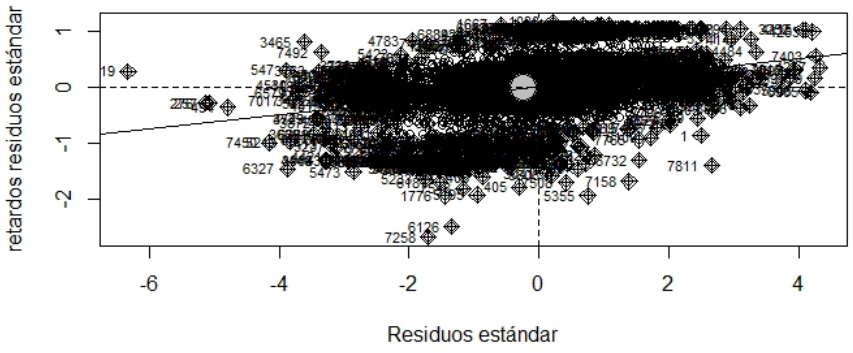

Figura 3-4.: Gráfico de Moran de los residuos versus sus rezagos, para una matriz de pesos estandarizada de contigüidad basada en vecinos de 0-1000 metros

3.4.2. Modelo jerárquico

Modelo mixto basado en datos de barrios

El modelo lineal jerárquico se plantea para el nivel 1 de la siguiente manera:

$$\begin{aligned}\ln[Precio_{ik}] =& \beta_{0k} + \beta_1 \text{TipoCasa}_{ik} + \beta_{2k} m2_{ik} + \beta_3 Hab_{ik} + \beta_4 Baños_{ik} + \beta_5 Parqueaderos_{ik} \\ &+ \beta_6 \text{Estrato4}_{ik} + \beta_7 \text{Estrato5}_{ik} + \beta_8 \text{Estrato6}_{ik} + \beta_9 \text{Antigüedad5-10años}_{ik} \\ &+ \beta_{10} \text{Antigüedad10-20años}_{ik} + \beta_{11} \text{AntigüedadMayor20años}_{ik} \\ &+ \epsilon_{ik}, \quad \text{con} \quad \epsilon_{ik} \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)\end{aligned}$$

donde i se refiere a la i-ésima vivienda dentro del k-ésimo barrio, con $k = 1, \ldots, 506$, $i = 1, \ldots, n_k$; β_{0k} es el intercepto asociado al k-ésimo barrio; β_1 representa la diferencia porcentual entre el valor en el mercado de una casa y un apartamento manteniendo todas las demás características constantes, asociado a la variable TipoCasa_{ik} la cual toma el valor de uno si el inmueble es una casa y cero en cualquier otro caso.
Al igual que en el modelo lineal, los coeficientes se interpretan como semi-elasticidades, es decir, β_{2k} mide el cambio porcentual en el precio del inmueble ante una variación de un metro cuadrado en el k-ésimo barrio, lo que implica cambios en la variable $m2_{ik}$; β_3 representa el cambio porcentual en el precio ante el aumento o disminución de una habitación dentro de las viviendas, siendo el coeficiente asociado a la variable Hab_{ik}; β_4 representa el cambio porcentual en el precio ante el aumento o disminución del número de baños $Baños_{ik}$ que tiene la vivienda; β_5 mide el cambio porcentual en el precio deseado de venta ante una variación de una unidad para el estacionamiento de vehículos, lo cual implica cambios en la variable $Parqueaderos_{ik}$. β_6, β_7 y β_8 miden la diferencia porcentual en el precio de un inmueble ubicado en el estrato 4, estrato 5 y estrato 6, respectivamente, frente a los inmuebles ubicados en el denominado grupo base (el cual agrupa los estratos 1, 2 y 3), dado que las variables Estrato4_{ik}, Estrato5_{ik} y Estrato6_{ik} son variables que toman el valor de uno si la vivienda pertenece al estrato indicado en el nombre de la misma y cero en caso contrario. Por su parte, β_9, β_{10} y β_{11}, miden la diferencia porcentual en el precio de un inmueble cuya antigüedad esta entre 5-10 años, 10-20 años, o es mayor a 20 años, respectivamente, frente a viviendas de antigüedad entre 0-5 años, siendo que las variables $\text{Antigüedad5-10años}_{ik}$, $\text{Antigüedad10-20años}_{ik}$ y $\text{AntigüedadMayor20años}_{ik}$ al igual que en el caso de los estratos, variables binarias cuyo valor es uno si pertenece al grupo especificado en el nombre de la variable y cero en otros casos.
En este caso, el modelo cuenta con un intercepto aleatorio y una pendiente aleatoria para la variable relacionada con los metros cuadrados de la i-ésima vivienda, $m2_{ik}$. El primero se considera puesto que cada barrio cuenta con una mayor o menor valoración inicial de acuerdo a múltiples factores en la mente de los consumidores (como la seguridad del barrio, zonas verdes, o simplemente una concepción general del valor de un inmueble dentro de dicho

barrio), mientras la segunda se considera porque existe una mayor o menor valoración de los metros cuadrados dependiendo el barrio al cual pertenece la vivienda. Lo anterior genera un modelo lineal para el nivel 2 como se muestra a continuación:

$$\begin{aligned}\beta_{0k} =& \beta_{00} + U_{0k}, \quad \text{con} \quad U_{0k} \overset{\text{iid}}{\sim} N(0, \sigma^2_{U0})\\ \beta_{2k} =& \beta_{20} + U_{2k}, \quad \text{con} \quad U_{2k} \overset{\text{iid}}{\sim} N(0, \sigma^2_{U2})\end{aligned}$$

donde β_{00} es un intercepto común para todas las observaciones, mientras β_{20} es el efecto común que tiene un cambio en una unidad en los metros cuadrados de una vivienda sobre el cambio porcentual del precio de venta deseado. Cabe aclarar que los efectos aleatorios U_{0k}, U_{2k} y el error ϵ_{ik} tienen covarianza cero, es decir, $Cov(U_{0k}, \epsilon_{ik}) = Cov(U_{2k}, \epsilon_{ik}) = 0$, $\forall i, k$. Reemplazando tanto β_{0k} como β_{2k} se tiene que el modelo se puede expresar como

$$\begin{aligned}\ln[Precio_{ik}] =& \beta_{00} + \beta_1 \text{TipoCasa}_{ik} + \beta_{20} m2_{ik} + \beta_3 Hab_{ik} + \beta_4 Baños_{ik} + \beta_5 Parqueaderos_{ik}\\ &+\beta_6 \text{Estrato4}_{ik} + \beta_7 \text{Estrato5}_{ik} + \beta_8 \text{Estrato6}_{ik} + \beta_9 \text{Antigüedad5-10años}_{ik}\\ &+\beta_{10} \text{Antigüedad10-20años}_{ik} + \beta_{11} \text{AntigüedadMayor20años}_{ik}\\ &+U_{0k} + U_{2k} m2_{ik} + \epsilon_{ik}\end{aligned} \tag{3-2}$$

En este caso, los primeros doce términos son componentes fijos o determinísticos de la ecuación, mientras los últimos tres términos son aleatorios. Se supone con anterioridad que los términos aleatorios no tienen correlación con los componentes determinísticos. Adicionalmente, se supone que $Cov(U_{0k}, \epsilon_{ik}) = Cov(U_{2k}, \epsilon_{ik}) = 0$ para todo i y k, y que $Cov(U_{0k}, U_{2k}) = \sigma_{U0U2}$, con lo cual, $Var(\ln[Precio_{ik}]) = \sigma^2_{U0} + (m2_{ik})^2 \sigma^2_{U2} + 2m2_{ik}\sigma^2_{U0U2} + \sigma^2_\epsilon$. Con todo lo anterior, se puede plantear el modelo de manera matricial como

$$\begin{aligned}\mathbf{Y} =& \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_0\mathbf{U}_0 + \mathbf{Z}_2\mathbf{U}_2 + \boldsymbol{\epsilon}\\ \mathbf{Y} =& \mathbf{X}\boldsymbol{\beta} + \begin{pmatrix}\mathbf{Z}_0 & \mathbf{Z}_2\end{pmatrix}\begin{pmatrix}\mathbf{U}_0\\ \mathbf{U}_2\end{pmatrix} + \boldsymbol{\epsilon}\\ \mathbf{Y} =& \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U} + \boldsymbol{\epsilon},\end{aligned}$$

donde $\mathbf{X}$ es una matriz diseño asocida a los efectos fijos del modelo, $\boldsymbol{\beta}$ es un vector de parámetros asociados a los efectos fijos del modelo, $\mathbf{Z}_0$ y $\mathbf{Z}_2$ son matrices diseño asociados a los vectores aleatorios del intercepto $\mathbf{U}_0$ y de la pendiente $\mathbf{U}_2$. De igual manera, se tiene que

$$\boldsymbol{\epsilon} \sim NMV(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_{7850\times 7850}) \;\; \text{y} \;\; \mathbf{U} \sim NMV(\mathbf{0}, \boldsymbol{\Sigma}_{2\times 2}),$$

Por lo cual,

$$\mathbf{Y}|\mathbf{X} \sim NMV(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$$

con

$$\mathbf{V} = \mathbf{Z}\boldsymbol{\Sigma}\mathbf{Z}^t + \sigma_\epsilon^2\mathbf{I}$$

El uso de este tipo de modelamiento se basa en el hecho de que se puede tomar el valor de venta deseado de las viviendas como un experimento con muestras repetidas, donde cada individuo posee características únicas y propias que pueden llegar a afectar este valor, es decir, se tiene que la variable puede ser afectada por la variación entre los individuos tomados. En este caso particular, se tiene una desagregación bastante específica al contar con los barrios en los cuales se encuentran ubicadas las viviendas.
Siguiendo a Lee et al. (2017), tal como se expuso en el marco teórico, se procede a calcular los parámetros del modelo, tanto los asociados a la media como los parámetros de dispersión de los efectos aleatorios. En este caso, la estimación de los parámetros de dispersión $\boldsymbol{\tau}$ dieron significativas; sin embargo, tal como se muestra en la tabla **3-3**, la desviación estándar asociada a la pendiente aleatoria es muy cercana a cero, por lo que se realiza una transformación a la variable metros cuadrados trabajando con su logaritmo. De esta manera, la interpretación económica de esta variable cambia, puesto que su coeficiente asociado es el cambio porcentual en el precio deseado por el vendedor ante un cambio porcentual en los metros cuadrados del inmueble.
Como se observa en la tabla **3-4** sobre los efectos aleatorios del modelo, los intervalos de confianza al 95 % sobre el valor de la desviación estándar de dichos efectos no incluyen el cero, por lo que la variación de los grupos es significativa, afectando de manera diferente tanto el intercepto como la pendiente asociada al logaritmo de los metros cuadrados. Por su parte, los efectos fijos del modelo dieron significativos (ver tabla **3-5**).

Tabla 3-3.: Tabla reporte efectos aletorios del modelo mixto antes de aplicar logaritmo a la variable metros cuadrados

Efectos aleatorios intervalo de confianza 95 %			
Efecto	Limite inferior	Estimado	Limite superior
σ_{U0}	0.2950	0.3270	0.3626
σ_{U2}	0.0012	0.0014	0.0016
σ_{U0U2}	-0.7717	-0.6933	-0.5941

Tabla 3-4.: Tabla reporte efectos aletorios del modelo mixto después de aplicar logaritmo a la variable metros cuadrados

Efectos aleatorios intervalo de confianza 95 %			
Efecto	Limite inferior	Estimado	Limite superior
σ_{U0}	0.6164	0.7182	0.8269
σ_{U2}	0.1329	0.1542	0.1788
σ_{U0U2}	-0.9691	-0.9556	-0.9363

Tabla 3-5.: Tabla reporte efectos fijos del modelo mixto

Coeficientes	**Estimación**	**EE**	**Estadístico t**	**valor-p**
Intercepto	16.2700	0.0561	289.94	<0.0000
TipoCasa	-0.0736	0.0092	-8.04	<0.0000
$\ln(m2)$	0.7659	0.0132	58.12	<0.0000
Hab	-0.0640	0.0036	-17.88	<0.0000
Baños	0.0776	0.0042	18.65	<0.0000
Parqueaderos	0.0406	0.0036	11.14	<0.0000
Estrato 4	0.1855	0.0110	16.93	<0.0000
Estrato 5	0.2340	0.0124	18.87	<0.0000
Estrato 6	0.2943	0.0129	22.90	<0.0000
Antigüedad5-10años	-0.0742	0.0114	-6.49	<0.0000
Antigüedad10-20años	-0.1448	0.0099	-14.63	<0.0000
AntigüedadMayor20años	-0.1722	0.0099	-17.36	<0.0000

Ahora bien, con respecto a los supuestos de los efectos aleatorios y los errores del modelo, se tiene que condicionando estos últimos sobre valores fijos de los primeros, la distribución resultante para cada observación es idéntica e independiente de las demás de manera normal, es decir, los errores tomando los efectos aleatorios **U** como fijos deberían cumplir los supuestos de un modelo de regresión clásico. En este caso, pese a trabajar los datos de diferentes formas (los datos sin manipular, eliminando datos atípicos, usando un vector de pesos para dar menor relevancia a aquellos datos fuera de los patrones comunes), no parece presentarse una normalidad sobre los errores, puesto que con las pruebas utilizadas (Kolmogórov-Smirnov, Jarque-Bera, Lilliefors, Anderson-Darling, Cramér-von Mises), se obtienen valores-p menores a 0.05, rechazando la hipótesis nula de normalidad en la distribución de los errores. Del mismo modo, se probaron diferentes transformaciones como Box-Cox y raíz cuadrada sin cambio alguno (ver anexo D).

Aunque este supuesto se rechaza por medio de las pruebas, se decide trabajar con este supuesto de manera asintótica, haciendo uso del logaritmo del precio deseado dado la inter-

pretación económica que poseen las variables, siendo esto posible debido a que se cuenta con una buena cantidad de observaciones que permiten mantener este supuesto por lo menos de esta manera. Adicionalmente al graficar la densidad estimada de los residuales estándar del modelo trabajando con el logaritmo y compararlos con una distribución normal estándar se observa un comportamiento simétrico, centrado en cero pero con un comportamiento más leptocúrtico (ver figura **3-5**).
A pesar de no contar con la normalidad, se realizan pruebas para validar la homocedasticidad de los errores tanto paramétricas como no paramétricas, mostrando en la mayoría de casos valores-p por encima del 0.05, dando evidencia estadística para no rechazar la hipótesis nula de que el error es constante a través de las observaciones (ver tabla **3-6**).

Tabla 3-6.: Tabla reporte sobre el supuesto de homocedasticidad de los errores condicionados del modelo mixto

Pruebas para la homocedasticidad de los errores condicionados del modelo	
Prueba	valor-p
Levene	0.5166
Fligner-Killeen	0.5149
Kruskal-Wallis	0.5326

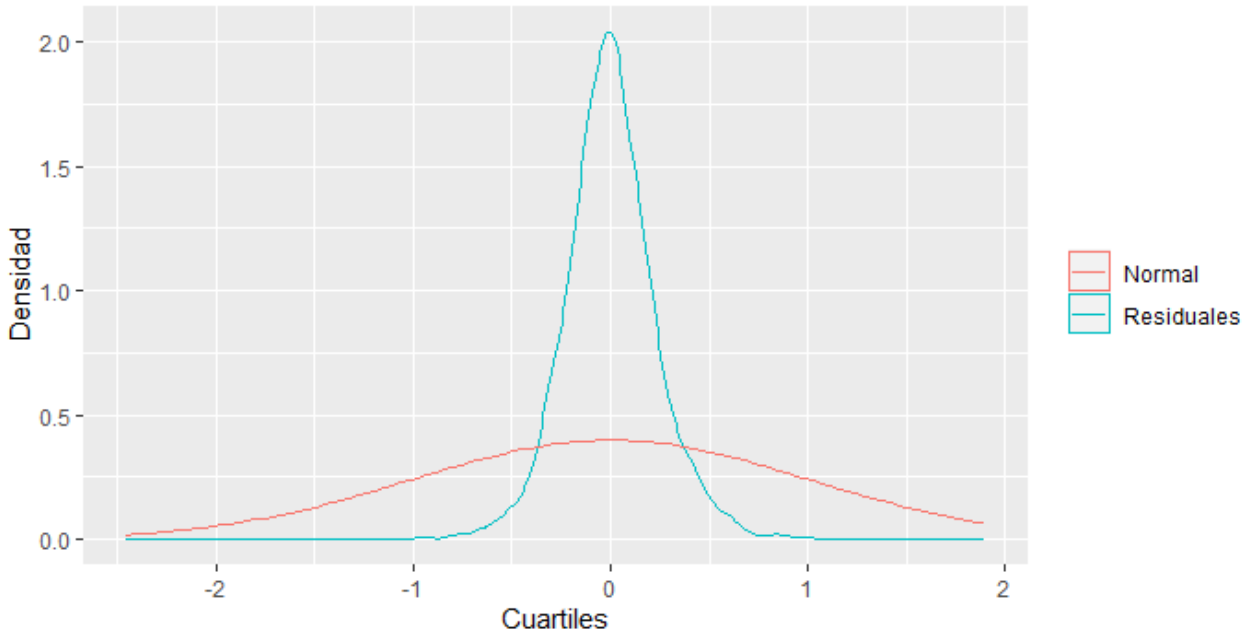

Figura 3-5.: Distribución estimada de los residuales estándar condicionados del modelo (azul) y función de densidad de una distribución normal estándar (rojo)

De forma similar, al validar los supuestos asociados a los efectos aleatorios del modelo (EAM),

se tiene que el supuesto de homocedasticidad no se rechaza (ver tabla **3-7**). Por otro lado, al poner a prueba el supuesto de normalidad, se obtuvo con las pruebas aplicadas (prueba de Kolmogórov-Smirnov, de Jarque-Bera, de Anderson-Darling y la de Cramér-von Mises), que no se tiene tal característica en estos efectos, puesto que con todas las pruebas aplicadas se rechaza la hipótesis nula de que los efectos de manera marginal vienen de una distribución normal (el valor-p es inferior a 0.00001 en todos los casos).

Tabla 3-7.: Tabla reporte sobre los supuestos de los efectos aleatorios del modelo mixto

Supuestos sobre efectos aleatorios del modelo				
Prueba	EAM	Supuesto a validar	Hipótesis nula	valor-p
Levene	U_0	Homocedasticidad	El comportamiento de los residuales entre los grupos es igual	0.4478
Fligner-Killeen	U_0	Homocedasticidad	Varianzas de los grupos son, iguales, no se asume normalidad	0.4357
Kruskal-Wallis	U_0	Homocedasticidad	Varianzas de los grupos son, iguales, no se asume normalidad	0.5235
Levene	U_2	Homocedasticidad	El comportamiento de los residuales entre los grupos es igual	0.4861
Fligner-Killeen	U_2	Homocedasticidad	Varianzas de los grupos son, iguales, no se asume normalidad	0.4799
Kruskal-Wallis	U_2	Homocedasticidad	Varianzas de los grupos son, iguales, no se asume normalidad	0.5206

Modelo mixto basado en coordenadas

El modelo lineal jerárquico se plantea para el nivel 1 de la siguiente manera:

$$\begin{aligned}\ln[Precio_i] =& \beta_0(\mathbf{S}_i) + \beta_1 \text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 Baños_i + \beta_5 Parqueaderos_i \\ &+\beta_6 \text{Estrato4}_i + \beta_7 \text{Estrato5}_i + \beta_8 \text{Estrato6}_i + \beta_9 \text{Antigüedad5-10años}_i \\ &+\beta_{10} \text{Antigüedad10-20años}_i + \beta_{11} \text{AntigüedadMayor20años}_i \\ &+\epsilon_i, \quad \text{con} \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma_\epsilon^2)\end{aligned}$$

donde i se refiere a la i-ésima vivienda, $\beta_0(\mathbf{S}_i)$ es el intercepto aleatorio que dependerá del vector de coordenadas de la i-ésima vivienda. Los coeficientes estimados se interpretan como semi-elasticidades, por ende β_1 representa la diferencia porcentual entre el valor en el mercado de una casa y un apartamento, asociado a la variable TipoCasa_i, la cual toma un valor de uno cuando la i-ésima vivienda es una casa y toma el valor de cero en cualquier otro caso.

En este modelo, β_2 mide el cambio porcentual en el precio del inmueble ante una variación de un metro cuadrado de la variable $m2_i$, es decir, ante cambios en el i-ésimo inmueble manteniendo todas las demás variables constantes; β_3 representa el cambio porcentual en el precio ante el aumento o disminución de una habitación dentro de las viviendas, siendo el coeficiente asociado a la variable Hab_i; β_4 representa el cambio porcentual en el precio ante el aumento o disminución del número de baños $Baños_i$ que tiene la vivienda; β_5 mide el cambio porcentual en el precio deseado de venta ante una variación de una unidad para el estacionamiento de vehículos, lo cual implica cambios en la variable $Parqueaderos_i$. β_6, β_7 y β_8 miden la diferencia porcentual en el precio de un inmueble ubicado en el estrato 4, estrato 5 y estrato 6, respectivamente, frente a los inmuebles ubicados en el denominado grupo base (el cual agrupa los estratos 1, 2 y 3), dado que las variables Estrato4_i, Estrato5_i y Estrato6_i son variables que toman el valor de uno si la vivienda pertenece al estrato indicado en el nombre de la misma y cero en caso contrario. Por su parte, β_9, β_{10} y β_{11}, miden la diferencia porcentual en el precio de un inmueble cuya antigüedad esta entre 5-10 años, 10-20 años, o es mayor a 20 años, respectivamente, frente a viviendas de antigüedad entre 0-5 años, siendo que las variables $\text{antigüedad5-10años}_i$, $\text{antigüedad10-20años}_i$ y $\text{antigüedadMayor20años}_i$ al igual que en el caso de los estratos, variables binarias cuyo valor es uno si pertenece al grupo especificado en el nombre de la variable y cero en otros casos.
En este caso, se plantea un modelo lineal para el intercepto en el nivel 2 como se muestra a continuación:

$$\beta_0(\mathbf{S}) = \mathbf{1}\beta_{00} + \mathbf{U}_0(\mathbf{S}), \quad \text{con} \quad \mathbf{U}_0(\mathbf{S}) \sim NMV(\mathbf{0}, \sigma^2_{U0}\boldsymbol{\varrho}(d_{ij}, a,)),$$

donde β_{00} es un intercepto común para todas las observaciones, $\mathbf{1}$ es un vector columna de unos cuyas dimensiones son de 7850×1, mientras $\mathbf{U_0}(\mathbf{S})$ es un vector aleatorio que se distribuye de manera normal multivariante cuyo objetivo es absorber la dependencia espacial de los datos, con un vector de medias cero y una matriz de varianzas y covarianzas que se compone de un término común σ^2_{U0} y una matriz de correlaciones $\boldsymbol{\varrho}$, cuyos valores dependen de la distancia d_{ij} entre la observación i y la observación j, y un parámetro a asociado al rango de dependencia espacial. Cabe aclarar que el efecto aleatorio $U_0(\mathbf{S}_i)$ y el error ϵ_i tienen covarianza cero, es decir, $Cov(U_0(\mathbf{S}_i), \epsilon_i) = 0$, para todo i.
Reemplazando el valor asociado al intercepto de cada observación, se tiene que el modelo se puede expresar como:

$$\begin{aligned}\ln[Precio_i] =& \beta_0(\mathbf{S}_i) + \beta_1\text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 Baños_i + \beta_5 Parqueaderos_i \\ &+\beta_6\text{Estrato4}_i + \beta_7\text{Estrato5}_i + \beta_8\text{Estrato6}_i + \beta_9\text{Antigüedad5-10años}_i \\ &+\beta_{10}\text{Antigüedad10-20años}_i + \beta_{11}\text{AntigüedadMayor20años}_i \\ &+U_0(\mathbf{S}_i) + \epsilon_i\end{aligned} \tag{3-3}$$

En este caso, al igual que en el modelo de efectos mixtos basado en datos de barrios, se puede representar de manera matricial como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{U}_0(\mathbf{S}) + \boldsymbol{\epsilon},$$

donde

$$\boldsymbol{\epsilon} \sim NMV(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{7850\times 7850}) \;\text{ y }\; \mathbf{U}_0(\mathbf{S}) \sim NMV(\mathbf{0}, \sigma_{U0}^2 \boldsymbol{\varrho}(d, a,))$$

En este caso particular $\mathbf{Z} = \mathbf{I}$, donde $\mathbf{I}$ es una matriz identidad de $n \times n$, por lo cual,

$$\mathbf{Y}|\mathbf{X} \sim NMV(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}),$$

donde

$$\mathbf{V} = \boldsymbol{\Sigma} + \sigma_\epsilon^2 \mathbf{I} \quad \text{y,} \quad \boldsymbol{\Sigma} = \sigma_{U0}^2 \boldsymbol{\varrho}(d, a).$$

Entonces, repitiendo el procedimiento del modelo mixto dado en el marco teórico, se tiene que la distribución conjunta de la variable dependiente y los efectos aleatorios es:

$$\begin{aligned}
f_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{U}(\mathbf{S})) =& f_{\boldsymbol{\theta}}(\mathbf{Y}|\mathbf{U}(\mathbf{S})) f_{\boldsymbol{\theta}}(\mathbf{U}(\mathbf{S})) \\
=& \frac{1}{\sqrt{(2\pi)^n (\sigma^2)^{\frac{n}{2}}}} \exp\left\{ \frac{-1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}(\mathbf{S}))^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}(\mathbf{S})) \right\} \\
& \frac{1}{\sqrt{(2\pi)^2}} \left|\boldsymbol{\Sigma}\right|^{\frac{-1}{2}} \exp\left\{ \frac{-1}{2} \mathbf{U}^t(\mathbf{S}) \boldsymbol{\Sigma}^{-1} \mathbf{U}(\mathbf{S}) \right\} \\
=& \frac{1}{\sqrt{(2\pi)^{n+2}}} \left|\sigma^2 \mathbf{I}\right|^{\frac{-1}{2}} \left|\boldsymbol{\Sigma}\right|^{\frac{-1}{2}} \exp\left\{ \frac{-1}{2\sigma^2} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}(\mathbf{S}))^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{U}(\mathbf{S})) \right. \\
& \left. - \frac{1}{2} \mathbf{U}^t(\mathbf{S}) \boldsymbol{\Sigma}^{-1} \mathbf{U}(\mathbf{S}) \right\} \\
=& L(\boldsymbol{\theta}, \mathbf{U}(\mathbf{S}); \mathbf{Y}),
\end{aligned}$$

en el cual, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\tau})$ con $\boldsymbol{\tau} = (\sigma_\epsilon^2, \sigma_{U0}^2, a)$. Obtenidos todos estos elementos, se puede proceder al igual que se hizo en el modelo de efectos mixtos basado en datos de barrios, estimando los parámetros $\boldsymbol{\beta}$ y los efectos aleatorios $\mathbf{U}(\mathbf{S})$ mediante la log-verosimilitud conjunta, mientras que para los parámetros de dispersión se utiliza la log-verosimilitud restringida, que es una aproximación a la log-verosimilitud marginal de los mismos. Las estimaciones de los parámetros fijos se presentan en la tabla **3-8**. De estos, se destaca que la variable que indica si el inmueble es casa o apartamento es significativa con signo positivo, lo cual refleja que

bajo este modelo las casas se les otorga un mayor precio deseado inicial manteniendo el resto de variables en valores fijos.

Tabla 3-8.: Coeficientes estimados modelo de efectos mixtos basado en las coordenadas

Coeficientes	**Estimación**	**EE**	**Estadístico t**	**valor-p**
Intercepto	18.9900	0.0381	497.81	<0.0000
TipoCasa	0.0489	0.0102	4.80	<0.0000
$m2$	0.0032	0.00006	56.85	<0.0000
Hab	-0.0317	0.0040	-7.96	<0.0000
Baños	0.1407	0.0046	30.83	<0.0000
Parqueaderos	0.0569	0.0042	13.66	<0.0000
Estrato4	0.1915	0.0129	14.84	<0.0000
Estrato5	0.2342	0.0146	16.07	<0.0000
Estrato6	0.2988	0.0151	19.83	<0.0000
Antigüedad5-10años	-0.0468	0.0132	-3.55	0.0004
Antigüedad10-20años	-0.1101	0.0114	-9.66	<0.0000
AntigüedadMayor20años	-0.1334	0.0114	-11.68	<0.0000
σ^2_{U0}	0.0955			
a	0.0011			

En este modelo, se aclara que se utilizó una función de correlación exponencial, por lo cual, el parámetro a se interpreta como el inverso de la dependencia espacial y se relaciona con el punto en que la correlación espacial tiende a caer a valores cercanos a cero. En este caso, con el valor estimado $\hat{a}$, se obtuvo que ese punto está a una distancia de aproximadamente 924 metros, siendo éste el rango de dependencia espacial aproximado para el precio de venta inicial deseado por los vendedores de viviendas usadas en la ciudad de Bogotá (ver tabla **3-8**).

Ahora, para comprobar la pertinencia del modelo estimando efectos aleatorios con respecto a un modelo lineal estimado por mínimos cuadrados, siguiendo a Lee et al. (2017), se realiza una prueba de razón de verosimilitud, teniendo en cuenta que los modelos cuenten con los mismos efectos fijos (condición primordial para que la prueba sea válida). Los resultados arrojan que la inclusión de estos efectos aleatorios son significativas, siendo de fondo la hipótesis nula igual a que los parámetros de dispersión asociados a estos efectos son iguales a cero (ver tabla **3-9**).

Tabla 3-9.: Prueba de verosimilitud para los efectos aleatorios del modelo mixto basado en coordenadas

Modelo	df	loglik	Estadístico	valor-p
Efecto mixto espacial	15	-1962.42		
lineal	13	-3038.90	2152.96	<0.0000

En este caso, dado que la estimación de los efectos aleatorios se da por medio de las coordenadas, es posible estimar los valores de estos efectos en ubicaciones fuera de los datos recolectados para este estudio, teniendo en cuenta que dentro de las limitaciones que este método se encuentra que es eficiente para interpolar datos entre ubicaciones más no para extrapolar fuera del rango de la longitud y latitud de los mismos.
Al igual que en el modelo de efectos mixtos basado en datos de barrios se validan los supuestos sobre el error condicionado del modelo, encontrando que no se puede afirmar que vengan de una distribución normal, dado que las pruebas realizadas muestran valores-p inferiores a 0.05, rechazando la hipótesis de normalidad, aunque al igual que en el modelo basado en barrios, se trabaja con el supuesto de manera asintótica. Por su parte, no hay evidencia estadística para rechazar la hipótesis nula de homocedasticidad en los errores, puesto que las pruebas robustas ante la falta de normalidad muestra que este supuesto no se rechaza (ver tabla **3-10**).

Tabla 3-10.: Tabla reporte sobre el supuesto de homocedasticidad de los errores condicionados del modelo mixto basado en coordenadas

Pruebas para la homocedasticidad de los errores condicionados del modelo	
Prueba	valor-p
Levene	0.4931
Fligner-Killeen	0.4826
Kruskal-Wallis	0.5147

3.4.3. Modelos espaciales

Modelo de rezago en la variable dependiente (SAR)

Siguiendo lo presentado por Urrea-Zora & Cárdenas-Mazo (2019) en su artículo sobre calidad de barrios en Medellín, posterior a la realización del modelo lineal y las pruebas de dependencia espacial, se realizó un primer modelo con la matriz de pesos espaciales antes expuesta, incluyendo un efecto del logaritmo del precio de venta deseado de las viviendas vecinas sobre una observación especifica i. Se propone:

$$\begin{aligned}\ln[Precio_i] =& \rho \sum_{p=1}^{n} w_{ip} \ln[Precio_p] + \beta_0 + \beta_1 \text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 Baños_i \\ &+ \beta_5 Parqueaderos_i + \beta_6 \text{Estrato4}_i + \beta_7 \text{Estrato5}_i + \beta_8 \text{Estrato6}_i \\ &+ \beta_9 \text{Antigüedad5-10años}_i + \beta_{10} \text{Antigüedad10-20años}_i \\ &+ \beta_{11} \text{AntigüedadMayor20años}_i + \epsilon_i, \quad \text{con} \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)\end{aligned} \tag{3-4}$$

donde $\ln[Precio_p]$ es el logaritmo del precio de la p-ésima vivienda; la variable TipoCasa_i es una variable dicotómica que indica con el valor uno si la i-ésima vivienda es una casa o con el valor cero en caso de que no; $m2_i$ son los metros cuadrados de la i-ésima vivienda; Hab_i hace referencia al número de habitaciones de la vivienda i; $Baños_i$ se refiere al número de baños de la vivienda i; $Parqueaderos_i$ indica el número de parqueaderos con los que cuenta la vivienda; variables de estratos, siendo variables dicotómicas que indican si la vivienda i pertenece ya sea al estrato 4, 5 o 6; variables dicotómicas referentes a la antigüedad de la vivienda i, la cual puede estar en la categoría de 5-10 años, en la categoría de 10-20 años, o estar en la categoría para inmuebles de más de 20 años de antigüedad.
En este caso, en la ecuación (3-4), el término estimado $\hat{\rho}$, es el efecto promedio que tiene los vecinos sobre una observación específica, w_{ip} es el valor que está en la i-ésima fila y p-ésima columna de la matriz de pesos espaciales. Por su parte, se debe aclarar que los coeficientes $\boldsymbol{\beta}$ siguen teniendo una interpretación económica de semi-elasticidad, es decir, representan cambios porcentuales en el precio de venta inicial deseado ante cambios unitarios en las variables.
Para este caso, se puede ver que las variables son significativas (ver tabla **3-11**), pero los coeficientes no representan el efecto total sobre el valor del precio deseado, puesto que existe un impacto asociado a cambios en el valor de las variables de los vecinos, así como cambios en el valor de las variables de la vivienda en sí, causados por la dependencia espacial. Por eso, es indispensable separar los efectos directos e indirectos de las variables. Los efectos directos representan como la variación de una variable explicativa afecta porcentualmente el valor del precio deseado de la i-ésima observación, mientras los efectos indirectos son la afectación que tiene cambios en las variables de los j vecinos sobre la observación i. Para este ejercicio especifico, por ejemplo, se observa en la tabla **3-12**, como ante cambios en la variable de número de baños, si aumenta en una unidad la cantidad de baños en la vivienda i, el precio deseado de la misma aumenta un 16 % aproximadamente, mientras, que si el número de baños aumenta en las j viviendas vecinas, el precio deseado aumenta un 4 % aproximadamente, manteniendo todas las demás condiciones constantes.

Tabla 3-11.: Coeficientes estimados modelo SAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros

Coeficientes	**Estimación**	**EE**	**Estadístico t**	**valor-p**
Intercepto	15.6200	0.1800	84.23	<0.0000
TipoCasa	-0.0490	0.0110	-4.45	<0.0000
m^2	0.0036	0.00006	56.80	<0.0000
Hab	-0.0632	0.0044	-14.52	<0.0000
Baños	0.1696	0.0051	33.19	<0.0000
Parqueaderos	0.0605	0.0047	12.80	<0.0000
Estrato4	0.3104	0.0118	26.47	<0.0000
Estrato5	0.3658	0.0139	26.38	<0.0000
Estrato6	0.5176	0.0152	34.09	<0.0000
Antigüedad5-10años	-0.0733	0.0150	-4.88	<0.0000
Antigüedad10-20años	-0.1275	0.0131	-9.63	<0.0000
AntigüedadMayor20años	-0.1235	0.0131	-9.45	<0.0000
ρ	0.1942	0.0090	21.56	<0.0000

Tabla 3-12.: Efectos directos e indirectos estimados modelo SAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros

Coeficientes	**Directo**	**Indirecto**	**Total**
TipoCasa	-0.0490	-0.0118	-0.0608
m^2	0.0036	0.0009	0.0044
Hab	-0.0632	-0.0152	-0.0784
Baños	0.1697	0.0408	0.2105
Parqueaderos	0.0605	0.0145	0.0751
Estrato4	0.3106	0.0746	0.3852
Estrato5	0.3660	0.0879	0.4539
Estrato6	0.5180	0.1244	0.6423
Antigüedad5-10años	-0.0733	-0.0176	-0.0909
Antigüedad10-20años	-0.1258	-0.0302	-0.1560
AntigüedadMayor20años	-0.1236	-0.0297	-0.1533

Modelo de rezago en el error (SEM)

Dado que la dependencia espacial puede estar asociada a variables no observadas del problema, es necesario al igual que hacen Urrea-Zora & Cárdenas-Mazo (2019), plantear un modelo que tome en cuenta esta situación. Se presenta:

$$\begin{aligned}\ln[Precio_i] =& \beta_0 + \beta_1 \text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 Baños_i + \beta_5 Parqueaderos_i \\ &+ \beta_6 \text{Estrato4}_i + \beta_7 \text{Estrato5}_i + \beta_8 \text{Estrato6}_i + \beta_9 \text{Antigüedad5-10años}_i \\ &+ \beta_{10} \text{Antigüedad10-20años}_i + \beta_{11} \text{AntigüedadMayor20años}_i \\ &+ u_i\end{aligned} \tag{3-5}$$

donde

$$u_i = \epsilon_i + \lambda \sum_{p=1}^{n} w_{ip} u_p, \quad \text{con} \quad \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$$

En este modelo, la variable Tipo es una variable dicotómica que indica si la vivienda es una casa o no; $m2_i$ son los metros cuadrados de la i-ésima vivienda; Hab_i, hace referencia al número de habitaciones de la vivienda; $Baños_i$, se refiere al número de baños de la vivienda; $Parqueaderos_i$ indica el número de parqueaderos con los que cuenta la vivienda; variables de estratos, siendo variables dicotómicas que indican si la i-ésima vivienda pertenece ya sea al estrato cuatro, cinco o seis; variables dicotómicas referentes a la antigüedad de la vivienda, indicando si la observación i pertenece al grupo de inmuebles con 5-10 años de antigüedad, 10-20 años de antigüedad o más de 20 años de antigüedad. Adicionalmente, el modelo cuenta con un término λ, que representa el efecto promedio que tienen todos los términos no adicionados al modelo por dificultad de medición u otras causas, de las observaciones vecinas.

Tras la realización del modelo con rezago en el error, se destaca que el parámetro estimado asociado a la dependencia espacial del error, $\hat{\lambda}$, que mide el grado de correlación entre los errores del modelo, no solo es significativo sino que su valor es 0.78, lo cual implica que existe una fuerte dependencia espacial entre los factores no considerados de valorización (anteriormente definidos en el marco teórico) de viviendas cercanas entre sí cuando se utiliza una matriz de pesos espaciales de contigüidad estandarizada.

En este caso, los coeficientes estimados para la regresión se interpretan como semi-elasticidades al igual que en el modelo lineal y de efectos mixtos, por ejemplo, si se desea saber el impacto de que la vivienda cuente con un baño extra, se tiene que el precio deseado aumenta en un 16 % aproximadamente. Por otro lado, la variable de tipo de vivienda no parece ser significativa, lo cual puede deberse a que la correlación entre los factores no considerados de valorización es más importante que la distinción entre casas y apartamentos (ver tabla **3-13**).

Siguiendo lo expuesto por Urrea-Zora & Cárdenas-Mazo (2019), se realizaron las pruebas de multiplicadores de Lagrange, tanto la prueba clásica como sus versiones robustas, para determinar si existe evidencia estadística para realizar un modelo que tenga en cuenta la dependencia espacial de la variable explicada y el error al tiempo, obteniendo que es sugerido estimar un modelo con que tenga en cuenta estas dos simultáneamente (ver tabla **3-14**).

Tabla 3-13.: Coeficientes estimados modelo SEM usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros

Coeficientes	**Estimación**	**EE**	**Estadístico t**	**valor-p**
Intercepto	19.1000	0.0237	806.79	<0.0000
TipoCasa	-0.0074	0.0190	-0.68	0.4952
$m2$	0.0035	0.00006	57.63	<0.0000
Hab	-0.0518	0.0043	-12.17	<0.0000
Baños	0.1605	0.0049	32.46	<0.0000
Parqueaderos	0.0598	0.0045	13.16	<0.0000
Estrato4	0.2784	0.0129	21.60	<0.0000
Estrato5	0.3515	0.0154	24.02	<0.0000
Estrato6	0.4764	0.0154	30.99	<0.0000
Antigüedad5-10años	-0.0675	0.0145	-4.65	<0.0000
Antigüedad10-20años	-0.1248	0.0126	-9.94	<0.0000
AntigüedadMayor20años	-0.1353	0.0126	-10.76	<0.0000
lambda	0.7815	0.0187	41.71	<0.0000

Tabla 3-14.: Pruebas de Lagrange para determinar la pertinencia de un modelo SARAR

Pruebas de Lagrange			
Prueba	Estadístico	grados de libertad	valor-p
Lagrange error	3891.80	1	<0.0000
Lagrange rezago	598.61	1	<0.0000
Lagrange Robusto error	3517.50	1	<0.0000
Lagrange Robusto rezago	224.35	1	<0.0000
Lagrange SARAR	4116.10	2	<0.0000

Modelo SARAR

Mediante las pruebas de multiplicadores de Lagrange se pudo determinar que es necesario tener en cuenta tanto la dependencia espacial con las otras observaciones, así como la dependencia espacial en el término de error, por ende, se plantea el siguiente modelo:

$$\begin{aligned}\ln[Precio_i] =& \rho \sum_{p=1}^{n} w_{ip} \ln[Precio_p] + \beta_0 + \beta_1 \text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 baños_i \\ &+ \beta_5 parqueaderos_i + \beta_6 \text{Estrato4}_i + \beta_7 \text{Estrato5}_i + \beta_8 \text{Estrato6}_i \\ &+ \beta_9 \text{antigüedad5-10años}_i + \beta_{10} \text{antigüedad10-20años}_i \\ &+ \beta_{11} \text{antigüedadMayor20años}_i + u_i\end{aligned} \tag{3-6}$$

donde

$$u_i = \epsilon_i + \lambda \sum_{p=1}^{n} w_{ip} u_p, \quad \text{con} \quad \epsilon_i \overset{\text{idd}}{\sim} N(0, \sigma^2)$$

Al igual que en los modelos presentados en las ecuaciones (3-4) y (3-5), se cuenta con variables de tipo dicotómicas, como lo son la variable Tipo casa$_i$, las variables de estratos Estrato4$_i$, Estrato5$_i$ y Estrato6$_i$, y las variables asociadas a la antigüedad del inmueble, antigüedad5-10años$_i$, antigüedad10-20años$_i$ y antigüedadMayor20 años$_i$. De igual manera, se incluyen variables con información sobre las características de la vivienda como lo son $m2_i$, Hab_i, $baños_i$ y $parqueaderos_i$.

En este caso, el modelo toma del modelo de la ecuación (3-4) la idea de que la dependencia espacial está vinculada al valor de observaciones vecinas, incluyendo por lo tanto un término ρ que mide el efecto de los vecinos sobre una observación i, ponderando cada observación por un w_{ip} que es el elemento asociado a la i-ésima fila y a la p-ésima columna de la matriz de pesos espaciales. De igual forma, incluye la dependencia espacial de los errores del modelo al igual que el modelo de la ecuación (3-5), mediante el término λ.

Se resalta que el coeficiente de la variable indicativa del tipo de vivienda (Tipo casa), al igual que en el modelo SEM, no es significativa (ver tabla **3-15**). Adicionalmente, como se ha podido observar en las diferentes tablas de los modelos espaciales, existe una mayor correlación entre los factores no observados de valorización que entre el logaritmo de los valores observados de la variable dependiente, indicando que el precio deseado por los vendedores es mayormente influenciado por factores difíciles de medir cuyo efecto solo es apreciable de manera conjunta a través de los errores.

Tal como ocurre en el modelo SAR, para tener una visión completa del efecto de las variables, es necesario observar qué tanto del efecto total es atribuible a efecto directo y qué tanto a un efecto indirecto. De igual manera, al tener un coeficiente ρ muy pequeño, la mayor cantidad del efecto total es causado por los efectos directos, es decir, es más importante las propias

características de la vivienda para el valor deseado de venta que las características de las viviendas vecinas (ver tabla **3-16**).

Tabla 3-15.: Coeficientes estimados modelo SARAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros

Coeficientes	**Estimación**	**EE**	**Estadístico t**	**valor-p**
Intercepto	18.59	0.2244	82.83	<0.0000
TipoCasa	-0.0075	0.0109	-0.69	0.4893
m^2	0.0035	0.00006	57.66	<0.0000
Hab	-0.0524	0.0043	-12.29	<0.0000
Baños	0.1610	0.0049	32.54	<0.0000
Parqueaderos	0.0596	0.0045	13.13	<0.0000
Estrato4	0.2785	0.0129	21.63	<0.0000
Estrato5	0.3518	0.0146	24.05	<0.0000
Estrato6	0.4770	0.0154	31.04	<0.0000
Antigüedad5-10años	-0.0672	0.0145	-4.63	<0.0000
Antigüedad10-20años	-0.1249	0.0126	-9.94	<0.0000
AntigüedadMayor20años	-0.1354	0.0126	-10.75	<0.0000
rho	0.0253	0.0111	2.28	0.0227
lambda	0.7637	0.0204	37.65	<0.0000

Tabla 3-16.: Efectos directos e indirectos estimados modelo SARAR usando una matriz de pesos espaciales igual a vecinos estandarizados, donde se considera vecino si la distancia es menor a 1000 metros

Coeficientes	**Directo**	**Indirecto**	**Total**
Tipo casa	-0.0075	-0.0002	-0.0077
m^2	0.0035	0.00009	0.0036
Hab	-0.0524	-0.0014	-0.0538
Baños	0.1609	0.0042	0.1651
Parqueaderos	0.0596	0.0015	0.0612
Estrato 4	0.2785	0.0072	0.2858
Estrato 5	0.3518	0.0091	0.3609
Estrato 6	0.4770	0.0124	0.4894
Antigüedad 5-10 años	-0.0672	-0.0017	-0.0690
Antigüedad 10-20 años	-0.1249	-0.0032	-0.1281
Antigüedad mayor 20 años	-0.1354	-0.0035	-0.1389

Modelo con filtrado espacial mediante vectores propios

Existen otras formas de modelar la dependencia espacial de los datos, por lo cual, se propone:

$$\begin{aligned}\ln[Precio_i] =& \beta_0 + \beta_1 \text{TipoCasa}_i + \beta_2 m2_i + \beta_3 Hab_i + \beta_4 baños_i + \beta_5 parqueaderos_i \\ &+ \beta_6 \text{Estrato4}_i + \beta_7 \text{Estrato5}_i + \beta_8 \text{Estrato6}_i + \beta_9 \text{antigüedad5-10años}_i \\ &+ \beta_{10} \text{antigüedad10-20años}_i + \beta_{11} \text{antigüedadMayor20años}_i \\ &+ \sum_{a=1}^{A} E_{ia} + \epsilon_i, \quad \text{con} \quad \epsilon_i \overset{\text{idd}}{\sim} N(0, \sigma^2)\end{aligned} \tag{3-7}$$

En este caso E_{ia} es el i-ésimo valor del a-ésimo vector propio de la matriz de pesos espaciales tras sustraer la media tanto de la fila como de la columna respectiva a cada dato. En la ecuación anterior, se toma A=157 vectores propios como parte de un modelo semiparamétrico en el cual estos son los encargados de absorber la dependencia espacial. En este caso, la selección de los vectores se realiza tomando solo aquellos vectores que tengan un valor propio mayor a cero de la matriz de pesos espaciales de distancias inversas doblemente centrada. Con estos se realiza una regresión, y siguiendo lo expuesto por Griffith (2003), se realiza una selección de los vectores propios mediante un stepwise de selección backward, eliminando vectores hasta encontrar un mínimo de vectores que permite minimizar el AIC del modelo, iniciando con 225 vectores y un AIC de 3787.96, terminando con los 157 vectores mencionados anteriormente y un AIC de 3686.32.

Tabla 3-17.: Coeficientes estimados de las variables usando filtrado espacial, seleccionando vectores propios mediante stepwise backward

Coeficientes	**Estimación**	**SE**	**Estadístico t**	**valor-p**
Intercepto	19.18	0.0160	1199.97	<0.0000
Tipo Casa	0.0445	0.0103	4.32	<0.0000
$m2$	0.0033	0.00006	57.26	<0.0000
Hab	-0.0363	0.0040	-9.08	<0.0000
baños	0.1422	0.0047	30.55	<0.0000
parqueaderos	0.0595	0.0042	14.03	<0.0000
Estrato4	0.2245	0.01242	18.08	<0.0000
Estrato5	0.2711	0.0141	19.26	<0.0000
Estrato6	0.3412	0.0149	22.92	<0.0000
antigüedad5-10años	-0.0526	0.0135	-3.90	<0.0000
antigüedad10-20años	-0.1119	0.0117	-9.56	<0.0000
antigüedadMayor20años	-0.1348	0.0117	-11.52	<0.0000

Del modelo se resalta que el coeficiente estimado para la variable que índica si la vivienda es casa o apartamento es positivo, siendo el único modelo, junto al de efectos mixtos basado en

coordenadas, que determina que las casas poseen un precio significativamente mayor a los apartamentos. De igual manera se estiman coeficientes tanto para las variables como para los vectores propios. Los primeros se interpretan como semi-elasticidades (ver tabla **3-17**), mientras los últimos, carecen de interpretación puesto que su principal función es absorber la dependencia espacial.

3.5. Comparación de modelos

En esta sección se presentan algunas medidas de bondad de ajuste de los datos, esto con el fin de seleccionar aquel modelo que mejor desempeño tiene dada la información con la que se disponen para este ejercicio especifico. Estas medidas están resumidas en la tabla **3-18**, en la cual se reportan el criterio de información de Akaike (AIC, por sus siglas en inglés), el criterio de información bayesiana (BIC, por sus siglas en inglés), el valor de la log-verosimilitud (LV, por sus siglas en inglés), el tiempo utilizado para la estimación de los parámetros, los errores cuadráticos medios de entrenamiento y de testeo, estos dos últimos tras la separación del conjunto de datos, en un conjunto para entrenamiento conformado por el 70 % de las observaciones totales y el otro 30 % para el testeo de los modelos.
Se aclara, que las estimaciones de los modelos se llevaron a cabo mediante la versión 4.1.2 del software R (R Core Team, 2023) y la versión 2021.09.1 de RStudio (RStudio Team, 2023). Adicionalmente, se debe tener en cuenta que el tiempo de cómputo de los diferentes modelos está sujeto a las características del dispositivo usado. En este caso, se utilizó un computador portátil de 8Gb de memoria RAM para la realización de este trabajo.
Por un lado, al enfocarse en las medidas calculadas, es posible ver en la tabla **3-18** que en términos de AIC, ECM de testeo (ECMTs) y entrenamiento (ECMEn), el modelo de filtrado espacial está dentro de los de mejor desempeño de todos. Sin embargo, utilizando el criterio de información bayesiana, el cual penaliza de manera más fuerte el número de parámetros estimados, se puede ver que el modelo que mejor se adapta al problema trabajado es el modelo de efectos mixtos basado en coordenadas, siendo este uno de los que menor tiempo de cómputo utilizó para la estimación de sus parámetros.
Por otro lado, se debe tener en cuenta que el problema central de este trabajo es generar una herramienta de valoración de los inmuebles, con el objetivo de tener una recomendación sobre el precio inicial de venta, razón por la cual, se decidió ponderar las diferentes métricas usadas, utilizando una manera más objetiva para la selección del mejor modelo, dando esta denominación a aquel modelo que tenga la puntuación final más baja, dado que la ponderación se da a manera de ranking, donde uno es la puntuación para la métrica más importante y cinco la menos importante, puntualizando el hecho que no se pondera la log-verosimilitud puesto que esta se utiliza para el cálculo del AIC y el BIC.
La importancia dada a cada métrica es dada por las siguientes razones:

- Se considera que el modelo debe tener una precisión alta en su predicción, por lo

tanto, el error cuadrático medio en el conjunto de testeo se considera la métrica más importante.

- Se debe conseguir, en la medida de lo posible, cumplir con el principio de parsimonia, por lo cual, una métrica que penaliza de manera más fuerte el número de parámetros a estimar como el BIC se considera que es la siguiente en importancia.
- El tiempo de cómputo está relacionado a la limitación que se pueda tener en términos de tecnología, donde un tiempo más alto también implica que es posible que el modelo no sea estimable con menos capacidad de memoria o procesamiento, por lo cual se toma el tiempo de estimación como la tercera métrica más importante.
- El AIC no castiga de manera rigurosa el número de parámetros como el BIC, por lo cual se toma está métrica como la segunda de menor importancia.
- Debido a que los algoritmos se basan en la disminución del error entre los datos y las predicciones, el error cuadrático medio de entrenamiento se pondera como la métrica de menor importancia, en contraste con el error cuadrático medio de testeo.

Ahora bien, por cada métrica se realizó un ranking donde 1 es el mejor desempeño y 7 el peor desempeño, esto se puede ver en la tabla **3-19**. Posterior a esta clasificación, se pondera el ranking por el valor de cada métrica y se suman los resultados por cada modelo, donde el menor valor dentro de esta suma pertenece al modelo con un desempeño más parejo en todas las métricas, especialmente en las de mayor importancia.

Tabla 3-18.: Tabla de reporte de las métricas calculadas para los modelos utilizados

Modelos	LV	AIC	BIC	Tiempo	ECMEn	ECMTs
Lineal	-3038.90	6103.63	6194.22	0.02 seg	0.1749	0.1790
Mixto Barrios	-2086.04	4204.08	4315.54	0.52 seg	0.1143	0.1168
Mixto Coords.	-1962.42	3954.84	4059.36	1 min 34seg	0.0538	0.0718
SAR	-2946.19	5920.40	6017.95	33 min 54 seg	2.3812	2.3967
SEM	-2385.32	4798.60	4896.19	20 min 6 seg	0.1339	0.1315
SARAR	-2352.28	4734.60	4839.09	1 hrs 54 min	0.1718	0.1710
Filtrado espacial	-1673.16	3686.32	4870.95	1 hrs 23 min*	0.1000	0.0936
*Se toma en cuenta el tiempo para la descomposición de la matriz de pesos espaciales en vectores propios usados en la regresión y la selección backward para el modelo final						

Tabla 3-19.: Tabla comparativa entre modelos ponderando las métricas

Modelos	AIC (4)	BIC (2)	Tiempo (3)	ECMTs (1)	ECMEn (5)	resultados
Lineal	7	7	1	6	6	81
Mixto Barrios	3	2	2	3	3	34
Mixto Coords.	2	1	3	1	1	25
SAR	6	6	5	7	7	93
SEM	5	5	4	4	4	66
SARAR	4	3	7	5	5	73
Filtrado espacial	1	4	6	2	2	42

Dado lo expuesto, se recomienda el uso del modelo mixto basado en coordenadas para la estimación del precio de venta inicial, dado que mediante sus efectos aleatorios, este recoge aquellos factores de valorización que son difíciles o imposibles de medir, ofreciendo dentro de sus ventajas la posibilidad de predecir el precio de viviendas fuera de la muestra recolectada siempre y cuando, se cuente con la información asociada a los efectos fijos y las coordenadas de la vivienda, y se tenga presente las limitaciones expuestas sobre el uso del modelo fuera del rango de las observaciones.
Se resalta también que el tiempo de cómputo fue de los más rápidos, por lo cual se esperaría que el modelo pueda ser estimado en equipos con menor capacidad de procesamiento o en su defecto, se pueda adaptar a una mayor cantidad de observaciones. Adicionalmente, en términos de parsimonia y precisión en el testeo, el modelo estuvo entre los mejores resultados en ambos, razón por la cual se usa este modelo para determinar el valor asociado a los factores no considerados de valorización de los barrios de Bogotá que están dentro de la muestra recolectada, siguiendo lo expuesto por Urrea-Zora & Cárdenas-Mazo (2019).

3.6. Factores no considerados de valorización de los barrios de Bogotá

Tal como se expuso en el marco teórico, es factible agrupar dentro de los errores del modelo aquellos factores que afectan el valor deseado por el propietario de la vivienda, ya sea porque dentro de estos se encuentran variables no observadas, de difícil medición o porque dentro de estos se agrupan razones emocionales y de especulación, que afectan a este precio.
En este caso, se supondrá que estos factores se relacionan de una u otra forma con la comodidad y los beneficios que ofrece vivir en un lugar, por lo cual es posible usar los efectos mixtos asociados al intercepto para encontrar en qué barrios se cuenta con una menor cantidad de estos factores, entendiendo que estos afectan el precio base deseado por el vendedor.

Los barrios incluidos dentro de la muestra que presentan valores por debajo de cero serían aquellos cuyos factores no considerados de valorización están por debajo de la media de la ciudad, y que por ende, podrían necesitar obras de intervención, con el fin de nivelar estos factores a lo largo de la ciudad (figura **3-6**).

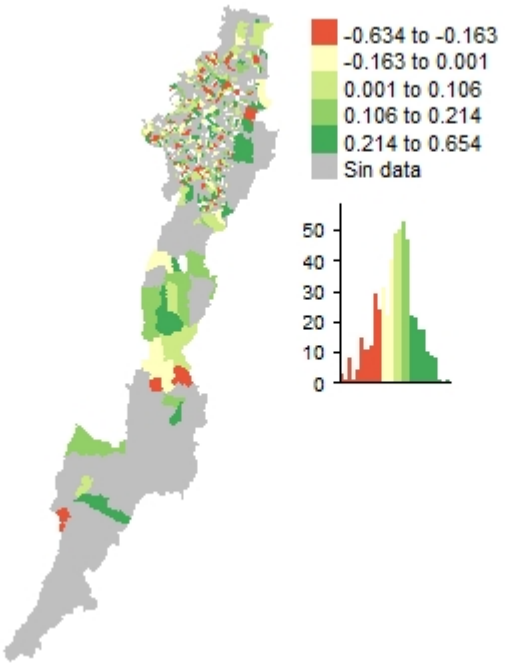

Figura 3-6.: Valor de los factores no considerados de valorización

En ese orden de ideas, se espera que mediante el modelo se pueda llegar a estimar el precio inicial de negociación de una vivienda, ayudando tanto a compradores y vendedores a tener un precio de referencia. De igual manera, se espera que mediante los efectos aleatorios del modelo se pueda generar algún ranking o herramienta que permita a las personas interesadas en comprar vivienda, una visión sobre los mejores barrios, y darse una idea inicial sobre las condiciones del mismo (barrio de condiciones inferiores al promedio, en el promedio o superior a este).

Por otro lado, se espera que, mediante la metodología usada para estimar estos factores no considerados de valorización en este trabajo, se pueda brindar una ayuda a los entes distritales y territoriales para generar herramientas y métricas que permitan focalizar esfuerzos, intervenciones en infraestructura, proyectos sociales, entre otros, influyendo positivamente en la efectividad de las decisiones y en el manejo eficiente de los recursos públicos.

4. Conclusiones y limitaciones

4.1. Conclusiones

Con respecto a la dependencia espacial, se puede observar que los errores del modelo lineal estimado por mínimos cuadrados ordinarios están correlacionados positivamente, indicando que viviendas de precios altos se rodean de viviendas de precios altos. Adicionalmente, en los modelos autorregresivos SAR, SEM y SARAR, así como el modelo de efectos mixtos basado en coordenadas, los componentes espaciales son significativos, lo que afianza la idea de que es necesario modelar la dependencia espacial de los datos.

Por su parte, los modelos muestran una clara tendencia a valorar de menor manera un inmueble a medida que presenta una mayor antigüedad. De manera contraria, se aprecia que a mayor el nivel del estrato socioeconómico, mayor es el precio base deseado por el vendedor. Se resalta que, en todos los modelos, la variable asociada al número de parqueaderos de la vivienda es significativa y positiva, siendo congruente con los constantes cambios que ha tenido la norma del pico y placa en la ciudad de Bogotá, al igual que la tendencia del aumento del número de vehículos.

Al enfocarse en el resultado de la variable de tipo de vivienda (casa o apartamento), solo con el modelo mixto basado en coordenadas y con el modelo de filtrado espacial esta variable presenta una estimación de valor positivo, siendo una posible explicación la manera en que se estiman los efectos aleatorios y fijos en el primero, mientras que las variables sintéticas en el segundo pueden absorber alguna clase de efecto que convierte el signo de la estimación de negativo en otros modelos a positivo en estos dos.

Pasando a la comparación de modelos, al analizar los resultados del error cuadrático medio de los distintos modelos, a excepción del modelo SAR, todos presentan valores inferiores a la unidad, lo que sugiere que los modelos son bastante precisos para predecir el precio deseado por el vendedor. Igualmente, al valorar cada modelo según las ponderaciones establecidas para cada medida se encuentra que el modelo de efectos espaciales basado en coordenadas es el mejor modelo bajo los datos recolectados para este ejercicio, razón por la cual, se utilizó para estimar los factores no considerados de valorización de cada observación i, utilizando esta información para construir el mapa de estos factores mostrando la media de este valor por los barrios de Bogotá dentro de la muestra.

Para finalizar, se observa que la metodología empleada, siguiendo las ideas de Dubin (1992) y Urrea-Zora & Cárdenas-Mazo (2019), permite estimar factores que afectan el valor de las viviendas y que por diferentes razones no son posibles estimar directamente (factores asociados

a la seguridad, movilidad, cercanía a parques y centros comerciales, entre otros), brindando la oportunidad de visualizar cómo se comportan éstos a través de la ciudad, siendo que para este caso, se agregaron los resultados por los barrios con los cuales se contaba en la muestra extraída a través de los portales web utilizados, resaltando que con el modelo de efectos mixtos basado en coordenadas es posible además predecir lugares fuera de la muestra recolectada.

4.2. Limitaciones

Dentro de las limitaciones se menciona el hecho de que la extracción de las coordenadas no es exacta, lo cual puede generar errores. De igual manera, se debe tener en cuenta que la extracción de la información de los portales web tiende a ser demorada debido a que existen límites sobre la cantidad de anuncios que se pueden obtener antes de que el sitio web detecte que se está haciendo web scraping y detenga la ejecución del programa mediante el bloqueo de componentes que no dejan cargar correctamente la página y por ende generar errores en el programa creado.

Una limitación importante es el equipo tecnológico usado para realizar el cómputo de los modelos, ya que es muy posible que modelos como el SARAR y el de filtrado espacial (que requieren una gran cantidad de procesamiento), ante una mayor cantidad de datos no sea posible generar estimaciones de los parámetros de estos modelos o en su defecto, consuman mucho más tiempo, aumentando costos asociados a la energía eléctrica.

Por parte de los modelos, cuando los efectos mixtos se basan en datos de barrios se limita solo a aquellos barrios dentro de la muestra, siendo imposible estimar un efecto aleatorio de un barrio que no se encuentre dentro de los datos recolectados, siendo esto posible con el modelo de efectos mixtos basado en coordenadas. Sin embargo, es necesario resaltar que este último se basa en una técnica de interpolación, por lo cual, para puntos fuera del rango de la longitud y latitud de la muestra, no se cuenta con predicciones confiables del efecto aleatorio. De igual manera, se resalta que la distribución de los efectos mixtos tiene un comportamiento normal solo de manera asintótica.

5. Recomendaciones

Se puede destacar que para la recolección de datos mediante técnicas de web scraping, el lenguaje de programación Python (Van Rossum & Drake (2009)), cuenta con diferentes librerías que permiten la extracción de información de una manera razonablemente fácil. De igual modo, dentro de la extracción de información, se destaca que cada portal web tiene su propia estructura, razón por la cual, fue necesario realizar un programa por cada portal utilizado en el trabajo.

Como recomendación, se propone incluir otros portales enfocados en la promoción de ofertas inmobiliarias y buscar la manera de incluir más datos de los estratos 1 y 2, para tener una mejor representación de todos los estratos en general.

Se incentiva que en futuros trabajos se pueda realizar una investigación de tipo espacio-temporal para poder observar cómo cambian o han cambiado las estimaciones de estos modelos a través del tiempo, siendo esto algo que puedan usar los entes territoriales y las personas para visualizar si han ocurrido cambios en los factores no considerados de valorización (sea esto de manera positiva nivelando los mismos o de manera negativa, aumentando la brecha entre barrios).

Se espera que este trabajo sirva como guía para aquellas personas que deseen trabajar precios hedónicos, mostrando de una manera detallada cómo se obtuvieron los datos empleados en este trabajo. Adicionalmente, se espera que la metodología empleada sirva como insumo para posteriores investigaciones, así como para incentivar a otras personas a realizar trabajos similares en otras ciudades.

A. Anexo: Estadísticas y gráficas descriptivas adicionales de las variables de la base de datos

Se presentan en este anexo algunas gráficas y tablas que permiten conocer de manera descriptiva variables que están dentro de la base de datos y que permiten un entendimiento más profundo de las variables sin llegar a influir en el entendimiento del trabajo en sí.

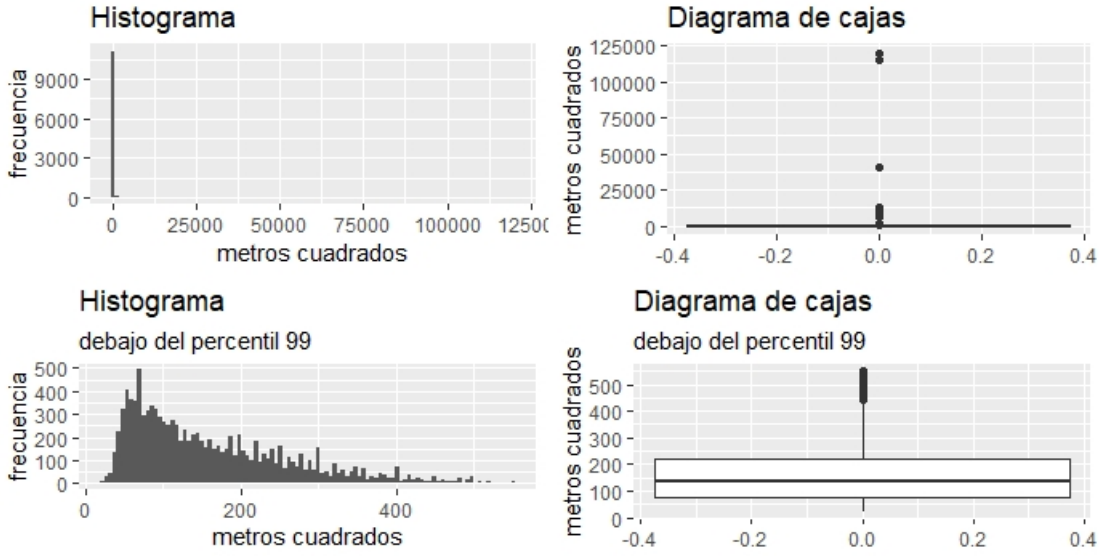

Figura A-1.: Histograma y diagrama de caja para la variable metros cuadrados y para la base reducida debajo del percentil 99.

B. Anexo: Código para realizar web scraping en finca raíz mediante Python

```
from selenium import webdriver
from bs4 import BeautifulSoup
import pandas as pd
import time
import geocoder
import re

### abrir firefox con python
### webdriver.firefox() abre firefox descargar geckodriver firefox

driver=webdriver.Firefox(executable_path=r'/home/daniel/proyecto/geckodriver')

#url para scrapear aptos
url_to_scrap='https://www.fincaraiz.com.co/apartaestudios-apartamentos-casas/venta'

driver.get(url_to_scrap)

### codigo html pagina scrapear
html_code=driver.page_source


### crear objeto tipo soup para parsear
soup=BeautifulSoup(html_code,'lxml')

#tag article, cada anuncio esta dentro de una etiqueta article
all_house_li=soup.find_all("article")

#para uno solo
#obtener vinculo inmueble
```

```
uno=all_house_li[0]
b="https://www.fincaraiz.com.co"+uno.a["href"]

#obtener precio inmueble
p_inm=float(uno.a.contents[2].find("b").text.split("$")[1].replace(".",""))

metraje=""
cuartos=""
baño=""
parqueadero=""

if uno.find("section").find_all(string=re.compile("m"))==[]:
    metraje=""
else:
    metraje=float((uno.find("section")
                   .find_all(string=re.compile("m"))[0]
                   .split("m")[0]))

if uno.find("section").find_all(string=re.compile("ha"))==[]:
    cuartos=""
else:
    cuartos=float((uno.find("section")
                   .find_all(string=re.compile("ha"))[0]
                   .split("h")[0]))

if uno.find("section").find_all(string=re.compile("ba"))==[]:
    baño=""
else:
    baño=float((uno.find("section")
                .find_all(string=re.compile("ba"))[0]
                .split("b")[0]))

if uno.find("section").find_all(string=re.compile("pa"))==[]:
    parqueadero=""
else:
    parqueadero=float((uno.find("section")
                       .find_all(string=re.compile("pa"))[0]
                       .split("p")[0]))
```

```
tipo_inm=uno.find("footer").find("b").text.split(" ")[0]

#Función para las coordenadas
coord=[]
def latlong (ub):
    g=geocoder.osm(ub)
    coordenadas=g.latlng
    print(coordenadas)
    coord.append(coordenadas)

# Aplicar para las ubicaciones obtenidas en los anuncios
contador=0
for ub in data["Ubicacion"]:
    latlong(ub)
    print(contador)
    contador=contador+1
    time.sleep(3)
```

C. Anexo: Código para la estimación de los modelos en R

```
    # Preparacion del espacio ------------------------------------------------

rm(list=ls())
setwd("~/maestria_unal")
# Librerias ---------------------------------------------------------------

library(gridExtra)
library(spatialEco)
library(sf)
library(tidycensus)
library(tmap)
library(spdep)
library(tigris)
library(rmapshaper)
library(broom)
library(car)
library(ade4)
library(spatialreg)
library(knitr)
library(geojsonio)
library(stargazer)
library(readr)
library(geojsonR)
library(sp)
library(rgdal)
library(rgeos)
library(dbscan)
library(spgwr)
library(lattice)
library(nlme)
library(lmerTest)
```

```
library(EMSaov)
library(GGally)
library(gdata)
library(olsrr)
library(MASS)
library(lmtest)
library(normtest) ###REALIZA 5 PRUEBAS DE NORMALIDAD###
library(nortest) ###REALIZA 10 PRUEBAS DE NORMALIDAD###
library(tidyverse)
library(spaMM)

# Datos ------------------------------------------------------------------

consolidado <- read_csv("consolidado.csv")

consolidado$Tipo <- factor(consolidado$Tipo)
consolidado$estratos <- factor(consolidado$estratos)
consolidado$antiguedad <- factor(consolidado$antiguedad)
summary(consolidado)

consolidado2 <- consolidado[consolidado$Precio>=50000000 &
                             consolidado$Precio<=2841200000,]

consolidado3 <- consolidado2[consolidado2$Metros_cuad>=20 &
                              consolidado2$Metros_cuad<= 550,]
spdf <- geojson_read('avaluo.json', what = "sp")
manz_bogota <- spTransform(spdf,CRS('+init=epsg:9377'))

spdf <- geojson_read('DAILoc.geojson', what = "sp")
crimen_bogota <- spTransform(spdf,CRS('+init=epsg:9377'))


spdf <- geojson_read('DAISCAT.geojson', what = "sp")
barrio <- spTransform(spdf,CRS('+init=epsg:9377'))


spdf <- geojson_read('DAIUPZ.geojson', what = "sp")
upz <- spTransform(spdf,CRS('+init=epsg:9377'))

viviendas <- consolidado3
```

```
coordinates(viviendas) <- c("longitud","latitud")
proj4string(viviendas) <- CRS('+proj=longlat +datum=WGS84')
viviendas <- spTransform(viviendas,CRS('+init=epsg:9377'))
viviendas

vivi_manz <- point.in.poly(viviendas,manz_bogota)
viavrobos <- point.in.poly(vivi_manz,crimen_bogota)
masupz <- point.in.poly(viavrobos,barrio)
basef <- point.in.poly(masupz,upz)

#Variable de hurtos totales 21
basef@data$hurtos21<-basef@data$CMHP21CONT.x+basef@data$CMHR21CONT.x+
  basef@data$CMHA21CONT.x+basef@data$CMHB21CONT.x+
  basef@data$CMHCE21CONT.x+basef@data$CMHM21CONT.x+
  basef@data$CMHF21CONT.x+basef@data$CMHC21CONT.x

vusadas <- as.data.frame(basef)

vusadas <- vusadas%>% select(Precio, Tipo, Metros_cuad, Hab, baños,
                             parqueaderos, estratos, antiguedad,
                             AVALUO_COM, AVALUO_CAT,hurtos21, CMH21CONT.x,
                             MANZANA_ID, CMNOMSCAT,CMNOMLOCAL,CMNOMUPLA,
                             coords.x1,coords.x2)

vusadas <- na.omit(vusadas)
vusadas$estratos <- as.numeric(vusadas$estratos)
vusadas[which(vusadas$estratos=="1" | vusadas$estratos=="2" |
                vusadas$estratos=="3"),"estratos"] <- "base"
vusadas$estratos <- factor(vusadas$estratos)
vusadas$estratos <- relevel(vusadas$estratos,"base")

vusadas$CMNOMSCAT <- factor(vusadas$CMNOMSCAT)
vusadas$CMNOMUPLA <- factor(vusadas$CMNOMUPLA)
vusadas$logmetcuad <- log(vusadas$Metros_cuad)
summary(vusadas)

# MODELOS ------------------------------------------------------------------

lineal <- lm(log(Precio)~Tipo+Metros_cuad+Hab+baños+parqueaderos+
               estratos+antiguedad,
```

```
             data=vusadas)
summary(lineal)

#matriz 1 si esta entre 0 y 500 metros; cero en caso contrario
#Matriz necesaria para estimar modelos espaciales

distmax <- dnearneigh(residuales2,0,1000)
distnbls <-nb2listw(distmax, style = "W",zero.policy = T)

viviendas <- vusadas
coordinates(viviendas) <- c("coords.x1","coords.x2")
proj4string(viviendas) <- CRS('+init=epsg:9377')

#Modelo SAR
y_lag3 <- lagsarlm(log(Precio)~Tipo+Metros_cuad+Hab+baños+parqueaderos+
                     estratos+antiguedad,
                   data=viviendas@data, listw=distnbls,
                   zero.policy=T)

#Modelo SEM
error_lag3 <- errorsarlm(log(Precio)~Tipo+Metros_cuad+Hab+baños+
                           parqueaderos+estratos+antiguedad,
                         data=viviendas@data, listw=distnbls,
                         zero.policy=T)

#Modelo SARAR
ev <- eigenw(distnbls)
sarar_mod3 <- sacsarlm(log(Precio)~Tipo+Metros_cuad+Hab+baños+
                          parqueaderos+estratos+antiguedad,
                       data=viviendas@data, listw=distnbls,
                       zero.policy=T,
                       control=list(pre_eig1=ev, pre_eig2=ev))

#Modelo con filtrado espacial
y=log(viviendas@data$Precio)
x=viviendas@data[,c(3:6)]
x$tipo <- ifelse(viviendas@data$Tipo=="Casa",1,0)
x$ant5_10 <- ifelse(viviendas@data$antiguedad=="Entre 5 y 10 años",1,0)
x$ant10_20 <- ifelse(viviendas@data$antiguedad=="Entre 10 y 20 años",1,0)
x$ant20 <- ifelse(viviendas@data$antiguedad=="Más de 20 años",1,0)
```

```
x$s4 <- ifelse(viviendas@data$estratos=="4",1,0)
x$s5 <- ifelse(viviendas@data$estratos=="5",1,0)
x$s6 <- ifelse(viviendas@data$estratos=="6",1,0)
inicio <- Sys.time()
coords=coordinates(viviendas)
eigf=meigen(coords = coords)
esfD <- esf(y=y,x=x,meig=eigf, fn="aic")

#Modelo mixto basado en barrios
Modelo1=lme(log(Precio)~Tipo+logmetcuad+Hab+baños+parqueaderos+
              estratos+
              antiguedad,
            data=vusadas
            ,random=~1+logmetcuad|CMNOMSCAT)

#Modelo mixto basado en coordenadas
m_spamm0 <- fitme(log(Precio)~Tipo+Metros_cuad+Hab+baños+parqueaderos+
                  estratos+antiguedad+
                  Matern(1|coords.x1+coords.x2),
                data=vusadas4,
                method = "REML",
                fixed=list(nu=0.5))
```

D. Anexo: Graficas de transformaciones probadas para el modelo lineal mixto basado en barrios

Debido a que no es posible validar la hipótesis sobre el supuesto de normalidad de los errores del modelo, se probaron diferentes transformaciones con el fin de obtener una nueva variable que cumpliera con este supuesto.

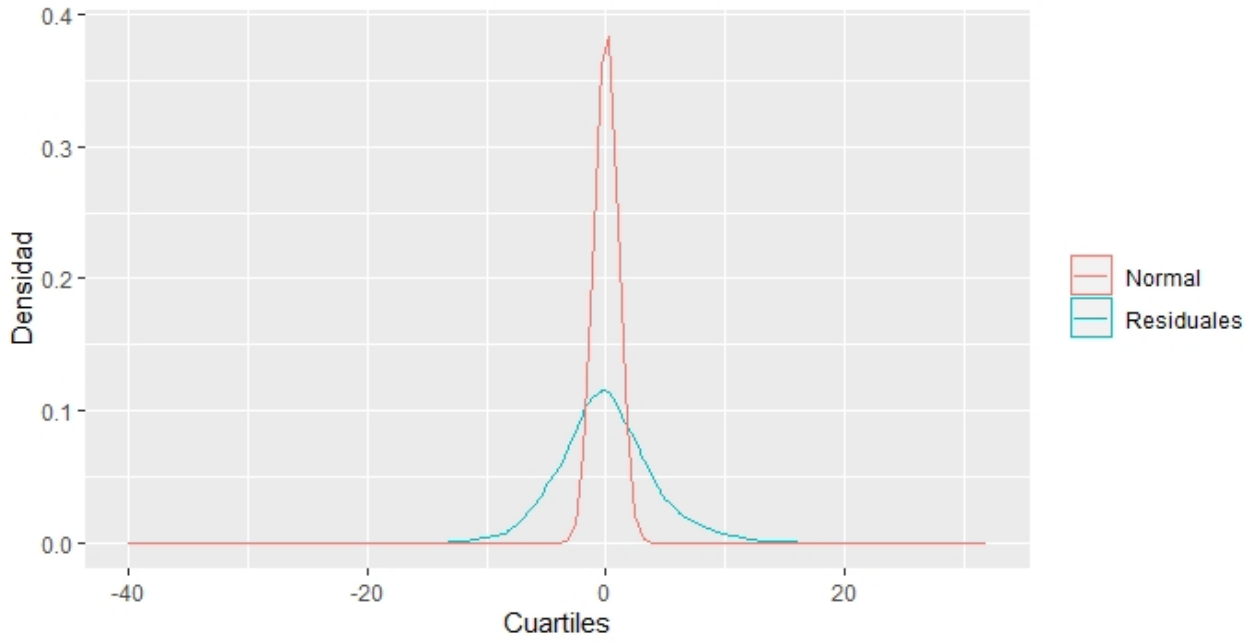

Figura D-1.: Distribución estimada de los residuales condicionados estándar del modelo (azul) después de transformar la variable Precio mediante Box-Cox con un $\lambda = 0{,}1414$. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

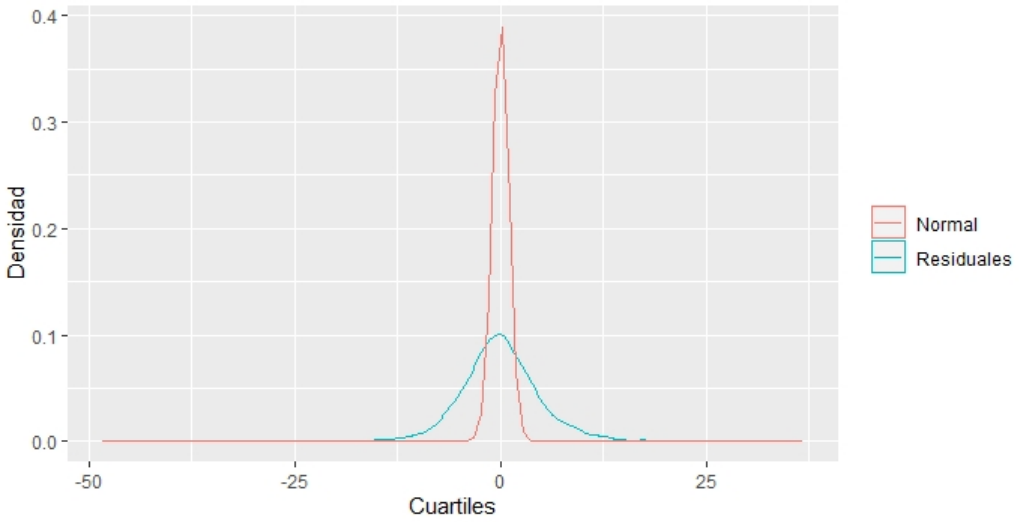

Figura D-2.: Distribución estimada de los residuales condicionados estándar del modelo (azul) después de transformar la variable logaritmo del Precio mediante Box-Cox con un $\lambda = 2$. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

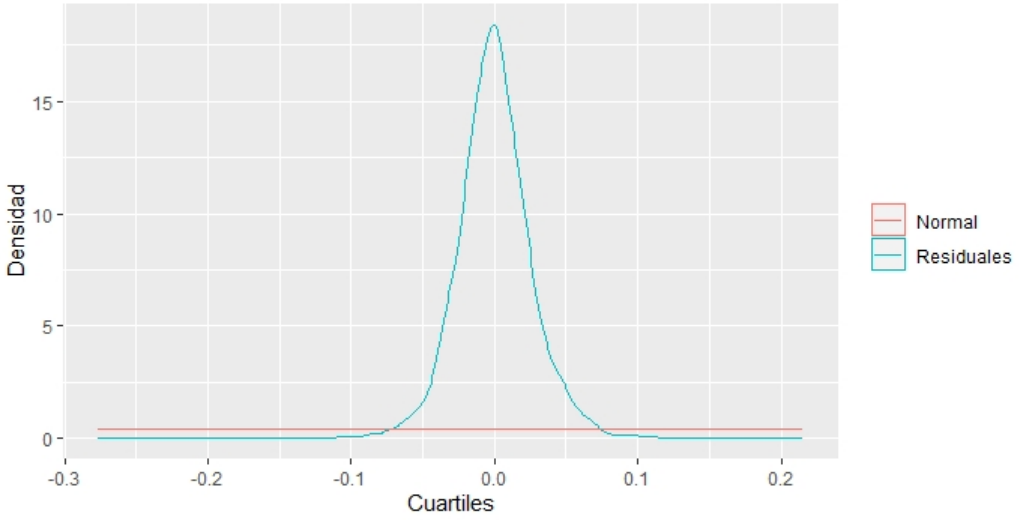

Figura D-3.: Distribución estimada de los residuales condicionados estándar del modelo (azul) después de aplicar raíz cuadrada a la variable logaritmo del Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

E. Anexo: Graficas de transformaciones probadas para el modelo lineal mixto basado en coordenadas

Debido a que no es posible validar la hipótesis sobre el supuesto de normalidad de los errores del modelo, se probaron diferentes transformaciones con el fin de obtener una nueva variable que cumpliera con este supuesto.

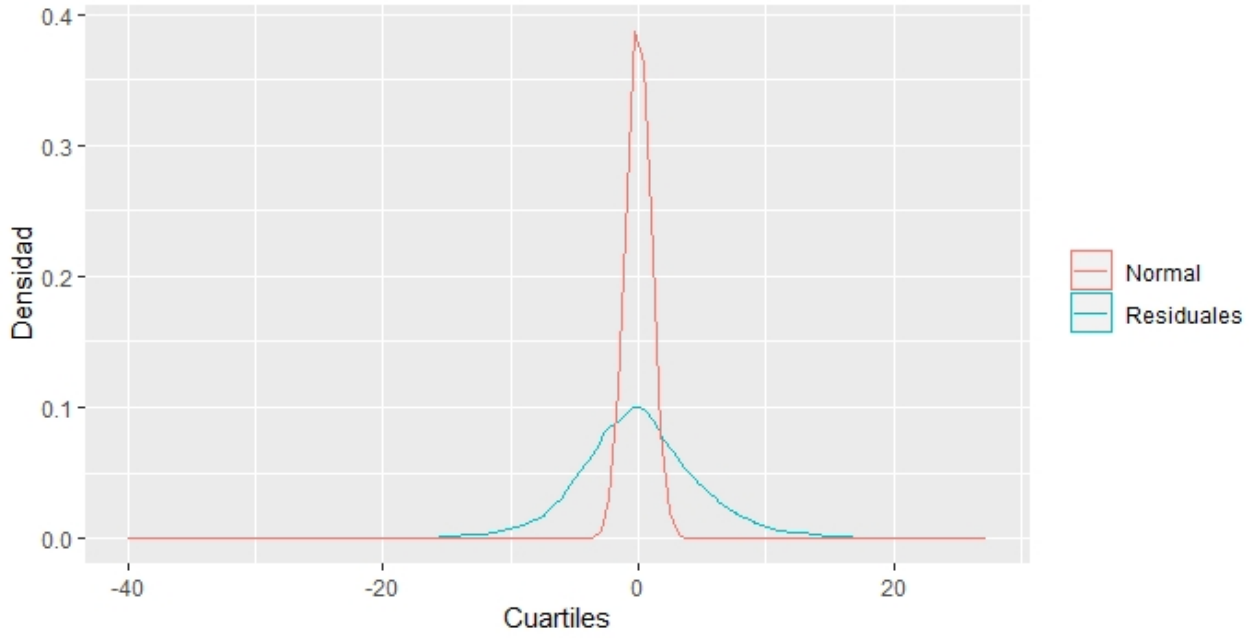

Figura E-1.: Distribución estimada de los residuales condicionados estándar del modelo (azul) después de transformar la variable Precio mediante Box-Cox con un $\lambda = 0{,}1414$. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

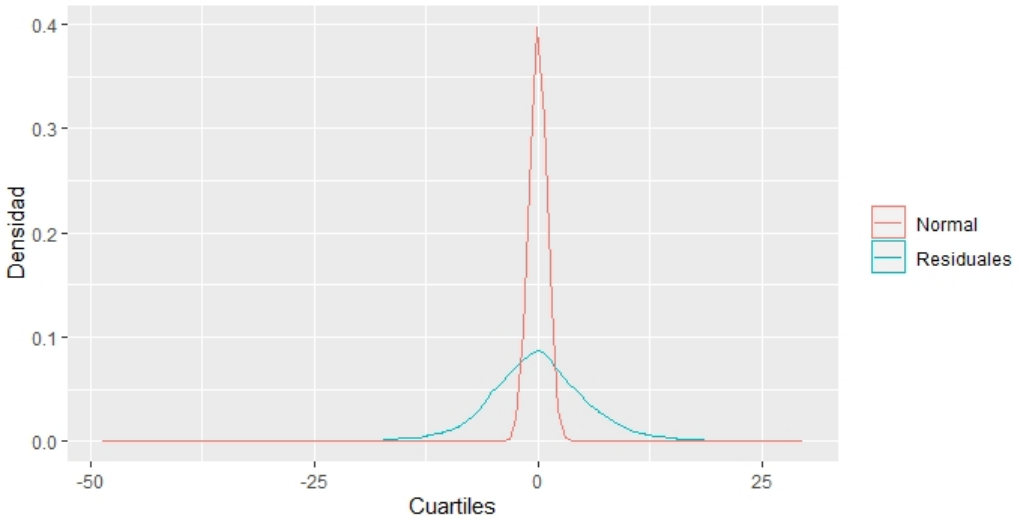

Figura E-2.: Distribución estimada de los residuales condicionados estándar del modelo (azul) después de aplicar raíz cuadrada a la variable logaritmo del Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

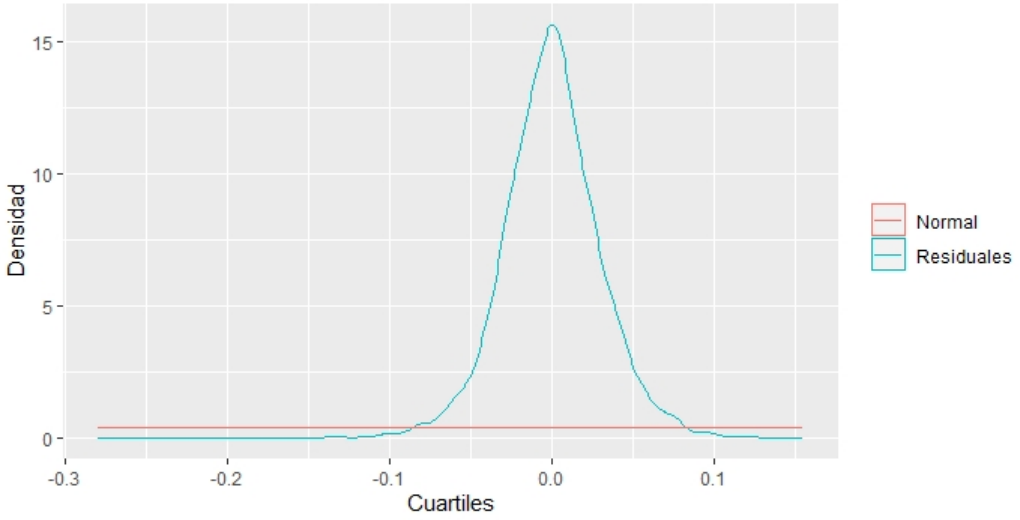

Figura E-3.: Distribución estimada de los residuales condicionados estándar del modelo (azul) después de aplicar raíz cuadrada a la variable logaritmo del Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

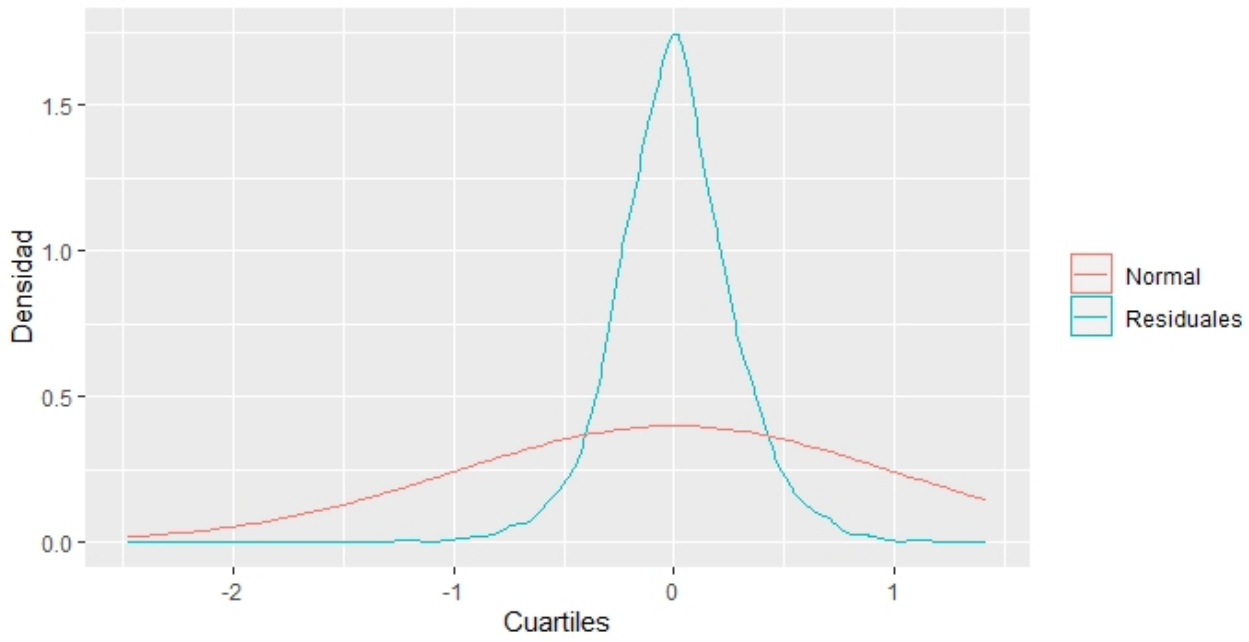

Figura E-4.: Distribución estimada de los residuales condicionados estándar del modelo (azul) utilizando el logaritmo de la variable Precio. Adicional, se muestra función de densidad de una distribución normal estándar (rojo)

Bibliografía

Araque-Solano, A. S., Gutiérrez-López, J. A., & Quenguan-López, L. F. (2017). El espacio público en la ciudad: Una aproximación desde los precios hedónicos. *Sociedad y economía*, 33, 77–98.

Arbia, G. (2006). *Spatial econometrics: statistical foundations and applications to regional convergence*. Berlin: Springer Berlin.

Bivand, R. S., Pebesma, E. J., & Gómez-Rubio, V. (2008). *Applied spatial data analysis with R*. New York; London: Springer.

Buendía-Azorín, J. D. & Sanchez, M. (2013). La distribución del desempleo en las provincias españolas: Un análisis con datos de panel mediante el filtrado espacial. *Investigaciones regionales*, 27, 143–154.

Casas-Lugo, R. (2021). Compraventa de vivienda usada está teniendo registros históricos. https://www.portafolio.co/mis-finanzas/vivienda/fedelonjas-hace-balance-del-sector-inmobiliario-en-2021-y-lo-que-se-espera-para-2022-557715.

Celemín, J. P. (2009). Autocorrelación espacial e indicadores locales de asociación espacial. importancia, estructura y aplicación. *Revista universitaria de geografía*, 18, 11–31.

Cliff, A. D. & Ord, K. (1970). Spatial autocorrelation: A review of existing and new measures with applications. *Economic Geography*, 46, 269–292.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley series in probability and mathematical statistics : Applied probability and statistics. New York: Wiley.

Dubin, R. A. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22(3), 433–452.

El nuevo siglo (2021). Lo que movió al sector de vivienda en 2021. https://www.elnuevosiglo.com.co/articulos/12-27-2021-lo-que-movio-al-sector-de-vivienda-en-2021.

Griffith, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering. Advances in Spatial Science*. New York: Springer-Verlag Berlin Heidelberg.

Hidalgo-Bucheli, G. E. (2020). Uso del Índice de Moran y LISA para explicar el ausentismo electoral rural en Ecuador. *Revista geográfica*, 160, 91–108.

Lee, Y., Nelder, J. A., & Pawitan, Y. (2017). *Generalized Linear Models with Random Effects Unified Analysis via H-likelihood*. New York: Chapman and Hall/CRC.

LeSage, J. & Pace, R. (2009). *Introduction to Spatial Econometrics*. Statistics: A Series of Textbooks and Monographs. London: Taylor & Francis, New York.

León, D. (2021). Real estate: sector de las posibilidades en Colombia. https://blog.fincaraiz.com.co/noticias-fincaraiz/sector-inmobiliario-colombia/.

López-Garcia, M. (2019). Vivienda y política pública: Objetivos e instrumentos. *Documentos de trabajo (FEDEA)*, 7, 1–266.

Melo-Martínez, C. E. & Melo-Martínez, O. (2003). Estimación de precios hedónicos para propiedades residencial y comercial en la ciudad de Bogotá. *Ciencia, investigación, academia y desarrollo*, 8, 10–18.

Mendieta-López, J. C. & Perdomo-Calvo, J. A. (2007). Especificación y estimación de un modelo de precios hedónico espacial para evaluar el impacto de transmilenio sobre el valor de la propiedad en Bogotá. *Documentos CEDE*.

Morales-Zurita, L. & Arias-Arbeláez, F. A. (2005). La calidad de la vivienda en Bogotá: Enfoque de precios hedónicos de hogares y de agregados espaciales. *Revista sociedad y economía*, 9, 47–80.

Pabón-Estrella, M. A. & Burbano-Ojeda, C. A. (2017). *Valoración económica ambiental de la vivienda en relación con el PM10 y ruido en la ciudad de Bogotá, aplicando la metodología de precios hedónicos*. Tesis de pregrado, Universidad Santo Tomas.

Perdomo-Calvo, J. A. (2010). Una propuesta metodológica para estimar los cambios sobre el valor de la propiedad: Estudio de caso para Bogotá aplicando propensity score matching y precios hedónicos espaciales (a propensity score matching and spatial hedonic prices approach for estimating property value fluctuations in Bogotá). *Lecturas de economía*, 73, 11–43.

Pontarollo, N., Mendieta, R., & Ontaneda, D. (2019). El crecimiento cantonal en Ecuador y el papel de la heterogeneidad espacial. *Revista CEPAL*, 129, 163–189.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rey-Hernandez, F. W. (2019). *Valoración del espacio público en la ciudad de Bogotá: Un modelo de precios hedónicos de frontera diferenciado por localidades.* Tesis de maestría, Universidad de los Andes.

Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Political Economy*, 82, 34–55.

RStudio Team (2023). *RStudio: Integrated Development Environment for R.* RStudio, PBC, Boston, MA.

Schabenberger, O. & Gotway, C. (2004). *Statistical Methods for Spatial Data Analysis.* Chapman & Hall/CRC Texts in Statistical Science. New York;London: Taylor & Francis.

Serrano, A. L. (2019). *Evolución de precios en el mercado inmobiliario (vivienda).* Tesis de pregrado, Universidad de Alicante.

Toloza-Delgado, J. D. (2020). *Modelación conjunta de media y varianza en modelos semiparamétricos autorregresivos espaciales.* Tesis de maestría, Universidad Nacional de Colombia.

Toloza-Delgado, J. D., Melo-Martínez, O., & Azcarate-Romero, J. S. (2021). Determinantes del precio de la vivienda nueva en Bogotá para el año 2019: una aproximación a través de un modelo semiparamétrico de regresión espacial. *Ingeniería y Ciencia*, 17, 23–52.

Urrea-Zora, M. & Cárdenas-Mazo, N. (2019). *Un modelo espacial de precios hedónicos y calidad de vecindario: una aplicación para Medellín, Colombia.* Tesis de pregrado, Universidad EAFIT.

Van Rossum, G. & Drake, F. L. (2009). *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace.

Zorrilla-Pérez, A. D. (2012). *Aplicación de la metodología de precios hedónicos para la valoración ambiental de las áreas verdes urbanas en la ciudad de Bogotá.* Tesis de pregrado, Universidad Santo Tomas.