PROPUESTA METODOLÓGICA BASADA EN REGRESIÓN ESPACIAL KRIGING PARA ESTIMAR EL VALOR POR METRO CUADRADO DE ÁREA PRIVADA EN APARTAMENTOS SOMETIDOS A PROPIEDAD HORIZONTAL CON FINES DE AVALÚO HIPOTECARIO (2023–2025): CASO UPL BRITALIA, BOGOTÁ D.C



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Autores

Meylin Sofia Huaman Pineda

Kevin Santiago Forero Silva

Universidad Distrital Francisco José De Caldas

Facultad Ingeniería

Ingeniería Catastral y Geodesia Bogotá D.C.

Bogotá D.C octubre 2 de 2025

PROPUESTA METODOLÓGICA BASADA EN REGRESIÓN ESPACIAL KRIGING PARA ESTIMAR EL VALOR POR METRO CUADRADO DE ÁREA PRIVADA EN APARTAMENTOS SOMETIDOS A PROPIEDAD HORIZONTAL CON FINES DE AVALÚO HIPOTECARIO (2023–2025): CASO UPL BRITALIA, BOGOTÁ D.C



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Autores

Meylin Sofia Huaman Pineda

Kevin Santiago Forero Silva

Modalidad

Monografía

Director

Carlos Eduardo Melo Martínez

Universidad Distrital Francisco José De Caldas

Facultad Ingeniería

Ingeniería Catastral y Geodesia Bogotá D.C.

Bogotá D.C octubre 2 de 2025

DEDICATORIA

" A mi familia, por su apoyo constante y por impulsarme; a quienes me acompañaron con su cercanía y calidez; a mis maestros, que hicieron parte de mi formación profesional y me ofrecieron guía en el camino; a mi amigo peludo, guardián de mis noches y fiel compañero de vida, cuya memoria vive en cada paso de este logro; y, sobre todo, a mi madre, por creer en mí y no dejarme caer." - Meylin Sofía Huamán Pineda

"A mis padres, por ser mi sustento y apoyo incondicional en cada paso de este camino. Gracias por su amor, comprensión y esfuerzo constante para que yo pudiera alcanzar este logro. A mi pareja, por estar siempre a mi lado, brindándome su apoyo y motivación incluso en los momentos más difíciles. Y, por supuesto, a mi fiel compañero de cuatro patas, mi perro, quien ha estado conmigo día y noche, haciendo de cada momento un poco más llevadero.

Este logro es de todos ustedes."-Kevin Santiago Forero Silva

AGRADECIMIENTOS

Primeramente, a la universidad pública que nos acogió y nos formó como profesionales, espacio en el que no solo adquirimos conocimiento, sino también valores, amistades y experiencias que marcarán nuestro camino. Agradecemos profundamente al Ingeniero Carlos Eduardo Melo Martínez, quien con paciencia y compromiso nos guió durante el proceso, siempre dispuesto a orientarnos, resolver nuestras inquietudes y motivarnos a seguir adelante.

RESUMEN

Este proyecto propone una metodología para la estimación del valor por metro cuadrado en apartamentos sometidos a propiedad horizontal, delimitada al caso de la Unidad de Planeación Local (UPL) Britalia en Bogotá D.C. durante el periodo 2023–2025. La propuesta surge de la necesidad de superar limitaciones de los métodos tradicionales de valoración urbana, los cuales suelen basarse en comparaciones directas o promedios zonales que no capturan de manera suficiente la heterogeneidad espacial ni las condiciones contextuales que inciden en el comportamiento del mercado inmobiliario. Para tal fin, se construye una base de datos georreferenciada que integra avalúos técnicos realizados en el periodo de estudio y registros de oferta inmobiliaria recolectados de manera sistemática mediante técnicas de web scraping en portales especializados, garantizando un insumo actualizado y homogéneo para el análisis. Sobre esta base se aplican modelos de regresión y econometría espacial, con el objetivo de evaluar el efecto de variables físicas, socioeconómicas y territoriales en la conformación del valor unitario.

Adicionalmente, se contempla la aplicación de técnicas geoestadísticas, en particular la regresión kriging, como estrategia para capturar la dependencia espacial no explicada por los modelos econométricos, lo que permite generar superficies de estimación y representaciones cartográficas coherentes con la distribución territorial de los valores. Los resultados de este enfoque se proyectan como una herramienta metodológica replicable que amplía las alternativas técnicas disponibles para la valoración urbana, facilitando la validación de avalúos, la interpretación de la estructura espacial del mercado y el fortalecimiento de la trazabilidad en procesos aplicados a gestión inmobiliaria. Asimismo, el carácter escalable de la metodología abre la posibilidad de adaptarla a otros contextos urbanos con dinámicas de alta heterogeneidad y disponibilidad parcial de información, siempre que se disponga de datos georreferenciados que permitan la estructuración del modelo.

Palabras clave: econometría espacial, geoestadística, regresión kriging, valoración inmobiliaria, análisis espacial, SIG, web scraping, Bogotá, avalúos.

ABSTRACT

This project proposes a methodological approach for estimating the square meter value of private area in apartments under condominium ownership, focusing on the Local Planning Unit (UPL) Britalia in Bogotá D.C. during the 2023–2025 period. The proposal arises from the need to overcome the limitations of traditional urban valuation methods, which often rely on direct comparisons or zonal averages that fail to adequately capture spatial heterogeneity and contextual conditions influencing real estate market behavior. To address this, a georeferenced database is constructed by integrating technical appraisal records from the study period with real estate listings systematically collected through web scraping from specialized portals, ensuring a consistent and up-to-date input for analysis. On this basis, regression models and spatial econometrics are applied to evaluate the impact of physical, socioeconomic, and territorial variables on unit values.

Additionally, geostatistical techniques, particularly residual kriging, are considered to capture spatial dependence not explained by econometric models, allowing the generation of estimation surfaces and cartographic representations consistent with the territorial distribution of property values. The outcomes of this approach are projected as a replicable methodological tool that broadens the technical alternatives available for urban valuation, supporting the validation of appraisals, the interpretation of spatial market structures, and the strengthening of traceability in real estate management processes. Furthermore, the scalable nature of the methodology opens the possibility of adapting it to other urban contexts characterized by high heterogeneity and partial data availability, provided that georeferenced information is available to support model structuring.

Keywords: spatial econometrics, geostatistics, kriging regression, real estate valuation, spatial analysis, GIS, web scraping, Bogotá, appraisal.

CONTENIDO

1. INTRODUCCIÓN 12

2. FORMULACIÓN 14

3. OBJETIVOS 16

3.1. OBJETIVO GENERAL 16

3.2. OBJETIVOS ESPECÍFICOS 16

4. MARCO REFERENCIAL 17

4.1. Delimitación Espacial 17

4.2. Antecedentes 19

4.3. Estado del Arte 25

4.4. Marco Teórico 29

5. METODOLOGÍA 40

5.1. Fase I: ESTRUCTURACIÓN DE LA ARQUITECTURA DE LA BASE DE DATOS ESPACIAL. 40

5.1.1. Diseño del modelo de base de datos espacial. 40

5.1.2. Definición de campos y tipos de datos para variables físicas, espaciales y socioeconómicas 41

5.1.3. Integración del sistema de coordenadas mediante proyección MAGNA-SIRGAS EPSG:9377. 43

5.1.4. Validación de la arquitectura mediante pruebas de integridad, topología y consistencia. 43

5.2. Fase II: Extracción, Transformación y Carga de los datos (ETL) 46

5.2.1. Recolección y organización de fuentes: catastro, DANE, SNR, normativas, Appraiser S.A.S y web scraping 46

5.2.2. Limpieza y transformación de datos 47

5.2.3. Estandarización de variables, eliminación de inconsistencias 48

5.2.4. Carga de datos al sistema espacial 49

5.2.5. Ingreso estructurado en la base SIG. 50

5.3. Fase III: Análisis exploratorio de datos. 51

5.3.1. Estadística descriptiva: medidas de tendencia central, dispersión y detección de outliers. 51

5.3.2. Pruebas de normalidad y correlación: Shapiro-Wilk, Kolmogorov-Smirnov, Pearson/Spearman. 54

5.3.3. Representación gráfica: histogramas, QQ-plots y diagramas de dispersión... 55

5.4. Fase IV: Discretización de datos 59

5.4.1. Clasificación de variables categóricas. 59

5.4.2. Codificación de variables: creación de dummys y codificación ordinal. 60

5.4.3. Agrupación de variables continuas en intervalos o cuantiles. 61

5.4.4. Transformación de atributos en clases según criterios de análisis espacial o socioeconómico. 62

5.5. Fase V: Normalización de datos 63

5.5.1. Estandarización de variables (z-score, min–max). 64

5.5.2. Escalado y transformación para comparabilidad entre unidades y magnitudes. 66

5.5.3. Evaluación de distribuciones con pruebas estadísticas y gráficas. 67

5.5.4. Aplicación de transformaciones (logarítmica, Box-Cox, Yeo-Johnson) para estabilizar varianzas y aproximar normalidad. 72

5.6. Fase VI: Transformación y Modelamiento espacial 75

5.6.1. Aplicación de modelos. 76

5.6.2. Validación de modelos y predicciones mediante métricas (MAE, RMSE) y validación cruzada. 100

5.6.3. Interpolación espacial mediante Regresión-Kriging. 108

5.6.4. Generación de mapas temáticos y análisis espacial de resultados. 118

6. RESULTADOS 121

7. ALCANCES 123

8. CONCLUSIONES 125

9. REFERENCIAS 127

10. ANEXOS 134

ILUSTRACIONES

Ilustración 1 Delimitación Espacial 17
Ilustración 2 Histograma de los datos crudos..... 52
Ilustración 3 Boxplot de los datos crudos 53
Ilustración 4 Matriz Pearson y Spearman 55
Ilustración 5 Histograma y Q-Q plot del LM de referencia 57
Ilustración 6 Diagrama de dispersión de residuos. 58
Ilustración 7 Transformación Box-Cox..... 68
Ilustración 8 Transformación Yeo-Johnson 69
Ilustración 9 Transformación Raíz Cuadrada..... 70
Ilustración 10 Transformación asinh..... 71
Ilustración 11 Transformación Logarítmica..... 72
Ilustración 12 Resumen del modelo regresión lineal ordinaria..... 78
Ilustración 13 Gráficos regresión lineal ordinaria..... 79
Ilustración 14 Moran regresión lineal ordinaria 80
Ilustración 15 Resumen del modelo de regresión robusta (RLM – Huber) 83
Ilustración 16 Pruebas conjuntas tipo Wald para bloques de variables 84
Ilustración 17 Gráficos Regresión lineal múltiple..... 85
Ilustración 18 Moran Regresión Lineal Múltiple..... 86
Ilustración 19 Resumen Modelo Aditivo Generalizado 88
Ilustración 20 Funciones Suavizadas del Modelo Aditivo Generalizado..... 91
Ilustración 21 Gráficos Modelo Aditivo Generalizado 92
Ilustración 22 Moran Modelo Aditivo Generalizado 93
Ilustración 23 Resumen Regresión Cuantílica 95
Ilustración 24 Gráficos Regresión Cuantílica 95
Ilustración 25 Moran Regresión Cuantílica 96
Ilustración 26 Resumen Random Forest 98
Ilustración 27 Gráficos Random Forest 99
Ilustración 28 Moran Random Forest..... 99
Ilustración 29 MAE Y RMSE modelos candidatos 104
Ilustración 30 Funciones suavizadas de GAM Final..... 108
Ilustración 31 Gráficos GAM Final 109
Ilustración 32 Variograma Empírico..... 111

Ilustración 33 Variograma direccional - Isotropía 113
Ilustración 34 Variograma..... 114
Ilustración 35 Kriging de Residuos y Varianza de Kriging 115
Ilustración 36Ilustración Validación cruzada Leave One Out 116
Ilustración 37 Predicción valor por metro cuadrado final..... 118
Ilustración 38 Redondeo de Cifras..... 118
Ilustración 39 Simbología y Corte 119
Ilustración 40 Salida Gráfica Final..... 120

TABLAS

Tabla 1 Definición de Dominios 41
Tabla 2 Estructura Relacional de la tabla principal. 45
Tabla 3 Comparación de transformaciones sobre los residuos 73
Tabla 4 Comparación Validación Cruzada 103

1. INTRODUCCIÓN

En el ámbito del financiamiento inmobiliario y la gestión del riesgo crediticio, la estimación objetiva del valor comercial de los inmuebles constituye un insumo clave, particularmente en contextos urbanos donde la dinámica constructiva, la diversidad tipológica y la complejidad socioespacial plantean desafíos para la valoración. Este valor, entendido como el precio más probable de transacción en condiciones de mercado libre, se apoya en marcos normativos internacionales como los del International Valuation Standards Council (IVSC, 2021), que promueven la consistencia y transparencia en los procesos valuatorios.

En Colombia, donde el crédito hipotecario representa uno de los principales mecanismos de acceso a la vivienda, los avalúos técnicos determinan tanto los montos financiables como la exposición al riesgo de las entidades crediticias (Banco de la República, 2023). En este escenario, firmas valuadoras han adoptado lineamientos internacionales como los de la Royal Institution of Chartered Surveyors (RICS, 2020), que enfatizan la competencia técnica y la trazabilidad de las metodologías aplicadas.

La ciudad de Bogotá concentra un mercado inmobiliario altamente heterogéneo, con predominio de proyectos en propiedad horizontal y una marcada variación en factores físicos, socioeconómicos y urbanísticos entre sus localidades. Estas condiciones limitan la eficacia de los métodos tradicionales de comparación directa o promedios zonales, que no siempre capturan adecuadamente las diferencias territoriales ni la influencia de variables contextuales en la conformación del valor. Frente a ello, resulta necesario adoptar enfoques analíticos que incorporen explícitamente la dimensión espacial y reconozcan la interdependencia entre predios y su entorno.

Diversos estudios han evidenciado que los valores inmobiliarios presentan autocorrelación espacial, lo que implica que el precio de un inmueble está relacionado con el de propiedades cercanas (Goodchild, 1986). Para modelar este comportamiento se han empleado técnicas de econometría espacial y geoestadística, entre ellas el kriging, que bajo supuestos estándar proporciona el mejor predictor lineal insesgado (Siglas en inglés BLUP)

y minimiza la varianza de predicción a nivel local del estimador (Joseph, 2025). Alternativamente, se han utilizado extensiones como el co-kriging, que incorpora variables auxiliares para mejorar la capacidad predictiva (Rossiter, 2018).

En este contexto, la integración de información proveniente de avalúos técnicos y registros de oferta inmobiliaria obtenidos mediante web scraping ofrece la posibilidad de construir bases georreferenciadas más completas y actualizadas. Este tipo de insumo, articulado con modelos de regresión espacial y técnicas de interpolación aplicadas a residuos, favorece una representación más coherente del mercado, la identificación de patrones territoriales y la reducción de la incertidumbre inherente a entornos urbanos complejos, contribuyendo a fortalecer la validez técnica de los avalúos y su aplicación en la gestión hipotecaria y en la planeación urbana (Schabenberger & Gotway, 2005).

2. FORMULACIÓN

En el marco de un mercado inmobiliario urbano cambiante, donde conviven ritmos de renovación, densificación y reemplazo tipológico, la cuestión central no es la constatación de la heterogeneidad —ya reconocida—, sino el modo en que esa diversidad se traduce en lecturas consistentes del valor por m². La práctica cotidiana muestra un panorama de fuentes disímiles (avalúos técnicos y registros de oferta reciente) con calendarios, coberturas y niveles de detalle heterogéneos; atributos no homologados que dificultan la comparación entre predios y subzonas; y criterios de calidad y validación aún poco alineados para un análisis territorial reproducible. En este escenario, la formulación del problema se orienta a precisar las brechas de integración y trazabilidad que condicionan la estimación del valor unitario y a delimitar el marco aplicado desde el cual se abordará su lectura territorial.

En términos operativos, la ausencia de una integración sistemática entre información técnica de avalúos y registros de oferta reciente limita la lectura territorial del mercado y restringe la posibilidad de contrastar, en una misma base, atributos del inmueble, condiciones socioeconómicas y factores territoriales asociados a accesibilidad y dotación. Esta desconexión favorece diagnósticos parciales y eleva la incertidumbre sobre el valor por metro cuadrado, en especial en áreas con dinámicas de renovación, densificación o reemplazo tipológico.

El proyecto adopta como ámbito de aplicación la Unidad de Planeación Local (UPL) Britalia de Bogotá D.C. para el periodo 2023–2025, con el propósito de delimitar un espacio de análisis donde se evidencian las situaciones descritas, sin considerar que el fenómeno sea privativo de dicha unidad. En este marco, se reconoce la necesidad de contar con insumos georreferenciados y criterios reproducibles que permitan: i) caracterizar el comportamiento del valor unitario en función de atributos físicos de la unidad inmobiliaria y condiciones del entorno; ii) reducir la incertidumbre derivada de la heterogeneidad espacial y de la calidad desigual de las fuentes; y iii) producir salidas interpretables que puedan ser utilizadas en revisión de avalúos, soporte a decisiones de crédito y ejercicios de planeación.

En síntesis, la combinación de heterogeneidad espacial, fuentes desarticuladas y enfoques convencionales se traduce en tres efectos operativos: i) sesgos en la lectura del valor por m², al no capturar adecuadamente diferencias territoriales y condiciones de entorno; ii) riesgo en decisiones de crédito, por estimaciones menos estables que impactan márgenes de garantía, exposición y políticas de otorgamiento; y iii) dificultad de comparabilidad inter-zonas, que limita la trazabilidad y la comunicación de resultados entre submercados urbanos. Estos efectos justifican la necesidad de consolidar insumos georreferenciados y criterios reproducibles que permitan una lectura territorial consistente del valor y una aplicación más segura en revisión de avalúos y apoyo a la gestión hipotecaria.

3. OBJETIVOS

3.1. OBJETIVO GENERAL

Diseñar una propuesta metodológica para estimar el valor por metro cuadrado de área privada en apartamentos sometidos a propiedad horizontal en la Unidad de Planeamiento Local (UPL) Britalia (Bogotá D.C.), integrando variables físicas, socioeconómicas y espaciales mediante técnicas de análisis econométrico y representación territorial, a partir de avalúos 2023–2025 y registros de oferta obtenidos mediante web scraping.

3.2. OBJETIVOS ESPECÍFICOS

- Consolidar una base de datos estructurada y georreferenciada (2023–2025) con información técnica de avalúos en la UPL Britalia, complementada con registros de oferta recolectados vía web scraping (Portal inmobiliario Finca Raíz), incorporando criterios explícitos de estandarización, depuración y control de calidad (validación de coordenadas, unidades/rangos, tipologías, tratamiento de atípicos y duplicados).
- Explorar y depurar las variables físicas, socioeconómicas y espaciales del inmueble y su entorno (p. ej., área, edad, estrato, estado, piso, dotación comunal, distancias a equipamientos y accesibilidad), dejando las transformaciones y recodificaciones listas para modelación (dummys normalizadas, inversa de área, edad invertida, etc.).
- Evaluar el efecto de las variables seleccionadas sobre el valor por m² mediante modelos de regresión y econometría espacial, verificando autocorrelación y especificación espacial; y, cuando proceda, determinar la aptitud para aplicar kriging (sobre residuos) como ajuste espacial complementario de la estimación.
- Elaborar representaciones cartográficas temáticas y superficies de estimación basadas en el modelo más consistente, incluyendo lineamientos de uso como herramienta técnica complementaria en avalúos urbanos en contextos de heterogeneidad territorial.

4. MARCO REFERENCIAL

4.1. Delimitación Espacial

La Unidad de Planeamiento Local Britalia está localizada entre las calles 138 y 192 o avenida Tibabita, y entre la autopista Norte – avenida Paseo de los Libertadores y la carrera 80 e incluye el suelo rural del cerro de La Conejera. Britalia tiene una extensión de 1.574 ha, de las cuales 1.411 ha están en suelo urbano, 0,26 ha hacen parte del suelo de expansión que se rige bajo los lineamientos del Plan Zonal del Norte – Ciudad Lagos de Torca (Decreto Distrital 088, 2017) y 162,6 ha corresponden a suelo rural. De acuerdo con las proyecciones DANE para 2021, la Unidad de Planeamiento Local (UPL) tiene una población total de 261.499 habitantes y una densidad poblacional de 166 habitantes por hectárea. (Decreto Distrital 557 de 2023). Está compuesta por los siguientes barrios:

Altos de Suba
Britalia
Casablanca Suba
El Plan
Granada Norte
Mazurén
San Antonio Noroccidental
San José V Sector
Suba Urbano
Victoria Norte
Atenas
Cantagallo
Delmonte
Escuela de Carabineros
Iragua
Portales del Norte
San José de Bavaria
Santa Helena
Tuna Alta
Villa del Prado
Barajas Norte
Casa Blanca Suba I
El Cerezo
Gilmar
Las Orquideas
Prado Pinzón
San José del Prado
Suba Cerros
Vereda Suba Cerros II

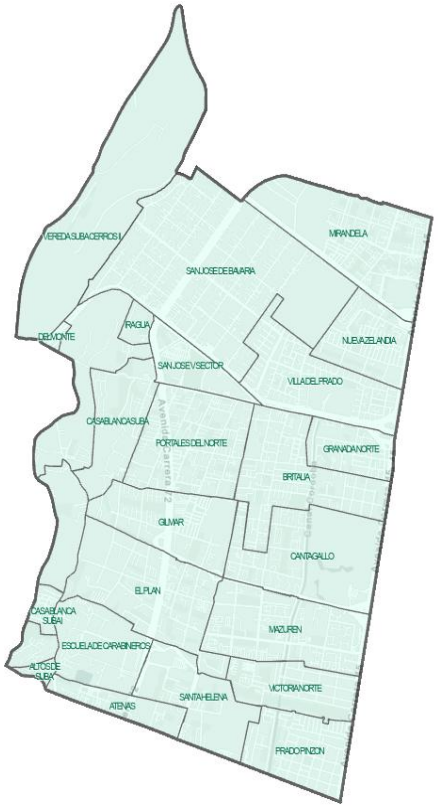


Ilustración 1 Delimitación Espacial

Conforme a la cartografía oficial y a la documentación de participación ciudadana del Distrito, sus límites generales son:

- **Norte:** límite aproximado con la **Calle 189** y la UPL Torca.
- **Sur:** límite aproximado con la **Calle 138** y la UPL Niza.
- **Oriente:** definido en gran parte por la **Autopista Norte** y la UPL Toberín.
- **Occidente:** limita con los cerros de la Conejera / Cerros de Suba y UPL Rincón de Suba (sector de cerros y renques naturales).

La UPL Britalia se caracteriza por ser predominantemente residencial con agrupaciones de vivienda en conjuntos cerrados y los equipamientos educativos e institucionales que allí se ubican son en su mayoría privados. Presenta un tejido urbano caracterizado por un predominio de vivienda multifamiliar organizada en conjuntos cerrados y edificios bajo régimen de propiedad horizontal, lo que la convierte en un escenario idóneo para el análisis del valor por metro cuadrado de área privada. Esta configuración responde a procesos de planeación urbana que han promovido la densificación en zonas con buena accesibilidad, tal como lo establece el Plan de Ordenamiento Territorial – POT Bogotá Reverdece 2022–2035 (Decreto 555 de 2021), en el que se consolida a Britalia como una de las unidades de planeamiento destinadas a fortalecer el uso residencial y optimizar el aprovechamiento del suelo. La norma urbana previa de la UPZ 18 Britalia ya había consolidado un patrón de ocupación predominantemente multifamiliar, base sobre la cual se ha mantenido la vocación residencial de la zona. De acuerdo con el Documento Técnico de Soporte para la Formulación de las UPL del Sector Norte (Decreto 557 de 2023), los equipamientos educativos e institucionales presentes en el área son en su mayoría de carácter privado, lo que refuerza su perfil como territorio de vocación habitacional planeada, complementado con servicios que incrementan su atractivo residencial. La oferta inmobiliaria actual está compuesta principalmente por apartamentos en conjuntos de estratos 3 y 4, con áreas entre 60 m² y 90 m², lo que confirma la prevalencia del régimen de propiedad horizontal y la convierte en una zona de estudio adecuada para explorar metodologías de estimación de valor unitario en contextos urbanos de alta densidad y relativa homogeneidad tipológica.

4.2. Antecedentes

La investigación sobre valoración inmobiliaria ha transitado desde enfoques convencionales hacia metodologías que asumen la complejidad urbana y la dependencia espacial entre observaciones. En el plano hedónico, trabajos recientes integran atributos intrínsecos y extrínsecos con medidas de accesibilidad y entorno, mostrando mejoras en ajuste y trazabilidad del modelo (Rey-Blanco, García-Palomares, & Gutiérrez, 2024) y actualizando el marco de equilibrio hedónico espacial en contextos con cambios de localización residencial y teletrabajo (Brueckner, Kahn, & Lin, 2021). En la vertiente de econometría espacial, se ha consolidado el uso de diagnósticos (Moran global/local) y modelos con dependencia espacial (SAR/SEM/SDM), con evidencia aplicada en mercados de vivienda que documenta autocorrelación espacial significativa y ganancias al modelarla explícitamente (Cellmer, Bełej, & Konowalczuk, 2020). En paralelo, síntesis metodológicas recientes y desarrollos de paneles espaciales y modelos dinámicos anidados han elevado el estándar para aplicaciones urbanas, al combinar rezagos espaciales, temporales y factores comunes en un marco unificado (Arbia, 2016; Elhorst, 2017, 2022).

La incorporación de variables ambientales, como la calidad del aire, la contaminación acústica, la presencia y accesibilidad de espacios verdes, la calidad del agua, y la proximidad a cuerpos de agua o áreas forestales, ha representado un avance significativo en la comprensión del valor inmobiliario. Leggett y Bockstael (2000) evidenciaron esta relación mediante el uso de modelos hedónicos de precios, analizando datos de transacciones inmobiliarias combinados con indicadores ambientales específicos, y demostraron que mejoras en la calidad del agua y la disponibilidad de espacios verdes aumentan de forma apreciable el valor de las propiedades residenciales. De manera complementaria, Tyrväinen y Miettinen (2000) constataron, a través de encuestas de percepción de valor y análisis econométrico de precios de vivienda en Finlandia, que la cercanía a áreas forestales urbanas tiene un impacto positivo y significativo sobre la valorización de los inmuebles, subrayando la creciente importancia de los servicios ecosistémicos en los entornos urbanos. En esta línea, Jim y Chen (2006) analizan la influencia de los parques urbanos sobre el precio de las viviendas en Guangzhou, China, concluyendo que la cantidad y calidad de las áreas verdes

cercanas se reflejan directamente en el valor de mercado de los inmuebles. Además de los espacios naturales, otras variables ambientales relevantes en contextos urbanos son la exposición a fuentes de ruido, la cercanía a fuentes de contaminación industrial, la disponibilidad de corredores ecológicos y la calidad visual del paisaje urbano, factores determinantes de la literatura en la percepción de calidad de vida y en la formación del valor inmobiliario.

La accesibilidad urbana, entendida como la facilidad con que los residentes acceden a bienes y servicios, se reconoce como un factor que incide de forma sistemática en la formación del precio de la vivienda, con efectos medibles cuando se incorporan índices de accesibilidad en modelos hedónicos y cuando cambian las condiciones de acceso por mejoras de transporte (Rey-Blanco, García-Palomares, & Gutiérrez, 2024; Rennert, 2022; Higgins, Arku, Farber, & Miller, 2024) aportaron evidencia empírica al demostrar que las viviendas situadas en proximidad a líneas de transporte público masivo, específicamente cercanas a estaciones de tren ligero, experimentan una valorización significativa, debido no solo a la reducción en los tiempos de desplazamiento, sino también a la mejora percibida en la conectividad regional. En este mismo sentido, Gibbons y Machin (2005), en su análisis del sistema ferroviario de Londres, profundizaron en la influencia de factores como la frecuencia de servicio, la densidad de las estaciones y la integración de los nodos de transporte con las redes peatonales circundantes, identificando que estas características amplifican el efecto positivo sobre los precios inmobiliarios. Más allá del simple acceso, investigaciones como la de Mohammad et al. (2013) han demostrado que la calidad del trazado vial, el mantenimiento de las vías y la percepción de seguridad vial también inciden en la formación de precios, ya que infraestructuras deterioradas o con altos índices de siniestralidad tienden a desincentivar la demanda residencial. Adicionalmente, estudios de Diao, Leonard y Sing (2017) evidencian que la accesibilidad mejora el atractivo de determinadas zonas urbanas no solo por facilitar la movilidad, sino también por generar corredores de desarrollo que atraen inversiones complementarias en comercio y servicios, dinamizando el valor del suelo y consolidando entornos urbanos más seguros y atractivos. De esta forma, la literatura actual reconoce que la influencia del sistema vial sobre la valorización inmobiliaria es multidimensional, combinando aspectos de acceso físico, conectividad estratégica, percepción de seguridad y

calidad de la infraestructura, todos ellos fundamentales para explicar la dinámica de precios en contextos urbanos.

Asimismo, los estudios sobre segmentación del mercado han resultado cruciales para comprender la diferenciación de precios dentro de contextos urbanos heterogéneos. Goodman y Thibodeau (2003) señalaron que los mercados de vivienda funcionan mediante submercados interrelacionados, donde el perfil socioeconómico de los vecindarios y las expectativas de desarrollo urbano configuran niveles diferenciados de precios. Esta afirmación se sustenta en la evidencia obtenida a partir de modelos hedónicos que incorporaron variables proxy como los proyectos de renovación urbana en curso, la planificación de nuevas infraestructuras y la asignación de inversiones públicas, constatando que estas expectativas, aún antes de materializarse físicamente, ya ejercen una influencia anticipada sobre las valoraciones inmobiliarias debido a la percepción de futura revalorización por parte de los agentes del mercado. Sin embargo, es importante reconocer que este fenómeno no es universal ni homogéneo. Como advierten Grigsby y Rosenburg (1975), la respuesta del mercado a las iniciativas de desarrollo urbano depende de múltiples factores contextuales, tales como la reputación histórica del área, la persistencia de problemáticas sociales estructurales, o la desconfianza de los compradores respecto a la sostenibilidad de las intervenciones urbanísticas. Existen casos en los que, pese a los esfuerzos de inversión y mejora del entorno, la percepción negativa del barrio persiste, limitando el impacto positivo esperado en los precios de la vivienda. La evidencia reciente señala que la estigmatización barrial puede amortiguar o retrasar la capitalización de intervenciones urbanas en los precios y en la demanda: la reputación histórica del área, la persistencia de problemáticas sociales y la desconfianza frente a la sostenibilidad de los cambios tienden a neutralizar parcialmente los efectos esperados de las inversiones públicas, generando escenarios donde la oferta mejora pero la respuesta del mercado resulta limitada (Koster & van Ommeren, 2023). La literatura destaca que, aunque las expectativas de desarrollo urbano son relevantes en la configuración de precios, su efecto depende de las condiciones preexistentes del territorio y de la capacidad de las intervenciones para revertir percepciones arraigadas entre los compradores.

La evolución metodológica ha estado también marcada por el avance de técnicas de aprendizaje automático y el aprovechamiento de grandes bases de datos. Kok, Koponen y Martinez (2017) analizaron cómo la automatización de los procesos de valoración mediante algoritmos de machine learning ha incrementado la precisión y la eficiencia de las estimaciones. Esta tendencia fue reforzada por Antipov y Pokryshevskaya (2012), quienes aplicaron algoritmos de bosque aleatorio para la valoración masiva de apartamentos residenciales, logrando una notable reducción de los errores de predicción frente a los modelos convencionales.

Por otro lado, la incertidumbre en las estimaciones ha motivado la adopción de enfoques Bayesianos y jerárquicos, que permiten capturar tanto la dependencia espacial como la heterogeneidad no observada, mejorando así la capacidad explicativa de los modelos de precios de vivienda.

Desde la perspectiva de la heterogeneidad espacial, Helbich et al. (2014) ofrece una visión detallada al aplicar modelos de regresión espacial al mercado inmobiliario austríaco, evidenciando cómo la variabilidad local mejora sustancialmente la precisión de las predicciones. Finalmente, la integración de datos socioeconómicos detallados ha enriquecido el análisis de los precios inmobiliarios. Bajari y Kahn (2005) emplearon modelos estructurales para explicar cómo las preferencias individuales y las restricciones presupuestarias se reflejan directamente en la formación de precios, mientras que Glaeser y Gyourko (2018) identificaron la rigidez de la oferta como un factor decisivo en la dinámica de valorización urbana.

En el desarrollo de los métodos espaciales aplicados a la valoración inmobiliaria, los enfoques geoestadísticos comenzaron a consolidarse como una línea complementaria a la econometría espacial, orientada a representar de manera continua los patrones de variación territorial y a incorporar el tratamiento probabilístico de la incertidumbre. Durante las últimas décadas, el uso de técnicas derivadas del kriging se ha extendido desde la estimación de propiedades del suelo y variables ambientales hacia contextos urbanos, impulsado por el crecimiento de las bases de datos georreferenciadas y la disponibilidad de herramientas de análisis espacial de alta resolución (Hengl et al., 2018; Li & Heap, 2014). Estudios posteriores han documentado su utilidad para capturar estructuras de dependencia espacial y

reducir el sesgo de predicción en comparación con interpoladores determinísticos, especialmente en escenarios con alta heterogeneidad y distribución irregular de observaciones (Keskin & Grunwald, 2018). De este modo, la geoestadística se reconoce en la literatura reciente como un antecedente metodológico relevante dentro de la evolución de los modelos espaciales, marcando la transición hacia enfoques híbridos capaces de integrar dependencia espacial y cuantificación de incertidumbre en el análisis territorial.

En la evolución posterior de la geoestadística, el kriging universal (UK) se consolidó como una extensión del kriging ordinario que incorpora una deriva o tendencia sistemática expresada mediante variables auxiliares externas. Esta formulación posibilitó incluir información explicativa en el proceso de interpolación, ampliando su aplicabilidad a escenarios con gradientes espaciales no estacionarios o influenciados por factores exógenos observables. Diversas investigaciones destacan que la integración de covariables ambientales o socioeconómicas mejora la capacidad de representación espacial y la coherencia estructural del modelo en comparación con técnicas que asumen estacionariedad (Sampson et al., 2013; Baisad, Chutsagulprom, & Moonchai, 2023). A nivel operativo, el desarrollo de bibliotecas especializadas como PyKrige y gstat ha facilitado la implementación computacional del kriging universal, ofreciendo rutinas reproducibles que generan simultáneamente superficies de predicción y varianza de estimación (Gräler, Pebesma, & Heuvelink, 2016; PyKrige Developers, 2024). En conjunto, estos avances consolidan al kriging universal como una evolución metodológica orientada a la incorporación explícita de covariables en la estimación espacial, antecediendo a los esquemas híbridos contemporáneos de regresión-kriging y procesos Gaussianos.

Tras su consolidación teórica, el kriging universal (UK por sus siglas en inglés) ha sido adoptado en investigaciones aplicadas que buscan estimar variables urbanas y territoriales con estructura espacial compleja. En estos estudios, su capacidad para incorporar derivas informadas ha permitido generar superficies continuas de estimación y mapas de incertidumbre útiles para la planificación y el análisis espacial. Aplicaciones recientes evidencian que el UK mejora la estabilidad de las predicciones frente a métodos determinísticos, especialmente en redes de observación irregulares o de baja densidad (Lucas, Lamb, & Siciliano, 2022). En el ámbito inmobiliario, su extensión hacia enfoques de

universal cokriging y regression-cokriging ha permitido representar simultáneamente la dependencia espacial y la influencia de covariables exógenas, logrando estimaciones más coherentes del valor y su variabilidad en el territorio.

La regresión-kriging (RK) representa una evolución metodológica que combina un modelo explicativo global —frecuentemente hedónico, espacial o de aprendizaje automático— con la estructura geoestadística del kriging aplicada sobre los residuos. Este esquema híbrido permite estimar superficies continuas que integran la variabilidad explicada por las covariables y la dependencia espacial remanente, proporcionando una estimación más realista y una cuantificación explícita de la incertidumbre (Hengl, Heuvelink, & Stein, 2004). Diversos estudios recientes la consideran una técnica robusta y flexible en la cartografía predictiva y en la estimación de propiedades territoriales, particularmente cuando se dispone de covariables densas o derivadas de sensores remotos. En aplicaciones urbanas y de valoración inmobiliaria, la RK ha mostrado ventajas frente a modelos exclusivamente econométricos al permitir una representación espacial continua de los valores estimados, especialmente en áreas con distribución desigual de observaciones. Asimismo, las aproximaciones derivadas de los Procesos Gaussianos escalables, como el Nearest Neighbor Gaussian Process (NNGP), han permitido mantener el marco probabilístico del kriging con una reducción significativa del costo computacional, habilitando la generación de superficies e incertidumbres en conjuntos de datos de gran tamaño, como los que caracterizan a entornos urbanos densos (Datta et al., 2016).

En síntesis, la literatura muestra que la incorporación progresiva de estructuras espaciales dentro de los modelos econométricos ha permitido superar las limitaciones de los enfoques puramente globales, integrando la noción de dependencia espacial y la representación continua del territorio mediante técnicas como el kriging y sus variantes. La convergencia entre la econometría espacial y la geoestadística ha dado lugar a metodologías intermedias —como la regresión-kriging— que mantienen el carácter explicativo de los modelos tradicionales, al tiempo que introducen una formulación probabilística capaz de estimar la variabilidad local y la incertidumbre asociada. Esta trayectoria metodológica configura el marco de referencia a partir del cual se fundamenta el desarrollo del presente

estudio, en el que la regresión espacial con componente geoestadístico se entiende como una extensión natural de los avances previos en valoración y análisis territorial.

4.3. Estado del Arte

La valoración urbana en contextos densos y heterogéneos ha evolucionado desde enfoques comparativos/hedónicos hacia esquemas que integran dependencia espacial, no linealidades y cuantificación explícita de la incertidumbre. Este desplazamiento responde a tres necesidades: (i) representar la autocorrelación espacial de los precios y del entorno urbano, (ii) mejorar la validez externa de las predicciones evitando sobreoptimismo por particiones aleatorias, y (iii) ofrecer productos operables (mapas y bandas de incertidumbre) útiles para revisión de avalúos y apoyo a decisiones. En la última década se consolidan cuatro familias: econometría espacial/filtrado, geoestadística y procesos Gaussianos escalables, validación cruzada espacial y aprendizaje automático explicable aplicado a precios.

El Eigenvector Spatial Filtering (ESF) —y su extensión de efectos aleatorios (RE-ESF)— permite capturar la dependencia espacial sin imponer una forma dinámica específica, manteniendo la interpretabilidad de los coeficientes y reduciendo sesgos por correlación residual. Desarrollos recientes lo hacen viable con muestras grandes y coeficientes que varían en el espacio (SVC), con evidencia de mejoras en ajuste y estabilidad en estudios urbanos y de vivienda. En concreto, se han documentado implementaciones eficientes para grandes volúmenes de datos (Murakami & Griffith, 2019), aplicaciones a precios de vivienda con ganancias frente a alternativas globales (McCord et al., 2019) y variantes multiescala para mercados de arriendo que muestran heterogeneidad a distintas escalas espaciales (Peng, Z., & Inoue, R. 2022). Más recientemente, se proponen estrategias de sub-model aggregation que preservan patrones espaciales reduciendo coste computacional, reforzando la escalabilidad en entornos urbanos complejos (Murakami et al., 2024).

Para producir superficies continuas del valor y representar la incertidumbre de predicción, la combinación de modelos explicativos con kriging de residuos sigue siendo práctica habitual; sin embargo, el reto contemporáneo es la escalabilidad. La familia de

Nearest-Neighbor Gaussian Processes (NNGP, en español, Procesos Gaussianos de Vecinos Más Cercano) demostró que es posible realizar inferencia Bayesiana completa en decenas de miles de ubicaciones usando matrices de precisión dispersas, con resultados comparables al GP (Gaussian Processes, en español, *Procesos Gaussianos*) completo y costes manejables (Datta et al., 2016; revisión y extensiones posteriores). Estudios comparativos multi-grupo muestran el desempeño de métodos escalables (incluidos GP aproximados) en competiciones de predicción con datos simulados y observados, aportando guías sobre su robustez y costo (Heaton et al., 2019). Adicionalmente, se han propuesto marcos GP interpretables y variantes con restricciones en variedades/manifolds que mejoran la conexión con fenómenos espaciales y la automatización de kernels/variogramas, útiles cuando se requieren superficies y bandas de incertidumbre con fundamento probabilístico (Yang et al., 2021; Bourdais et al., 2024). En conjunto, estas líneas permiten llevar la geoestadística/GP a dominios urbanos con gran número de puntos, manteniendo la trazabilidad de la varianza de predicción y la coherencia espacial.

La evidencia reciente permite determinar que la validación k-fold aleatoria subestima el error cuando existe autocorrelación espacial o temporal; por ello, se recomiendan esquemas bloqueados/buffer y separaciones espaciales/ambientales para particionar entrenamiento-prueba de forma realista. La revisión de Roberts et al. (2017) sistematiza estas estrategias y muestra el sesgo de las particiones aleatorias; herramientas como blockCV (R) facilitan particiones reproducibles y diagnóstico previo del rango efectivo de autocorrelación, apoyando una selección honesta de modelos (Valavi et al., 2019). Trabajos recientes continúan evaluando el sesgo optimista de la CV (Cross -Validation, Validación Cruzada en español) aleatoria y afinan criterios de particionado espacial (Mahoney et al., 2023). Para valoración urbana, estos protocolos son críticos: permiten comparar de forma justa OLS/espaciales, kriging/GP y ML bajo la misma estructura de dependencia.

El Geographically Weighted Regression (GWR, en español, Regresión Geográficamente Ponderada) y su extensión MGWR (Multiscale Geographically Weighted Regression) han sido modernizadas con implementaciones paralelas y GPU (Compute Unified Device Architecture, CUDA, en español, Arquitectura Unificada de Computación de Dispositivos), lo que reduce sustancialmente los tiempos de cómputo; en paralelo, las

versiones Bayesianas permiten incorporar incertidumbre en los coeficientes locales y atenuar el riesgo de sobreajuste local. Estos avances han ampliado la aplicabilidad en datasets urbanos grandes y han refinado la lectura de no estacionaridad en los determinantes del valor (Lu et al., 2022; desarrollos y análisis de escala/correlación en MGWR 2024–2025; aproximaciones Bayesianas recientes). Para el ámbito de UPL, estas técnicas aportan diagnóstico espacial fino, siempre que su evaluación se someta a CV espacial y se controle la colinealidad local.

En escenarios con no linealidades e interacciones (por ejemplo, área × edad × accesibilidad), los ensambles de árboles —Random Forest y Gradient Boosting (incl. XGBoost/LightGBM)— suelen lograr mejor desempeño predictivo frente a interpoladores determinísticos o a especificaciones lineales rígidas. Estudios recientes en mercados de vivienda y ciudades sostenibles reportan mejoras en precisión y análisis de importancia mediante SHAP (explicabilidad a nivel global y de observación) y, cuando se requiere, estimación de bandas de incertidumbre con modelos cuantílicos (Como Quantile Random Forest , QRF, en español, Bosque Aleatorio Cuantílico) o conformal prediction; en todos los casos, se insiste en evaluar bajo validación espacial para evitar spatial leakage (Han et al., 2025; Sharma et al., 2024; Nikou et al., 2022; Hjort et al., 2022/2024; Dou et al., 2023). En valoración urbana aplicada, su rol recomendado es complementar a los modelos espaciales: explorar estructuras no capturadas por especificaciones paramétricas y corroborar señales con medidas de importancia/SHAP y mapas de errores por subzona.

Durante la última década, la geoestadística ha adquirido un papel protagónico dentro de los enfoques de modelación espacial aplicada, pasando de su uso tradicional en disciplinas ambientales hacia su incorporación en el análisis urbano y la valoración inmobiliaria. Esta evolución metodológica ha permitido integrar el tratamiento probabilístico de la incertidumbre con los marcos econométricos y de planificación territorial, favoreciendo interpretaciones continuas del territorio y una lectura más precisa de la variabilidad espacial. En este contexto, las técnicas derivadas del kriging se han consolidado como herramientas útiles para representar la distribución espacial del valor y estimar su incertidumbre asociada, fortaleciendo los procesos de análisis urbano y de valoración del suelo (Armijos-Toro, González, & Sarmiento, 2022).

El desarrollo metodológico contemporáneo se ha caracterizado por la incorporación de variantes del kriging que permiten incluir información auxiliar o corregir la no estacionariedad de los fenómenos espaciales. Entre ellas, el kriging universal y la regresión-kriging se destacan por combinar dependencias locales con covariables explicativas o tendencias globales, logrando estimaciones más estables frente a interpoladores determinísticos. Su aplicación se ha extendido a la generación de mapas continuos de precios de suelo, densidad edificatoria y calidad ambiental, demostrando su capacidad para capturar gradientes espaciales complejos en entornos urbanos heterogéneos (Chica-Olmo, Cano-Guervos, & Chica-Rivas, 2019).

En el ámbito urbano latinoamericano, la regresión-kriging ha sido explorada como una alternativa intermedia entre los modelos econométricos y los geoestadísticos puros, integrando covariables socioeconómicas e infraestructura vial para estimar superficies continuas de valor. Los resultados evidencian que este enfoque reduce el sesgo de predicción y mejora la coherencia espacial de las superficies generadas, particularmente en ciudades con alta densidad y heterogeneidad local (Villada-Medina, Hernández-Muñoz, & Arango-Londoño, 2022). Estas aplicaciones demuestran la madurez alcanzada por la geoestadística contemporánea, que trasciende la función meramente predictiva para consolidarse como un marco de inferencia espacial compatible con los modelos explicativos.

Finalmente, los estudios comparativos entre métodos hedónicos, kriging y enfoques de aprendizaje automático evidencian que la combinación de componentes explicativos y estocásticos produce resultados más consistentes en la estimación del valor del suelo urbano, siempre que exista una adecuada densidad de observaciones y una calibración espacial rigurosa (Derdouri, 2020). Estas tendencias reflejan una madurez metodológica orientada a integrar las capacidades descriptivas de la econometría espacial con la inferencia probabilística de la geoestadística, conformando un cuerpo metodológico híbrido que sustenta los análisis contemporáneos de la valoración urbana.

4.4. Marco Teórico

La valoración inmobiliaria, entendida como el proceso técnico que permite estimar con base objetiva y sistemática el valor de un bien inmueble en condiciones específicas de mercado, se sostiene sobre principios económicos, urbanísticos y estadísticos que se articulan para aproximar el valor de intercambio de un activo inmobiliario. Desde el enfoque tradicional, esta actividad ha estado regida por el método comparativo de mercado, que asume que el valor de un bien se puede inferir a partir del análisis de transacciones (ofertas) recientes de propiedades con características físicas similares. Sin embargo, este enfoque presenta limitaciones en contextos urbanos con alta heterogeneidad espacial y donde la disponibilidad de comparables es reducida o poco representativa. Frente a estas restricciones, la evolución teórica y técnica ha promovido la adopción de modelos de valoración masiva (Por sus siglas en inglés, AVM) que integran estructuras multivariadas y componentes espaciales para capturar con mayor fidelidad las variaciones del mercado (International Association of Assessing Officers [IAAO], 2025).

Los modelos hedónicos, introducidos formalmente por Rosen (1974), son una herramienta fundamental para analizar mercados inmobiliarios, al permitir la descomposición del precio de un bien inmueble según el valor atribuido a cada atributo observable. Este enfoque parte del principio microeconómico de que el precio de mercado de un bien compuesto refleja la suma de los precios implícitos de sus características individuales. En el caso del mercado de vivienda, estas características se agrupan comúnmente en variables estructurales (como el área construida, el número de habitaciones, el estado de conservación), espaciales (distancia a centros urbanos, cobertura de transporte público, accesibilidad a servicios) y contextuales (estrato socioeconómico, niveles de seguridad, calidad ambiental, entre otros). La formulación típica del modelo hedónico adopta una estructura funcional de regresión lineal o semilogarítmica, donde el precio o el logaritmo del precio se modela como función de estas variables, permitiendo cuantificar el efecto marginal de cada atributo sobre el valor final del inmueble. Para contextos actuales, integrar medidas de accesibilidad y entorno dentro de hedónicos y AVM mejora la trazabilidad y el ajuste de los modelos (Rey-Blanco, Zofío, & González-Arias, 2024).

Una de las principales virtudes del modelo hedónico es su capacidad para capturar la heterogeneidad del mercado y hacer visibles las preferencias reveladas de los compradores, expresadas a través del comportamiento observado en las transacciones. Sin embargo, este enfoque presenta ciertas limitaciones, especialmente en contextos donde existe dependencia espacial entre observaciones, ya que los precios de los inmuebles no son independientes unos de otros, sino que tienden a estar influenciados por la proximidad geográfica y las condiciones del entorno inmediato. Esta violación del supuesto de independencia de los errores puede sesgar las estimaciones y reducir la validez inferencial del modelo, lo que ha motivado el desarrollo de extensiones espaciales y la incorporación de técnicas geoestadísticas para mejorar su precisión. La literatura reciente sobre métodos para datos espaciales a gran escala compara enfoques econométricos, procesos Gaussianos y variantes escalables, mostrando escenarios de desempeño diferencial y criterios de validación (Heaton et al., 2019).

La valoración masiva apoyada en modelos explicativos y componentes espaciales exige comprobar condiciones mínimas sobre la distribución de la variable y, sobre todo, de los errores del modelo. En términos prácticos, estos diagnósticos orientan (i) la necesidad de transformaciones para estabilizar varianza y mejorar linealidad; (ii) el tratamiento de valores extremos; y (iii) la validez de la inferencia (intervalos, pruebas).

Como primer insumo, se calculan sesgo (skewness) y curtosis (kurtosis) de la distribución del valor por m². Un sesgo cercano a cero sugiere simetría; valores positivos/negativos indican colas a derecha/izquierda. La curtosis compara el "apuntamiento" respecto a la normal: valores altos evidencian colas pesadas (leptocurtosis), bajos colas ligeras (platicurtosis). Estos indicadores permiten anticipar si se requieren transformaciones antes del ajuste (Virtanen et al., 2020). Por otro lado es necesario analizar pruebas de normalidad como:

- **Shapiro–Wilk** : Contrasta la hipótesis de normalidad con elevada potencia en muestras pequeñas y medianas. Su estadístico W compara los datos ordenados con los cuantiles esperados bajo normalidad; valores de W próximos a 1 apoyan la normalidad. (SciPy, 2024a).

- **D'Agostino–Pearson:** Integra simultáneamente **sesgo** y **curtosis** en un único contraste de bondad de ajuste. El estadístico $K^2 = z_1^2 + z_2^2$(donde z_1 y $z_2$son puntuaciones z para sesgo y curtosis) sigue χ^2con 2 grados de libertad bajo H_0 (normalidad), siendo útil para detectar desviaciones combinadas de simetría y forma. (SciPy, 2024b).
- **Anderson–Darling:** Asigna mayor peso a las colas, resultando pertinente en contextos inmobiliarios donde pocos inmuebles de muy alto/ bajo valor impactan de forma desproporcionada métricas de ajuste. (SciPy, 2024c).

En todas las pruebas se establece un nivel α (Alpha). Si el *p*-value ≤α, se rechaza la normalidad; si es mayor, no se rechaza. Dado que con tamaños muestrales medios/grandes los contrastes pueden volverse muy sensibles, se recomienda complementar con inspección visual (histogramas, Q–Q plots) y considerar la normalidad de los residuos del modelo como condición más relevante que la normalidad de la variable bruta (GraphPad, 2024).

Cuando los diagnósticos indican asimetría, colas pesadas o heterocedasticidad, se justifica usar transformaciones (Box–Cox), especificaciones más flexibles o estimadores robustos. Para los componentes espaciales, asegurar residuos aproximadamente normales y covarianza espacial bien especificada favorece el desempeño del kriging de residuos y la coherencia de las bandas de incertidumbre. Estas verificaciones, junto con la validación cruzada con sensibilidad espacial, fortalecen la trazabilidad de las estimaciones del valor por m² y su uso en revisión de avalúos y soporte a decisiones.

Para los componentes espaciales, asegurar residuos aproximadamente normales y una covarianza espacial correctamente especificada favorece el desempeño del kriging de residuos y la coherencia de las bandas de incertidumbre. Estas verificaciones, junto con la validación cruzada con sensibilidad espacial, fortalecen la trazabilidad de las estimaciones del valor por metro cuadrado y su uso en la revisión de avalúos y soporte a decisiones.

En este contexto, la geoestadística constituye el marco teórico que permite modelar la dependencia espacial entre observaciones a través del concepto de variable regionalizada, definida como una función aleatoria cuya realización varía de forma continua en el espacio (Journel & Huijbregts, 1978). Esta formulación reconoce que las observaciones próximas

tienden a ser más semejantes que aquellas más alejadas, lo que introduce correlación espacial y continuidad en el fenómeno estudiado. Bajo este enfoque, se asumen propiedades que garantizan la regularidad del proceso: la estacionaridad de segundo orden, que presupone una media constante y una covarianza dependiente únicamente del vector de separación; la isotropía, que supone que la correlación depende solo de la distancia y no de la dirección; y la ergodicidad, que permite sustituir promedios teóricos por promedios espaciales observados, fundamento esencial para la estimación empírica del variograma (Stein, 1999; Wackernagel, 2003).

La dependencia espacial se representa mediante la covarianza espacial $C(h)$ o, de forma equivalente, a través de la semivarianza $\gamma(h)$, definida como la mitad de la varianza esperada de la diferencia entre dos puntos separados por una distancia h:

$$\gamma(h) = \frac{1}{2}E([Z(s) - Z(s+h)]2).$$

En procesos estacionarios de segundo orden se cumple que $\gamma(h) = C(0) - C(h)$, donde $C(0)$ corresponde a la varianza total del campo aleatorio. De este modo, la semivarianza refleja el grado de similitud espacial entre observaciones: valores pequeños de $\gamma(h)$ indican fuerte continuidad, mientras que valores altos denotan pérdida de correlación conforme aumenta la distancia. Los componentes clásicos del variograma —el efecto pepita (nugget, en inglés), la meseta (sill) y el rango (range)— describen respectivamente la microvariabilidad o error de medición, la varianza total del proceso y la escala de correlación espacial (Isaaks & Srivastava, 1989). Este marco conceptual permite caracterizar la estructura de dependencia y constituye la base formal para la interpolación geoestadística.

El kriging, derivado directamente de esta teoría, se define como el estimador lineal insesgado de mínima varianza (Best Linear Unbiased Estimator, por sus siglas en inglés BLUE), cuyo objetivo es obtener la mejor predicción posible del valor de una variable en una localización no muestreada a partir de una combinación ponderada de observaciones cercanas. Los pesos de interpolación se determinan mediante un sistema lineal que minimiza la varianza de estimación bajo la restricción de insesgamiento, utilizando la estructura de covarianza inferida del variograma (Cressie, 1993). A diferencia de los interpoladores determinísticos, el kriging no solo genera una superficie de predicción sino también una

medida explícita de incertidumbre espacial, representada por la varianza de krigeado, que permite identificar zonas con mayor o menor confiabilidad en las estimaciones (Oliver & Webster, 2015).

La estimación kriging se formaliza como un problema de optimización lineal sujeto a la restricción de insesgamiento. Dado un conjunto de observaciones $Z(s_1), Z(s_2), \dots, Z(s_n)$, el valor estimado en una localización no muestreada s_0 se expresa como una combinación lineal ponderada:

$$\hat{Z}(s_0) = \sum_{i=1}^{n} \lambda_i Z(s_i)$$

donde los pesos λ_i se determinan de manera que el estimador sea insesgado y que la varianza del error de predicción sea mínima. Para resolver este problema, se emplea el multiplicador de Lagrange *(μ),* que introduce la restricción $\sum_i \lambda_i = 1$ dentro de la función objetivo. El sistema resultante de ecuaciones puede expresarse como:

$$\begin{cases} \sum_{j=1}^{n} \lambda_i C(s_i, s_j) + \mu = C(s_i, s_0), i = 1, \dots, n \\ \sum_{j=1}^{n} \lambda_i C = 1 \end{cases}$$

Donde $C(s_i, s_j)$ representa la covarianza espacial entre los puntos (s_i, s_j). En forma matricial compacta, el sistema adopta la forma:

$$\begin{bmatrix} C & 1 \\ 1^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} C_0 \\ 1 \end{bmatrix}$$

donde C es la matriz de covarianza entre observaciones, λ el vector de pesos y C_0 el vector de covarianzas entre las observaciones y el punto a estimar. Una vez determinados los pesos, la varianza de krigeado se calcula como:

$$\sigma_K^2 = C(0) - 2\sum_{i=1}^{n} \lambda_i C(s_i, s_0) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j)$$

Esta formulación expresa cuantitativamente la incertidumbre asociada a cada predicción y depende únicamente de la geometría de los datos y del modelo de covarianza. Desde un enfoque probabilístico, el kriging puede entenderse como un caso particular de los procesos Gaussianos (Gaussian Processes), donde las observaciones $Z(s)$ se modelan como una distribución normal multivariada con media $m(s)$ y covarianza. En este contexto, la predicción s_0 corresponde al valor esperado de la distribución condicional:

$$\hat{\mathrm{Z}}(\mathrm{s}_0) \;=\; \mathrm{E}[\,\mathrm{Z}(\mathrm{s}_0) \;|\; \mathrm{Z}(\mathrm{s}_1),\, \mathrm{Z}(\mathrm{s}_2),\; \ldots,\, \mathrm{Z}(\mathrm{s}_n)\,]$$

lo que vincula el kriging clásico con la inferencia Bayesiana moderna y los enfoques computacionales de procesos Gaussianos (Rasmussen & Williams, 2006).

El análisis del variograma no se limita a describir la dependencia espacial, sino que constituye el paso clave para parametrizar el modelo de correlación que será utilizado en el proceso de krigeado. En la práctica, el variograma experimental se obtiene a partir de los datos empíricos mediante la agrupación de pares de observaciones separadas por intervalos de distancia, y su forma revela la estructura espacial del fenómeno. Un ascenso inicial rápido del variograma indica alta continuidad local, mientras que un aumento gradual sugiere una correlación de mayor alcance. La estabilización del variograma hacia una meseta, marca la distancia más allá de la cual las observaciones dejan de estar correlacionadas (Goovaerts, 1997).

Los modelos teóricos del variograma permiten formalizar esta relación mediante funciones continuas y positivas definidas. El modelo esférico, ampliamente utilizado en aplicaciones ambientales y urbanas, supone que la semivarianza crece linealmente hasta alcanzar el rango y luego se mantiene constante, lo que representa fenómenos con correlación espacial finita. El modelo exponencial describe una estructura en la que la correlación decrece de forma gradual e infinita, siendo adecuado para variables con transiciones suaves entre zonas vecinas. En contraste, el modelo Gaussiano asume una correlación muy alta a distancias cortas y una variación extremadamente suave, útil para fenómenos con alta continuidad espacial. Finalmente, el modelo Matérn generaliza los anteriores a través de un parámetro de suavidad (ν) que controla la forma del variograma, ofreciendo flexibilidad para

ajustar distintos grados de regularidad espacial en entornos heterogéneos (Banerjee, Carlin, & Gelfand, 2015).

El ajuste del variograma teórico requiere estimar sus parámetros —efecto pepita, meseta y rango— de manera que reproduzcan adecuadamente la estructura observada en los datos. Este proceso puede realizarse por ajuste visual, mínimos cuadrados ponderados o métodos de máxima verosimilitud (Kitanidis, 1997). En todos los casos, el criterio fundamental es que el modelo teórico refleje la tendencia general del variograma experimental y conserve la positividad de la matriz de covarianza. Una vez calibrado, su validación empírica se realiza mediante validación cruzada, comparando los valores predichos por kriging con los observados, y evaluando estadísticos de desempeño como el error medio (ME), el error cuadrático medio (RMSE) y la varianza estandarizada (Castrignanò, 2011).

En el campo de la valoración inmobiliaria, la elección del modelo variográfico tiene implicaciones directas sobre la suavidad y la coherencia espacial de las superficies estimadas. Un modelo con rango corto y efecto pepita elevado tenderá a producir superficies fragmentadas con alta variabilidad local, mientras que un modelo Gaussiano o Matérn ajustado con rango amplio generará predicciones más continuas y coherentes con gradientes urbanos. De este modo, la correcta parametrización del variograma se traduce en una representación más precisa de la dinámica espacial del valor del suelo, permitiendo identificar transiciones reales entre zonas de mercado y cuantificar de manera más realista la incertidumbre de estimación (Webster & Oliver, 2007).

La evolución del kriging ha dado lugar a extensiones que amplían su capacidad para representar relaciones espaciales complejas y combinarlo con modelos estadísticos y de aprendizaje más flexibles. Una de las más utilizadas en el ámbito urbano y ambiental es la regresión kriging, también conocida como kriging de residuos, la cual combina un modelo de regresión global —por ejemplo, un modelo hedónico de precios— con el krigeado de los residuos espaciales del modelo (Hengl et al., 2004). De esta manera, la componente determinista explica la variación sistemática asociada a los atributos del bien o del entorno, mientras que la componente estocástica captura la dependencia espacial residual no explicada. Este enfoque híbrido resulta especialmente útil en valoración inmobiliaria, donde

el precio de un inmueble depende tanto de sus características físicas como de su localización relativa dentro de la estructura urbana.

En paralelo, los enfoques Bayesianos y los métodos de inferencia integrada anidada por aproximación (Sus siglas en inglés, INLA) han permitido extender el kriging hacia el tratamiento explícito de la incertidumbre y la inferencia en grandes bases de datos espaciales. Modelos como el SPDE-INLA (Stochastic Partial Differential Equation) aproximan el proceso Gaussiano continuo mediante una representación discreta basada en mallas triangulares, lo que reduce drásticamente el costo computacional sin sacrificar la coherencia espacial (Lindgren, Rue & Lindström, 2011). Estos métodos han demostrado gran eficacia en aplicaciones urbanas, donde la densidad de observaciones y la heterogeneidad espacial imponen restricciones de escalabilidad.

Finalmente, los procesos Gaussianos escalables han posicionado al kriging dentro del campo del machine learning, al compartir la misma formulación probabilística de predicción condicional bajo una estructura de covarianza paramétrica. Este paralelismo ha permitido incorporar núcleos no estacionarios, estimación jerárquica de hiperparámetros y combinación con modelos no lineales, dando origen a enfoques como el deep kriging (Kriging profundo) o los Gaussian Process Regression Networks (Redes de regresión de procesos Gaussianos), (Liu et al., 2021). En conjunto, estas extensiones contemporáneas consolidan la geoestadística como un marco flexible y adaptable que integra métodos clásicos de interpolación con técnicas modernas de inferencia espacial y predicción predictiva basadas en datos masivos.

Para modelar esta estructura espacial, se han desarrollado principalmente dos enfoques: el modelo de rezago espacial (Spatial Autoregressive Model – SAR) y el modelo de error espacial (Spatial Error Model – SEM) (Lee & Yu, 2010). En el modelo SAR, la autocorrelación se introduce directamente en la variable dependiente, bajo la forma:

$$y = \rho W y + X\beta + \varepsilon$$

Donde y es el vector de la variable dependiente, X la matriz de variables explicativas, β el vector de coeficientes, ε el término del error, W la matriz de pesos espaciales y ρ el coeficiente de rezago espacial que mide el grado de influencia de los valores vecinos sobre

cada observación. Este modelo captura procesos de retroalimentación espacial, como ocurre en mercado inmobiliarios donde el precio de un inmueble tiende a ajustarse en función del comportamiento de precios en su entorno inmediato (Bivand, Pebesma, & Gómez-Rubio, 2013). En contraste, el modelo SEM introduce la dependencia en la estructura del error:

$$y = X\beta + u, \qquad u = \lambda W u + \varepsilon$$

Aquí la autocorrelación espacial se interpreta como el efecto de variable no observadas, pero especialmente estructurada que afectan simultáneamente a observaciones cercanas. El coeficiente λ mide la intensidad de esta dependencia en los residuos, y su estimación permite corregir los sesgos derivado de la omisión de variables relevantes que siguen un patrón espacial (Kelejian & Prucha, 2010).

En ambos casos, la matriz de pesos espaciales W desempeñan un papel central, esta matriz define el patrón de vecindad entre observaciones y puede construirse bajo diversos criterios: Contigüidad (adyacencia directa), distancias fijas, numero k de vecinos más cercanos, o funciones decrecientes de distancia. El tipo de matriz elegida incide directamente en los resultados del modelo y debe ser definida según la lógica espacial del fenómeno bajo estudio. Su forma más común es la matriz cuadrada, fila-normalizada, en la cual cada elemento w_{ij} representa la intensidad de la relación entre las unidades espaciales i y j, siendo cero si no son consideradas vecinas (Getis & Aldstadt, 2004).

Además de los modelos SAR y SEM, se han propuesto estructuras más flexibles como el modelo Durbin espacial (Spatial Durbin Model – SDM) que introduce rezagos especiales tanto de la variable dependiente como de las independientes, permitiendo capturar efectos directos e indirecto simultáneamente (Elhorst, 2010); por otro lado, la regresión ponderada geográficamente (Geographically Weighted Regression – GWR) relaja el supuesto de parámetros globales y permite que los coeficientes varíen espacialmente, adaptándose mejor a contexto urbanos altamente heterogéneos donde las relaciones entre variables no son constantes en todo el territorio (Fotheringham, Brunsdon, & Charlton, 2002).

La selección del modelo espacial adecuado debe considerar no solo la presencia estadística de autocorrelación (por ejemplo, diagnosticada mediante el índice de Moran sobre los residuos del modelo OLS), sino también la naturaleza del proceso subyacente: si el interés

es capturar contagios espaciales en la variable explicada, el modelo SAR será más apropiado; si se busca corregir sesgos por omisión espacialmente estructurada, el SEM es preferible; y si se desea explorar efectos de *spillover* entre explicativas, el SDM proporciona una solución intermedia (Waller & Gotway, 2004). En cualquier caso, los modelos de regresión espacial constituyen una herramienta imprescindible para el análisis riguroso de fenómenos georreferenciados, como la formación de precios en el mercado inmobiliario urbano (McMillen, 2010).

En este contexto, la validación cruzada se convierte en una herramienta útil, ya que complementa las pruebas de diagnóstico al ofrecer una medida objetiva de la capacidad predictiva del modelo. La validación cruzada busca reducir el riesgo de sobreajuste (*overfitting*), asegurando que el modelo seleccionado tenga un desempeño robusto frente a nuevas observaciones.

El enfoque más utilizado es la validación cruzada k-fold, que consiste en dividir el conjunto de datos en k particiones de tamaño similar. En cada iteración, se entrena el modelo con k−1 particiones y se valida con la restante, repitiendo el proceso k veces hasta que todas las observaciones hayan sido utilizadas al menos una vez como datos de validación. De esta forma, se calcula un error promedio que refleja el desempeño global del modelo. Matemáticamente, si E_i es el error del modelo en el pliegue i, el error de validación cruzada se expresa como:

$$CV(error) = \frac{1}{k}\sum_{i=1}^{k} E_i$$

Por otra parte, para los datos que alimentan los modelos es necesario establecer una base de datos espacial estructurada bajo principios normativos y técnicos que garanticen integridad, consistencia y trazabilidad. La literatura en información geográfica enfatiza que la calidad de los resultados econométricos y geoestadísticos depende directamente de la solidez en la organización de entidades, atributos y sistemas de referencia espacial, los cuales constituyen la infraestructura sobre la que se soportan tanto el análisis como la modelación.

Las bases de datos espaciales han evolucionado como un componente esencial en la gestión de información geográfica, pues permiten estructurar entidades georreferenciadas

con atributos físicos, espaciales y socioeconómicos de manera integrada. Este tipo de arquitectura se sustenta en los principios definidos por la especificación *Simple Features* del Open Geospatial Consortium, que establece cómo almacenar, consultar y analizar geometrías mediante operaciones estandarizadas en entornos de bases de datos relacionales y sistemas SIG (Open Geospatial Consortium, 2011). La incorporación de atributos temáticos vinculados a las geometrías ha permitido que las bases espaciales se consoliden como repositorios robustos para soportar procesos de modelación econométrica y geoestadística, garantizando consistencia y trazabilidad en proyectos de valoración masiva.

En el contexto colombiano, la unificación bajo el sistema de referencia MAGNA-SIRGAS ha sido determinante para garantizar compatibilidad espacial entre datos catastrales, cartográficos y socioeconómicos. En particular, la proyección oficial MAGNA-SIRGAS 2018 / Origen Nacional (EPSG:9377) constituye el marco de referencia adoptado para el levantamiento y la actualización catastral, asegurando precisión posicional y coherencia en análisis urbanos y regionales (IGAC, 2021). Esta base común respalda la generación de variables espaciales comparables en diferentes escalas de análisis, lo que refuerza la solidez de los modelos aplicados en valoración.

La validación de integridad y consistencia en bases espaciales se apoya en reglas topológicas y controles de calidad. Estas reglas, implementadas en motores SIG y geodatabases empresariales, aseguran que entidades como predios, manzanas y redes viales mantengan relaciones geométricas coherentes, evitando superposiciones indebidas, huecos o discontinuidades. Dichos controles están alineados con las guías de calidad de datos espaciales definidas por la ISO 19157, que establece principios para evaluar exactitud, completitud y consistencia lógica en información geográfica (International Organization for Standardization, 2023). Reforzando la confiabilidad de los insumos y contribuyendo a que los resultados de los modelos mantengan validez técnica y replicabilidad.

5. METODOLOGÍA

5.1. Fase I: ESTRUCTURACIÓN DE LA ARQUITECTURA DE LA BASE DE DATOS ESPACIAL.

La arquitectura de la base de datos espaciales, entendida como la estructura fundamental que organiza, normaliza y garantiza la integridad de la información requerida para el análisis. Busca establecer un modelo lógico coherente, capaz de integrar datos alfanuméricos y geográficos bajo un mismo esquema, apoyado en criterios de calidad, consistencia y compatibilidad con los sistemas de referencia oficiales. De esta manera, se sentaron las bases técnicas que permiten la posterior incorporación de fuentes heterogéneas y el desarrollo de procesos de análisis espacial y econométrico.

5.1.1. Diseño del modelo de base de datos espacial.

El diseño del modelo de base de datos espacial se estructuró bajo un enfoque relacional con soporte geográfico en PostgreSQL, utilizando la extensión PostGIS para el manejo de geometrías. La entidad central definida fue la tabla Datos, donde cada registro corresponde a una observación inmobiliaria caracterizada por atributos físicos, espaciales y socioeconómicos, además de su representación geométrica puntual.

A partir de esta tabla se establecieron relaciones de cardinalidad N:1 hacia catálogos normalizados, con el fin de garantizar consistencia semántica y evitar redundancias. Entre los catálogos más relevantes se encuentran los administrativos (departamentos y municipios) y los clasificatorios (tipo de bien, tipo de inmueble, etc…) complementado con restricciones de dominio aplicadas a campos críticos, con el fin de limitar los valores permitidos y prevenir inconsistencias comunes en el registro de datos inmobiliarios. Los dominios implementados en PostgreSQL se detallan en la siguiente tabla:

Dominio	Tipo de dato	Regla / Valores permitidos	Fuente / Alcance
Departamento	VARCHAR(100)	Lista cerrada de departamentos	DANE
Municipio	VARCHAR(100)	Lista cerrada de Municipios	DANE
Tipo_bien	VARCHAR(30)	Apartamento, Casa, Apartaestudio	Catálogo interno
Tipo_inmueble	VARCHAR(3)	Propiedad Horizontal / No PH	Catálogo interno
Estrato	SMALLINT	1,2,3,4,5,6	Normativa urbana
Nuevo_usado	SMALLINT	Usado, Nuevo	Catálogo interno

Tabla 1 Definición de Dominios

Adicionalmente, se aplicaron restricciones lógicas sobre atributos numéricos, como exigir que el área privada fuera mayor a cero, la edad del inmueble no admitiera valores negativos y que las coordenadas geográficas estuvieran dentro de los rangos válidos para latitud y longitud. Finalmente, la tabla central Datos se indexó con claves primarias y foráneas hacia sus catálogos de referencia, y se incluyeron índices espaciales GiST para el campo geométrico, con el fin de optimizar consultas geográficas y análisis territoriales.

5.1.2. Definición de campos y tipos de datos para variables físicas, espaciales y socioeconómicas.

La definición de campos que conforman la base de datos se realizó considerando los atributos que la literatura especializada y la práctica valuatoria han señalado como determinantes en la formación del valor unitario del metro cuadrado en propiedad horizontal. Este proceso se sustentó tanto en la experiencia del grupo avaluador como en estudios previos sobre dinámica inmobiliaria, lo cual permitió identificar un conjunto de variables que

posteriormente serían contrastadas y evaluadas dentro del modelaje econométrico. En este marco se identificaron tres grupos principales de atributos:

- **Variables físicas**: Comprenden características intrínsecas del inmueble como el área privada (m^2, variable continua), la edad (años), el piso (entera ordinal, asociada a la jerarquía vertical de las unidades en edificios), el estado de conservación (texto controlado: bueno, regular, malo), la condición nueva/usado (binaria 0–1), el estado de terminado (binaria) y las variables de dotación comunal: ascensor, vigilancia, garaje, gimnasio, piscina, salón comunal y ascensor (todas binarias). Adicionalmente, se construyó un índice compuesto de dotación comunal ponderada, que integra estas amenidades con pesos diferenciados (40% vigilancia, 30% garaje, 15% gimnasio y 15% piscina), con el fin de reflejar su aporte conjunto en la formación del precio.
- **Variables espaciales**: Agrupan atributos relacionados con la posición geográfica y la accesibilidad. Incluyen la localización puntual (coordenadas geográficas), así como distancias a equipamientos relevantes: parques, vías principales, colegios, hospitales, estaciones de transporte público (Transmilenio) y centros comerciales. Todas ellas se modelan como variables continuas (decimales en metros).
- **Variables socioeconómicas**: Recogen atributos vinculados al contexto social y a las amenidades de la copropiedad. Se incluye el estrato (ordinal, valores entre 1 y 6).

Cada variable fue tipificada según su naturaleza: enteros para edad, piso y estrato; decimales para área, valores unitarios y distancias; texto controlado para variables administrativas; y binarias para dotación y estado.

5.1.3. Integración del sistema de coordenadas mediante proyección MAGNA-SIRGAS EPSG:9377.

Las observaciones se registraron inicialmente en coordenadas geográficas WGS84 y se reproyectaron al sistema MAGNA-SIRGAS / Origen Nacional (EPSG:9377), adoptado en Colombia para cartografía y catastro. Esta decisión es una buena práctica cuando se requiere medir distancias y áreas en metros, ya que reduce ambigüedades en cálculos espaciales y permite comparar resultados entre capas que comparten el mismo marco de referencia.

La utilización de un ESPG estándar también es de gran ayuda para asegurar compatibilidad con insumos oficiales (capas catastrales, redes viales, zonificaciones) y para evitar errores derivados de mezclar sistemas distintos. En términos operativos, se mantuvo el almacenamiento de la geometría con SRID explícito, y las rutinas de análisis espacial (cálculo de distancias a equipamientos, buffers, áreas de influencia) se ejecutaron sobre la versión proyectada de las capas.

Como control de calidad, se verificó que todas las capas involucradas compartieran el mismo SRID antes de su uso, y se documentó la transformación aplicada (origen y destino, fecha y parámetros). Este registro garantiza la trazabilidad del proceso y reproducibilidad de resultados en etapas posteriores de análisis y modelación.

5.1.4. Validación de la arquitectura mediante pruebas de integridad, topología y consistencia.

La validación de la arquitectura se organizó en tres frentes complementarios: integridad de datos, consistencia referencial y controles espaciales/topológicos. Este conjunto de verificaciones sirve para asegurar que el esquema relacional y su componente espacial funcionen de forma estable antes de iniciar el ETL y el modelado.

Se aplicaron restricciones NOT NULL en campos clave como identificador, tipo de bien, estrato, geometría, y CHECK para reglas lógicas básicas: área privada mayor a cero,

edad mayor a cero, estrato perteneciente al valor dentro del dominio y coherencia de valor para piso que debe ser un entero positivo. Donde correspondía, se añadieron restricciones de unicidad sobre combinaciones que no deben repetirse. Estas restricciones previenen duplicados accidentales y entradas imposibles en términos de reglas de negocio.

Las columnas administrativas se validaron con claves foráneas hacia catálogos normalizados: departamento, municipio, tipo de bien, tipo inmueble, estrato y nuevo/usado. Además, se verificó la coherencia jerárquica es decir que el municipio pertenezca al departamento indicado mediante Foreign Keys (FKs) encadenadas o validación en inserción. Para mantener la limpieza de catálogos, se definieron índices B-tree sobre los códigos y nombres normalizados, lo cual también mejora el desempeño de consultas y uniones.

Las geometrías se almacenaron con SRID explícito y se verificó que todas las capas del esquema compartieran el EPSG:9377 antes de cálculos métricos. Se ejecutaron controles de extensión espacial es decir que los puntos caigan dentro del límite de la UPL o área de estudio y se descartaron registros fuera de dominio mediante una máscara espacial. También se revisó la validez de geometrías (ST_IsValid) y la ausencia de valores nulos en el campo geométrico. Para capas de referencia lineales y poligonales, se comprobaron reglas sencillas de topología operativa. Finalmente, la clase Dato se encuentra estructurada de la siguiente manera:

CLASE					
Dato	Clase especializada que describe la unidad espacial básica transaccional.				
Atributo	***Descripción***	***Cardinalidad***	***Tipo de dato***	***Dominio***	***Not Null***
ID	Identificador único consecutivo de la base de datos.	1	Serial	-	Si
Departamento	Departamento al cual pertenece.	1..1	Structure	Departamento	Si
Municipio	Municipio al cual pertenece.	1..1	Structure	Municipio	Si
Barrio	Barrio al cual pertenece.	1.. *	Varchar	Barrio	Si
Nomenclatura	Dirección.	1.. *	Varchar	-	Si
Fecha	Fecha del avalúo.	1	Date	-	Si

Nuevo / Usado	Estado del inmueble.	1..1	Boolean	Nuevousado	Si
Área Privada	M2 del área privada de inmueble	1..1	Double		Si
Valor Área Privada	Valor de m2 de área privada	1..1		-	Si
Estrato	Estratificación del inmueble	1..1	Structure	Estrato	Si
Longitud	Longitud del inmueble	1..1	Double	-	Si
Latitud	Latitud del inmueble.	1..1	Double	-	Si
Edad	Edad del inmueble	1..1	Int	-	Si
Piso	Nivel del inmueble dentro de la edificación.	1..1	Int	-	Sí
Año	Año de la transacción / avalúo.	1..1	Int	-	Sí
Vigilancia Privada	Si cuenta o no con dotación comunal.	1..1	Boolean	-	Sí
Portería		1..1	Boolean	-	Sí
Ascensor		1..1	Boolean	-	Sí
Garaje Visitante		1..1	Boolean	-	Sí
Gimnasio		1..1	Boolean	-	Sí
Salón Comunal		1..1	Boolean	-	Sí
Piscina		1..1	Boolean	-	Sí
Dotación Comunal	Índice de dotación de copropiedad (equipamientos y zonas comunes).	1..1	Double	-	Sí
Distancia a Parques	Distancia a parque (m).	1..1	Double	Distancia_POI	Sí
Distancia a Colegios	Distancia a colegio (m).	1..1	Double	Distancia_POI	Sí
Distancia a Centros de Salud	Distancia a centro de salud (m).	1..1	Double	Distancia_POI	Sí
Distancia a TransMilenio	Distancia a estación TM (m).	1..1	Double	Distancia_POI	Sí
Distancia a Centro Comercial	Distancia a centro comercial (m).	1..1	Double	Distancia_POI	Sí
Geom	Campo Geométrico del punto	-	Geometry	-	Si

Tabla 2 Estructura Relacional de la tabla principal.

5.2. Fase II: Extracción, Transformación y Carga de los datos (ETL)

Para consolidar fuentes heterogéneas en un esquema único y trazable: se definieron criterios de cobertura temporal y espacial, normaliza estructuras y unidades, resuelve duplicidades y discrepancias entre fuentes, y aplica reglas de calidad y consistencia antes de su integración en PostgreSQL/PostGIS con SRID homogéneo. El proceso incorpora metadatos y diccionario de datos, control de versiones y bitácoras de cambio, de modo que el conjunto resultante quede listo para análisis espacial y econométrico. A continuación, se detallan sus etapas. El tamaño de la muestra n es de 1203 observaciones.

5.2.1. Recolección y organización de fuentes: catastro, DANE, SNR, normativas, Appraiser S.A.S y web scraping.

Se formo una base de insumos heterogéneos en un esquema único de observaciones inmobiliarias. Para ello se definieron criterios comunes de cobertura temporal (2023–2025), ámbito espacial (UPL de estudio y áreas colindantes para variables de entorno), licenciamiento/uso (datos abiertos o datos con autorización) y formato (prioridad a estructuras tabulares y geoespaciales interoperables). Con base en esos criterios se integraron las siguientes fuentes:

- **Catastro**: capas geoespaciales y atributos prediales/urbanísticos utilizados como referencia territorial (límites administrativos, barrios/UPL, red vial, equipamientos).
- **DANE**: variables sociodemográficas y de contexto (estrato) alineadas a lo asignado por manzana.
- **SNR (Superintendencia de Notariado y Registro):** identificadores registrales y atributos mínimos vinculados a la dinámica transaccional como uso de los inmuebles, histórico de transacciones, esto con el fin de contar con un respaldo para la verificación de transacciones actuales.

- **Normativa catastral y técnica**: documentos que establecen definiciones y convenciones (zonas de protección ambiental, áreas de interés ecosistémico, etc..). Su función es verificar que puntos de las transacciones no se encuentren en zonas con posibles restricciones ambientales o ecosistémicas que pudieran influir en el valor.
- **Appraiser S.A.S.**: Observaciones primarias de avalúos (atributos físicos, espaciales y socioeconómicos) con metadatos internos sobre fecha de levantamiento y criterios valuatorios.
- **Web scraping (ofertas en el portal inmobiliario de Fincaraíz)**: registros de mercado publicados por propietarios o comercializadores (precios ofrecidos, características y ubicación). Se aplicaron filtros de calidad mínimos (coordenadas válidas dentro de la zona) y control de duplicados por medio de análisis de imágenes por métodos de campos vectoriales.

5.2.2. Limpieza y transformación de datos

Se realizó normalización operativa mínima (tipos de dato correctos, conversión de separadores decimales, trim y corrección de codificación) y detección de duplicados intra-fuente mediante claves compuestas acordes a la naturaleza de cada insumo (p. ej., combinación de dirección/nomenclatura con fecha en ofertas; o identificadores propios en registros técnicos). Los duplicados estrictos se eliminaron y los casos dudosos se resolvieron por consistencia interna del registro.

Respecto a valores faltantes, se definieron criterios por variable: eliminación cuando el campo es crítico para el modelaje (dotación comunal, coordenadas válidas, área) e imputación conservadora solo en atributos discretos no críticos (por ejemplo, inferir barrio a partir de la ubicación si el punto cae inequívocamente dentro de un polígono administrativo). Con el propósito de preservar observaciones útiles, se implementó un enriquecimiento cruzado entre ofertas y Appraiser S.A.S. para variables que en portales son incompletas. La vinculación combinó criterios espaciales (proximidad del punto y pertenencia al mismo

barrio/UPL) y textuales (concordancia de dirección normalizada o alta similitud), priorizando el valor de Appraiser cuando la oferta estaba vacía o era inconsistente y dejando trazabilidad de la procedencia campo a campo. Con esto, cada fuente quedó limpia, tipada y con completitud suficiente para pasar a la integración y estandarización global, sin adelantar transformaciones analíticas que se abordan en etapas posteriores.

5.2.3. Estandarización de variables, eliminación de inconsistencias

Con las fuentes ya integradas, el objetivo fue uniformar el significado de los valores y resolver discrepancias entre orígenes sin modificar la arquitectura de la base. Se construyeron tablas de mapeo y catálogos de equivalencias para unificar taxonomías y etiquetas: departamentos y municipios (con sus nombres), mientras que barrios y categorías descriptivas (tipo de bien, tipo de inmueble, estado de conservación) se alinearon a listas únicas, eliminando sinónimos y variantes ortográficas.

Para conflictos inter-fuentes en una misma observación, se aplicaron reglas de consolidación (survivorship): (i) prioridad de fuente por sensibilidad de la variable (p. ej., para edad y dotación prevaleció Appraiser frente a oferta), (ii) congruencia espacio-temporal (coincidencia territorial y cercanía de fechas) y (iii) consistencia interna del registro. El valor que sobrevivió quedó marcado con su origen a nivel de campo (oferta, Appraiser o derivado), y los alternativos se conservaron como traza cuando aportaban evidencia para auditoría.

Además, se abordó la normalización económica de ofertas para hacer comparables los precios reportados: (a) se aplicó un descuento del 5% como tasa estándar de negociación para aproximar el valor transado, (b) cuando existía garaje, se descontó su valor según tipología detectada en el scraping (p. ej., sencillo cubierto, doble cubierto), apoyando la validación del estrato con la capa oficial del IDECA y (c) en inmuebles en obra gris, se utilizó un factor de homogeneización de 1,10 para llevar el precio al referente de “terminado con acabados”. Con estos ajustes, el valor por m² resultante queda homogeneizado y es comparable entre ofertas de distinta procedencia.

5.2.4. Carga de datos al sistema espacial.

La carga se realizó en PostgreSQL/PostGIS siguiendo un flujo de dos etapas. Primero, los archivos consolidados se ingresaron a un esquema de staging (zona de preparación) mediante procesos de carga masiva (COPY … FROM … CSV HEADER). En esta fase se aplicaron verificaciones automáticas de tipos y un control preliminar de errores, manteniendo el proceso ligero (sin llaves foráneas ni restricciones complejas) para poder inspeccionar rápidamente los lotes. Cada carga quedó identificada con metadatos de origen, fecha y versión.

Superado este control inicial, los registros válidos se trasladaron al esquema core (núcleo) mediante inserciones controladas (INSERT … SELECT) que sí activan las reglas definidas en la Fase I: dominios, CHECK, NOT NULL y FOREIGN KEY. En este paso se poblaron las columnas administrativas con nombres normalizados (DIVIPOLA para departamento y municipio) y las categóricas se contrastaron contra sus catálogos. Los registros que no cumplieron alguna restricción se enviaron a una tabla de incidencias, con el motivo de rechazo documentado para su corrección posterior.

La geometría puntual se construyó a partir de longitud y latitud con ST_SetSRID(ST_MakePoint(longitud, latitud), 4326) y se reproyectó a EPSG:9377 usando ST_Transform, de modo que todas las capas compartieran el mismo sistema de referencia. Antes de habilitar la tabla para análisis, se verificaron el SRID; los registros fuera de ámbito quedaron marcados con un indicador de exclusión.

Para mejorar el desempeño, se creó un índice GiST sobre la columna geométrica. Al cierre de cada lote se ejecutó ANALYZE y, cuando fue pertinente, se aplicó particionamiento lógico por periodo (p. ej., año) o por ámbito territorial para acelerar consultas repetitivas. Toda la operación se ejecutó dentro de transacciones (BEGIN/COMMIT) para asegurar atomicidad. Ante cualquier error, se realizó ROLLBACK y se dejó registro en la bitácora operacional (archivo y fila afectados, restricción activada, fecha y responsable). Con ello, el conjunto quedó disponible en el esquema espacial para consulta SIG, análisis exploratorio y las etapas de discretización y modelado.

5.2.5. Ingreso estructurado en la base SIG.

El ingreso estructurado en el SIG consistió en preparar y organizar, dentro del proyecto geográfico, las capas necesarias para el cálculo consistente de métricas de accesibilidad por red. En esta etapa se cargaron las capas operativas con nombres normalizados y CRS homogéneo (EPSG:9377). Con ese marco, se incorporaron como capas de destino los parques, colegios, hospitales/centros de salud, estaciones de TransMilenio y centros comerciales, y como capa de tránsito la red vial de Bogotá. Para que el cálculo de rutas fuera válido, se saneó la topología de la red: eliminación de segmentos colgantes no justificados, inserción/corrección de nodos en intersecciones, ajuste de direccionalidades cuando aplicaba y unificación de geometrías bajo tolerancias de snapping. Este alistamiento topológico se registró como parte del ingreso, de modo que la red quedara lista para ser consumida por el motor de análisis.

Con el fin de reflejar la accesibilidad efectiva más allá del límite administrativo, el proyecto SIG incluyó un buffer de 500 m alrededor de la UPL, y se realizó el recorte (clip) de la red vial y de los equipamientos al ámbito “UPL + 500 m”. Esta ampliación es coherente con el objetivo del ingreso: asegurar que las capas disponibles en el proyecto cubran el espacio donde realmente se produce la interacción con los equipamientos más cercanos.

Una vez estructuradas las capas, se configuró la red de análisis (coste en metros, reglas de conexión y tolerancias) y se ejecutó la herramienta de “ruta más cercana” tomando como orígenes los puntos de observación inmobiliaria y como destinos cada una de las capas de equipamientos. Los resultados (distancias mínimas por red) se exportaron a una tabla de salida con campos como dist_parque, dist_tm, dist_colegio, dist_hospital y dist_cc, vinculados al identificador único de cada observación. Finalmente, estas métricas se reintegraron a la base espacial para su uso en etapas posteriores de análisis.

Como cierre del ingreso, se verificaron SRID, validez geométrica, ausencia de rutas nulas y coherencia del dominio espacial (rutas confinadas al ámbito definido). También se documentaron los parámetros de red y la fecha de las capas. Con ello, el proyecto SIG quedó

estructurado y operativo: contiene las capas, reglas y salidas necesarias para producir distancias de accesibilidad por red de forma reproducible y consistente con el área de estudio.

5.3. Fase III: Análisis exploratorio de datos.

Con el fin de caracterizar la variable objetivo y orientar decisiones de preprocesamiento y modelado, se efectuó un análisis exploratorio apoyado en estadística descriptiva, pruebas de normalidad y correlación, diagnóstico de heterocedasticidad y visualización.

5.3.1. Estadística descriptiva: medidas de tendencia central, dispersión y detección de outliers.

El comportamiento de valor_aprivada COP/m^2 evidencia un mercado concentrado en torno a un bloque central relativamente compacto, con una mediana cercana a los 5,6 millones y un núcleo intercuartílico que se sitúa aproximadamente entre 4,8 y 6,7 millones. Este intervalo intercuartílico o $RIQ = Q3 - Q1$ resume la "zona densa" de precios donde cae la mitad de las observaciones y es útil como referencia estable para comparaciones territoriales y temporales, porque es poco sensible a valores extremos. La media ($\approx 5{,}9 millones$) se ubica por encima de la mediana, lo que sugiere que la cola alta ejerce tracción al alza sobre el promedio, un patrón habitual en mercados urbanos donde una fracción pequeña de unidades de mayor categoría concentra precios elevados.

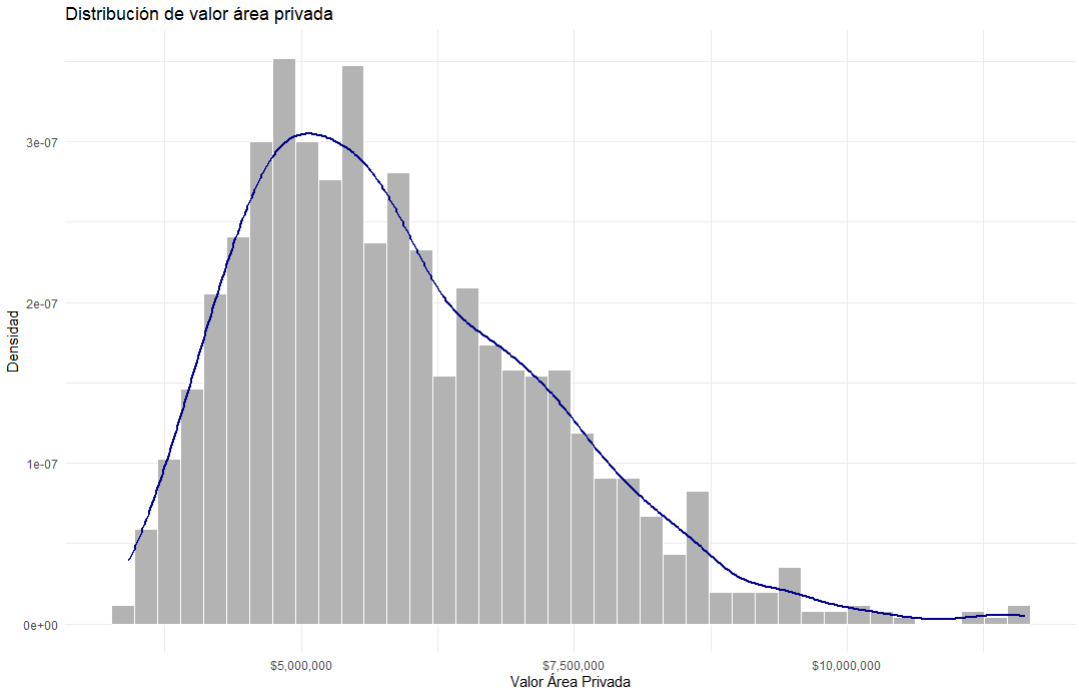


Ilustración 2 Histograma de los datos crudos.

Desde la perspectiva de forma de la distribución, los gráficos de histograma con densidad y boxplot apuntan a una distribución unimodal con asimetría positiva. Es decir, la mayor parte de los inmuebles se agrupa en el bloque intermedio, mientras que un subconjunto reducido se extiende hacia valores altos. Esta morfología acompañada de una curtosis ligeramente superior a la normal suele hallarse en mercados con heterogeneidad de calidades: combinaciones de atributos estructurales, amenidades de copropiedad y ubicaciones privilegiadas que, en conjunto, empujan el precio por encima del perfil típico.

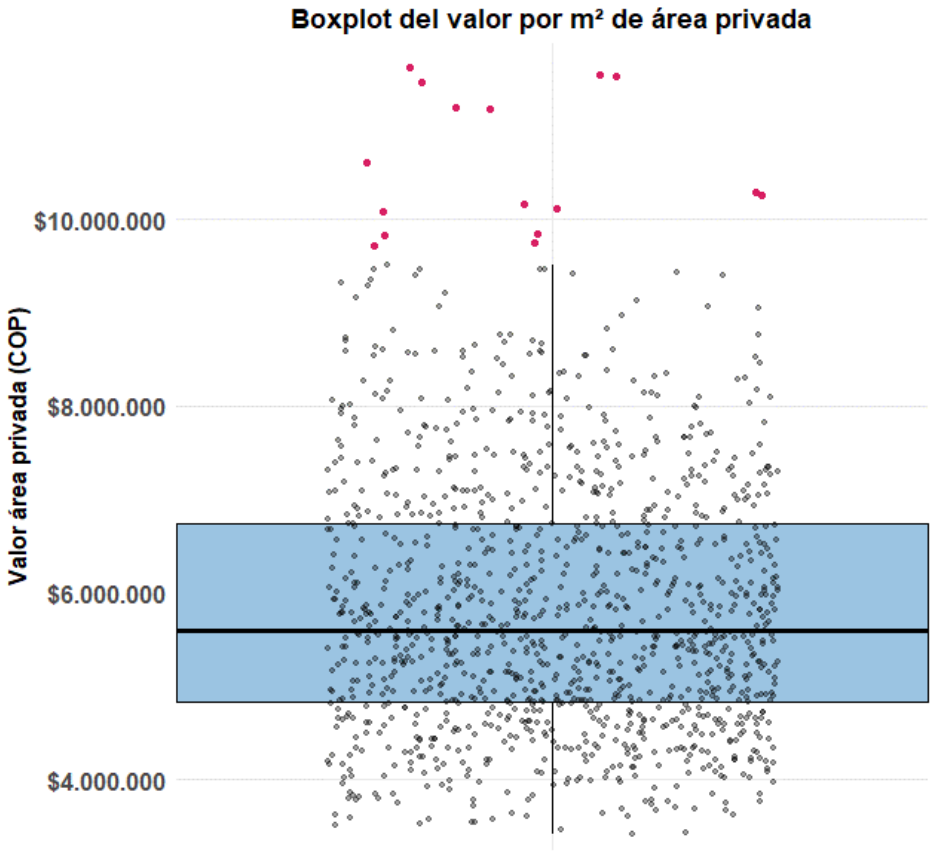


Ilustración 3 Boxplot de los datos crudos

Para la detección de atípicos se aplicó la regla de Tukey ($1{,}5 \cdot RIQ$) sobre el eje de COP/m^2. El límite superior de esta regla queda alrededor de 9,6 millones, y no se detectaron atípicos inferiores; por tanto, los outliers se concentran exclusivamente en la cola alta ($\approx 9{,}6 - 11{,}6 millones$). La inspección cruzada con metadatos sugiere que muchos de estos casos pertenecen a segmentos de alta gama (estratos superiores o localizaciones con ventajas claras), por lo que no deben asumirse como errores por defecto. En términos de tratamiento, se adoptó un criterio escalonado: (i) conservar observaciones extremas verosímiles cuando su contexto espacial y tipológico lo respalda; (ii) revisar/corregir registros con señales de captura anómala (separadores, áreas o coordenadas incoherentes); y (iii) marcar los extremos para análisis de sensibilidad y estimación robusta, evitando modificaciones irreversibles de la muestra base.

El valor interpretativo de este diagnóstico es doble. Por un lado, la mediana y el RIQ ofrecen un retrato resistente al sesgo de la cola y constituyen puntos de apoyo para describir

el mercado sin que las unidades premium oculten el comportamiento típico. Por otro, la cola derecha—aunque reducida—es informativamente relevante: alerta sobre nichos de alta disposición a pagar y anticipa la conveniencia de contrastar, más adelante, especificaciones que acomoden no linealidades y mecanismos de robustez frente a valores extremos.

5.3.2. Pruebas de normalidad y correlación: Shapiro-Wilk, Kolmogorov-Smirnov, Pearson/Spearman.

La evaluación de normalidad se aplicó a la variable dependiente principal (valor_aprivada, expresada en (COP/m^2) y a los residuos del modelo lineal de referencia. Los contrastes estadísticos implementados fueron Shapiro–Wilk, Kolmogorov–Smirnov y pruebas complementarias como Anderson–Darling y D'Agostino–Pearson, cada una sensible a distintos aspectos de la distribución.

Los resultados fueron consistentes entre sí. El test de Shapiro–Wilk arrojó un estadístico $W \approx 0{,}953$ con $p < 2{,}2 \times 10^{-16}$, lo que implica un rechazo claro de la hipótesis de normalidad. El test de Anderson–Darling reportó $A \approx 13{,}1$ con el mismo desenlace, y D'Agostino–Pearson obtuvo $X^1 \approx 171{,}9$ con significancia igualmente inferior a 0,001. El Kolmogorov–Smirnov corroboró este panorama al detectar diferencias significativas respecto a la distribución normal teórica. En conjunto, estos hallazgos confirman que tanto la variable bruta como los residuos presentan desviaciones significativas respecto al patrón Gaussiano.

Este comportamiento tiene implicaciones metodológicas. Aunque no invalida el uso de modelos lineales, sí indica que la variabilidad en los extremos puede afectar la estimación, lo cual sugiere dos rutas de ajuste para fases posteriores: aplicar transformaciones de tipo Box–Cox o Yeo–Johnson que estabilicen la varianza, o acudir a metodologías robustas menos sensibles a valores atípicos.

En paralelo al análisis de normalidad, se evaluó la correlación entre variables explicativas mediante la construcción de matrices de Pearson y Spearman. La primera se

enfocó en asociaciones estrictamente lineales, mientras que la segunda permitió identificar relaciones monótonas de carácter no lineal. Ambas matrices ofrecieron resultados concordantes: el valor_aprivada mostró correlaciones positivas notables con la edad invertida ($r \approx 0{,}72$), con la dotación comunal ($r \approx 0{,}65$) y con la presencia de ascensor ($r \approx 0{,}49$), lo que refleja que los inmuebles más recientes, con mayor oferta de servicios y accesibilidad, tienden a registrar precios unitarios más altos. El estrato socioeconómico presentó un nivel de correlación moderado ($r \approx 0{,}44$), en consonancia con la segmentación urbana.

Asimismo, se identificaron correlaciones negativas de interés. La distancia a colegios y a centros comerciales mostró asociaciones en torno a –0,25, lo que implica que a mayor lejanía de estos equipamientos, el valor por metro cuadrado tiende a reducirse. Aunque de magnitud baja a moderada, este patrón es coherente con la lógica de accesibilidad que caracteriza al mercado inmobiliario en entornos urbanos.

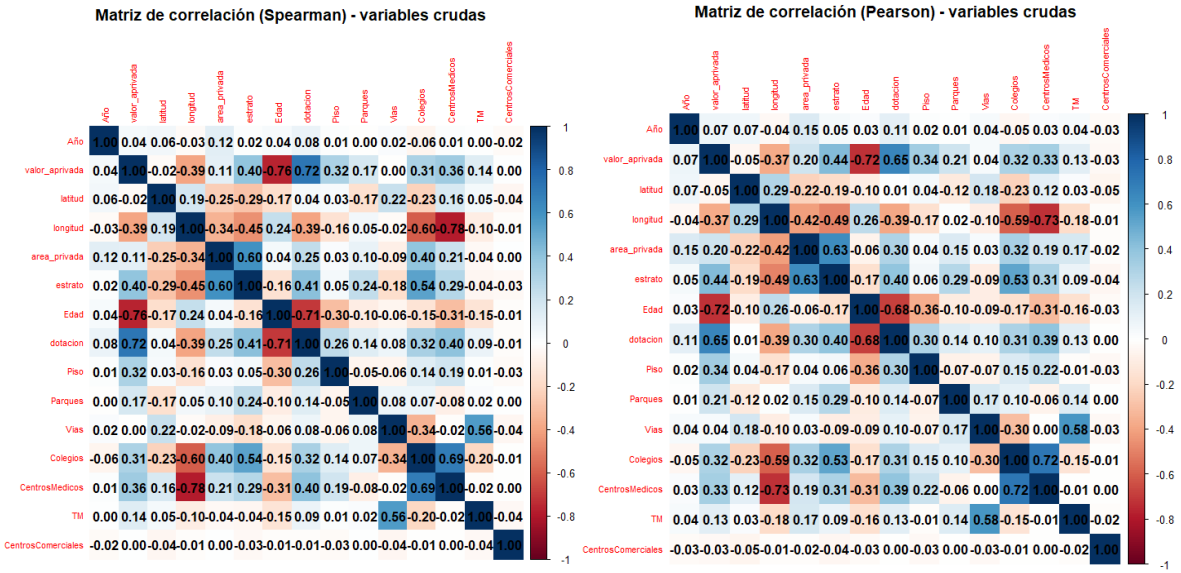


Ilustración 4 Matriz Pearson y Spearman

5.3.3. Representación gráfica: histogramas, QQ-plots y diagramas de dispersión.

Los gráficos se elaboraron a partir de los residuos del modelo lineal múltiple base, definido para el análisis preliminar como:

$$\begin{aligned}\text{Valor}_i = \beta_0 &+ \beta_1(\text{inv_area}_i) + \beta_2(\text{edad_invertida}_i) + \beta_3(\text{dotación}_i) + \beta_4(\text{Piso}_i) \\ &+ \beta_5(\text{nuevo_usado}_i) + \sum_{k=2}^{6} \gamma_k E_{ik} + \beta_6(\text{ascensor}_i) + \beta_7(\text{terminado}_i) \\ &+ \beta_8(\text{Año}_i) + \beta_9(\text{inv_parques}_i) + \beta_{10}(\text{inv_vías}_i) + \beta_{11}(\text{inv_colegios}_i) \\ &+ \beta_{12}(\text{inv_centros}_i) + \beta_{13}(\text{inv_TM}_i) + \varepsilon_i]\end{aligned}$$

donde $(E_{ik} = 1)$ si el inmueble pertenece al estrato $(k)((k = 1, \dots, 6))$ y (0) en otro caso. Este modelo incorpora atributos físicos, constructivos, socioeconómicos y de entorno, estimados mediante mínimos cuadrados ordinarios. Los residuos (ε_i) representan las diferencias entre los valores observados y los ajustados.

El histograma de residuos muestra una concentración alrededor de cero y colas más extendidas que las de una distribución normal teórica. Este comportamiento refleja la presencia de observaciones alejadas del centro que incrementan la dispersión en los extremos. Desde un enfoque estadístico, sugiere una distribución leptocúrtica con asimetrías leves; en términos prácticos, anticipa una mayor incertidumbre de predicción en determinados segmentos del mercado.

El Q–Q plot de residuos complementa esta lectura. En los cuantiles intermedios, los puntos se alinean razonablemente con la diagonal, indicando un comportamiento próximo al normal para la mayoría de las observaciones. En los cuantiles superiores, sin embargo, se evidencia una desviación sistemática respecto a la línea de referencia, lo que confirma la existencia de colas pesadas y de un subconjunto de valores atípicos elevados. Esta desviación sugiere que los estimadores basados en el supuesto de normalidad estricta pueden ser sensibles a tales observaciones; por tanto, se recomienda considerar ajustes robustos o transformaciones que mitiguen su influencia en la estimación de precisión.

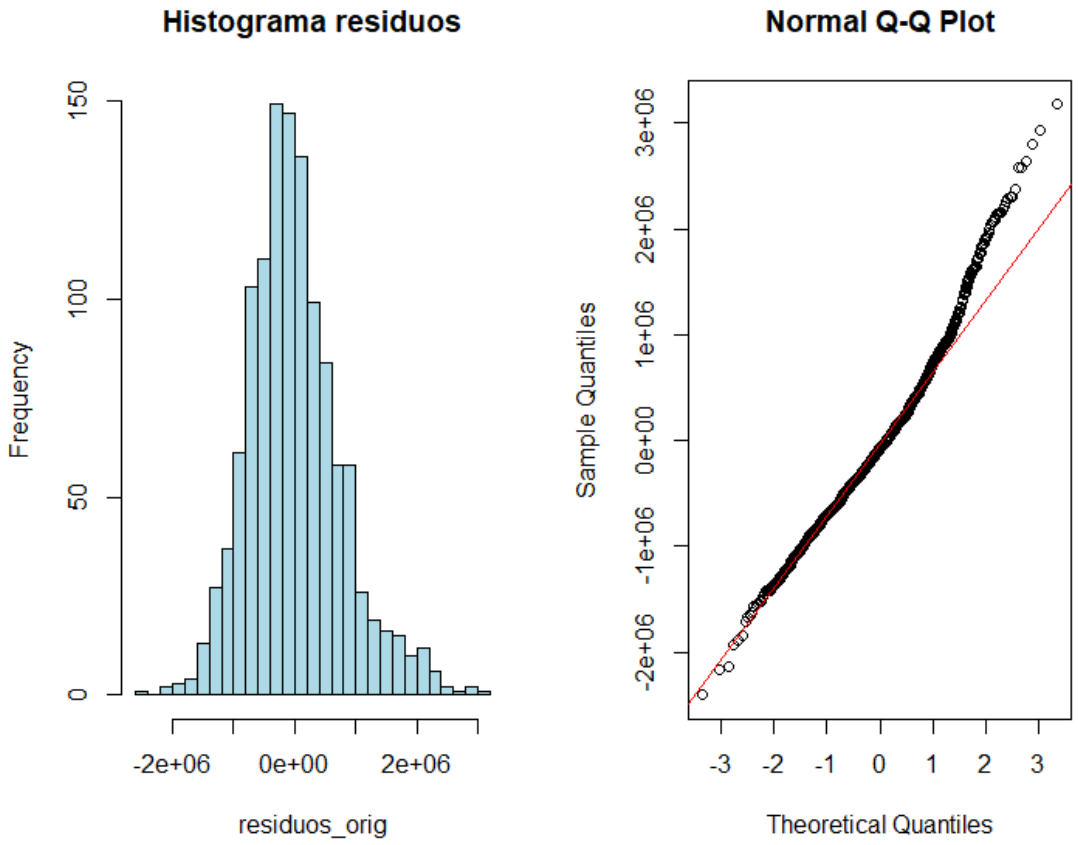


Ilustración 5 Histograma y Q-Q plot del LM de referencia

Por su parte, el diagrama de dispersión de residuos frente a valores ajustados muestra una nube sin tendencias estructuradas alrededor de la línea horizontal, lo que es coherente con la independencia media de los errores. No obstante, la amplitud de la nube crece a medida que se incrementan los valores ajustados. Este ensanchamiento es indicio de heterocedasticidad, es decir, la varianza del error no es constante en todo el rango de predicción y resulta mayor en los segmentos de valor más alto. Esto aconseja evaluar alternativas como la estimación con varianzas robustas tipo HC (Heteroskedasticity-Consistent), las ponderaciones por mínimos cuadrados ponderados o WLS (Weighted Least Squares), o la aplicación de transformaciones estabilizadoras en etapas posteriores, así como reportar intervalos de predicción que reflejen esta variabilidad diferencial.

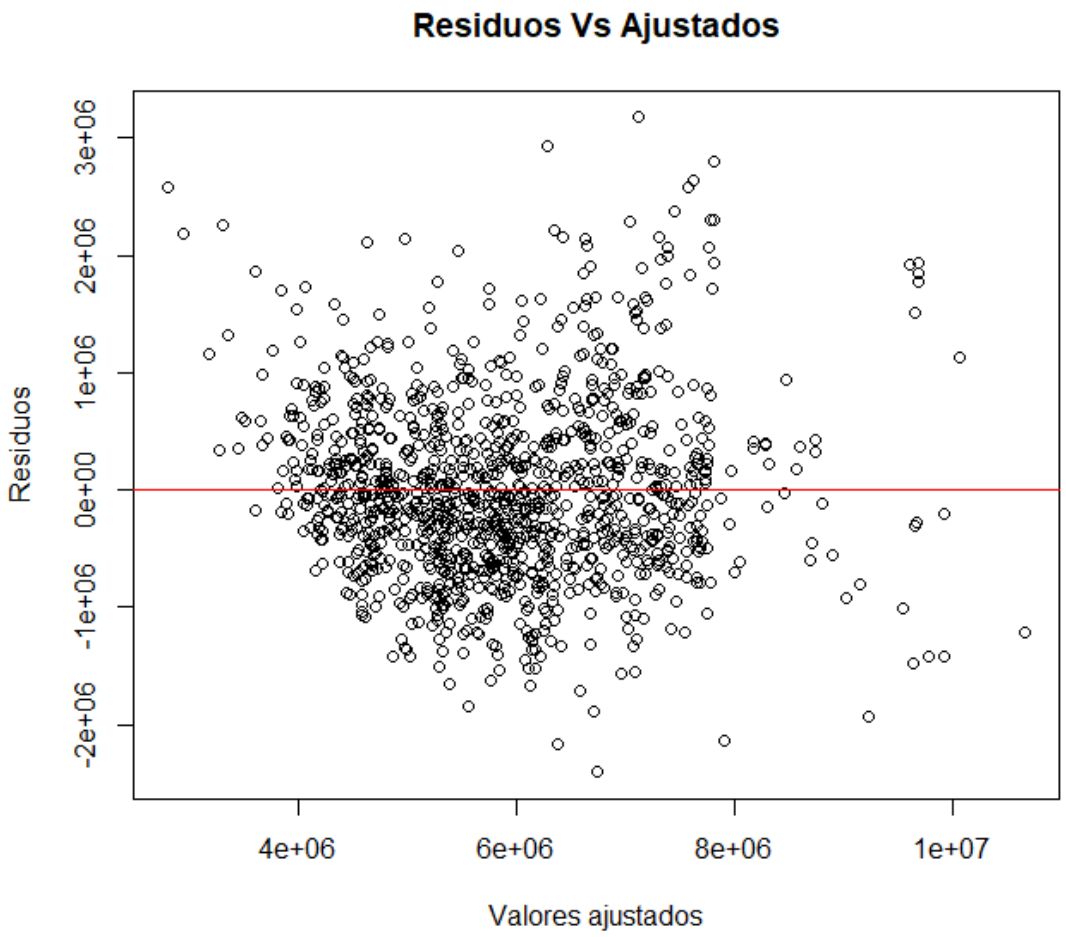


Ilustración 6 Diagrama de dispersión de residuos.

Estas tres piezas de evidencia, consideradas en conjunto, describen un mercado con un núcleo de transacciones de comportamiento relativamente estable y con dispersión acotada, junto a un segmento superior donde la variabilidad es mayor. En la práctica inmobiliaria, este patrón suele asociarse a inmuebles con atributos singulares (acabados, altura, vistas, dotaciones del edificio o del entorno inmediato) y, en menor medida, a localizaciones con demanda sostenida que amplifican la varianza, sin que ello suponga una ruptura del comportamiento general. La lectura es consistente con mercados urbanos heterogéneos: la predicción es más estable en el tramo medio y requiere cautela adicional en el tramo alto.

5.4. Fase IV: Discretización de datos

La discretización se utiliza como paso intermedio para incorporar adecuadamente ciertos atributos en los modelos econométricos y espaciales. Algunas variables continuas presentan escalas difíciles de interpretar o relaciones no lineales; otras variables categóricas tienen niveles muy específicos o poco frecuentes. Para resolverlo, se definieron reglas de recodificación que agrupan valores en rangos o categorías con soporte muestral suficiente y sentido sustantivo. Este procedimiento mejora la interpretabilidad de los coeficientes, reduce varianza por niveles raros y atenúa problemas de multicolinealidad y sobreajuste. Asimismo, facilita contrastes entre grupos y la inclusión de efectos fijos o indicadores espaciales cuando corresponda.

5.4.1. Clasificación de variables categóricas.

Se definieron reglas operativas para dos variables categóricas utilizadas en la estimación: la condición comercial (nuevo/usado) y el estado de terminación (terminado/obra gris). Estas reglas garantizan que cada categoría sea inequívoca, verificable con los insumos disponibles y consistente con la práctica inmobiliaria local, de modo que los contrastes entre niveles reflejen diferencias observables y no convenciones de codificación.

La condición comercial se determina mediante un criterio combinado temporal y transaccional. Un inmueble se clasifica como nuevo cuando su antigüedad constructiva es inferior a cinco años y no registra más de una compraventa desde la entrega formal (es decir, cero o una operación). La consideración simultánea de antigüedad y rotación evita clasificar como “nuevo” unidades con varios ciclos de reventa en un periodo corto o, en el extremo opuesto, inmuebles antiguamente construidos que no han sido transferidos recientemente. Cuando alguno de los dos criterios no se cumple, el inmueble se clasifica como usado. Las mejoras o remodelaciones dentro del quinquenio no modifican la categoría, pues esta variable describe el tramo inicial del ciclo de vida y de mercado de la unidad; las intervenciones se reflejan, en su caso, en variables de estado físico y no en la condición comercial.

El estado de terminación distingue entre terminado y obra gris con base en acabados y condiciones de habitabilidad. Se considera obra gris la unidad que carece de acabados definitivos —pisos sin instalar o en obra, muros y cielos sin terminación, baños sin enchapes ni remates— pero cumple mínimos de habitabilidad: dispone de al menos un aparato sanitario instalado, mobiliario básico de cocina (mesón y lavaplatos funcionales) y servicios públicos domiciliarios operativos (acueducto, energía y, cuando corresponda, gas). Se clasifica como terminado cuando presenta acabados completos y operativos en pisos, muros y techos, baños equipados (aparatos y grifería instalados y en funcionamiento), cocina completamente instalada (mobiliario, mesones y punto hídrico operativo) y servicios públicos activos.

5.4.2. Codificación de variables: creación de dummys y codificación ordinal.

La codificación se orientó a representar de manera operativa los atributos cualitativos relevantes, garantizando interpretabilidad directa de los parámetros y consistencia con los criterios definidos en la fase de clasificación. Para las variables binarias se empleó codificación dummy con asignación explícita 1/0, y para variables con jerarquía natural se preservó el orden mediante codificación ordinal, evitando reetiquetados que alteren la semántica original.

En las binarias se adoptó codificación estricta y reproducible. La presencia de ascensor se representó con 1 cuando el atributo está disponible y 0 en caso contrario, manteniendo los valores faltantes como ausentes para no introducir supuestos. La condición comercial se codificó con 1 para nuevo y 0 para usado, conforme al criterio operativo previamente establecido (antigüedad constructiva < 5 años y sin más de una compraventa desde la entrega formal). El estado de terminación se codificó con 1 para terminado y 0 para obra gris, atendiendo a los mínimos de habitabilidad y completitud de acabados definidos.

En el caso del estrato socioeconómico, la variable se declaró como factor en R mediante as.factor(estrato), de modo que el modelo lineal genera internamente un conjunto

de variables indicadoras (dummys) para cada nivel de la escala reglamentaria (1–6), tomando el primer nivel como referencia. Este procedimiento implica que cada categoría se representa mediante una variable binaria independiente, donde el valor 1 indica pertenencia al nivel correspondiente y 0 su ausencia. De esta forma, el modelo estima un conjunto de coeficientes que expresan la diferencia promedio en el valor por metro cuadrado respecto al estrato base, manteniendo la comparabilidad entre grupos sin suponer una relación estrictamente continua o equidistante entre ellos.

El tratamiento del estrato como variable categórica —y no como ordinal— responde a la naturaleza administrativa y socioeconómica de su asignación. Aunque el sistema de estratificación en Colombia presenta una secuencia numérica de 1 a 6, dicha numeración no representa un intervalo regular en términos de ingreso, infraestructura o dotación urbana, por lo que no sería metodológicamente adecuado modelarlo como una escala de razón o intervalo. La codificación dummy permite capturar variaciones discretas y heterogéneas entre niveles sin imponer una gradación lineal que distorsione la interpretación de los parámetros.

5.4.3. Agrupación de variables continuas en intervalos o cuantiles.

Tras el análisis del tratamiento de la variable no se implementó la agrupación de variables continuas en tramos. La opción se evaluó para la edad del inmueble, el área privada y los indicadores de proximidad o accesibilidad; sin embargo, se decidió conservar la escala continua —junto con las transformaciones ya documentadas, como edad invertida y recíprocos— para preservar la variación informativa completa y evitar la introducción de puntos de corte que no provienen de evidencia empírica concluyente. Mantener el continuo permite estimar gradientes y curvaturas sin imponer límites discretos que podrían segmentar artificialmente el comportamiento de la variable dentro de su soporte.

La discretización por intervalos o cuantiles tendió a reducir la potencia estadística al convertir observaciones distintas en un mismo grupo y puede genero resultados sensibles a pequeñas variaciones de los cortes, especialmente considerando que nuestra muestra tiene

una cola asimétrica, y donde la acumulación se encuentra en un rango estrecho o empate frecuente. Adicionalmente, la creación de tramos aumenta el número de parámetros si estos se incorporan como indicadores en el modelo, lo que consume grados de libertad y favorece la sobreparametrización sin ganancias claras en interpretabilidad.

El mantenimiento de la escala continua fue acompañado por controles de arquitectura que garantizan la validez de las transformaciones aplicadas. Se verificaron dominios de definición para funciones recíprocas, de modo que las variables utilizadas fueran estrictamente positivas en el conjunto de trabajo; se conservaron los valores faltantes sin imputaciones automáticas, evitando modificar la distribución original; y no se aplicaron reescalamientos que alteraran la lectura de los signos en la estimación. Con ello, la interpretación económica permanece alineada con el sentido de cada variable: aumentos en edad invertida reflejan mayor "novedad" relativa del activo, y aumentos en los recíprocos expresan mayor proximidad o intensidad del atributo medido.

La no discretización no impide contrastes por tramos cuando se consideren útiles para la exposición de resultados. En tales casos, las particiones se emplearán únicamente con fines descriptivos o de sensibilidad, dejando claro que no forman parte del insumo base de estimación. De este modo, la metodología conserva coherencia entre preparación de datos y objetivos de modelación, evita pérdidas de información por binning y mantiene abiertas alternativas de análisis estratificado cuando aporten claridad expositiva sin comprometer la consistencia inferencial.

5.4.4. Transformación de atributos en clases según criterios de análisis espacial o socioeconómico.

La transformación en clases se aplicó exclusivamente a la variable asociada a parques, con un propósito complementario de análisis y verificación de coherencia en la influencia sobre el valor, sin incorporarse al insumo base de modelación. No se realizaron reclasificaciones socioeconómicas ni de otros equipamientos.

La clasificación de parques se basó en la tipología original por rangos de área provista en la fuente (categorías del sistema de parques: p. ej., metropolitano, zonal, vecinal o de bolsillo, definidas por umbrales de metraje). Se respetó esa taxonomía sin modificar límites, de modo que cada observación heredó la clase tipológica correspondiente. Esta decisión permite contrastar si la relación observada entre la variable de valor y los parques mantiene un ordenamiento coherente con la jerarquía por tamaño: se espera que la influencia potencial de un parque metropolitano, por escala y oferta, difiera de la asociada a un parque de bolsillo.

Esta transformación no se utiliza en las especificaciones econométricas ni espaciales estimadas. Su función es estrictamente diagnóstica y expositiva: confirmar que las tendencias empíricas guardan sintonía con la jerarquía del sistema de parques definida por área y aportar una lectura territorial clara sin alterar la base continúa preparada en secciones previas. Con ello se mantiene la coherencia metodológica: el modelo opera sobre variables continuas (o continuo-transformadas), y las clases tipológicas de parques sirven únicamente para contrastes de coherencia y comunicación de resultados.

5.5. Fase V: Normalización de datos

La normalización se utilizó para llevar a una escala comparable atributos medidos en unidades heterogéneas y para facilitar diagnósticos gráficos y tabulares sin alterar su contenido informativo. Se aplicaron transformaciones acotadas (p. ej., estandarización z-score y normalización min–max) exclusivamente en etapas de preprocesamiento y presentación de resultados, preservando las unidades originales en los conjuntos destinados a estimación e inferencia. Los parámetros empleados en cada caso se consignaron para asegurar trazabilidad y lectura consistente entre variables y capas.

5.5.1. Estandarización de variables (z-score, min–max).

La estandarización se empleó como recurso de puesta en escala cuando fue necesario contrastar atributos medidos en unidades heterogéneas dentro de un mismo soporte gráfico o tabular, sin alterar su contenido informativo. Su uso se restringió a productos de lectura y diagnóstico; las bases destinadas a estimación conservaron las unidades originales para mantener la interpretación económica directa. Se trabajó con dos esquemas acotados y documentados que, en todo caso, se limitaron al diagnóstico exploratorio: z-score y min–max.

En el esquema z-score, cada observación x de la variable seleccionada se transforma como:

$$z = \frac{x - \bar{x}}{s},$$

Donde $\bar{x}$ es la media y s calculadas sobre el conjunto de referencia definido para esta fase. Esta operación centra la distribución en cero y expresa los valores en desviaciones estándar, lo que permite comparar magnitudes en una misma figura sin que la diferencia de unidades condicione la lectura.

La estandarización permitió comparar la magnitud relativa de la dispersión entre variables con unidades distintas, revelando diferencias en su grado de heterogeneidad. Las variables de distancia a equipamientos urbanos, como centros comerciales y estaciones de transporte masivo, presentaron una amplitud mayor en la escala estandarizada, reflejando una distribución espacial más diversa de los predios frente a dichos servicios.

El contraste entre amplitudes y concentraciones permitió identificar qué atributos poseen mayor capacidad de diferenciación dentro del conjunto, es decir, aquellos que discriminan con mayor fuerza entre los inmuebles por su variabilidad real. Las distancias, por su rango amplio y dispersión, evidencian diferencias sustantivas en accesibilidad y localización que influyen en la conformación del valor; mientras que variables con baja dispersión, como algunas dotaciones, reflejan comportamientos más estables o consolidados en el mercado. Este análisis comparativo aporta un criterio preliminar sobre el peso potencial

de cada atributo en la explicación del valor por metro cuadrado, al señalar qué variables introducen contraste entre observaciones y cuáles tienden a uniformar la muestra, sin alterar la interpretación económica original de los datos.

La normalización min–max proyecta la variable al intervalo [0,1] mediante

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}},$$

Donde x_{max} y x_{min} corresponden al mínimo y máximo observados en el mismo conjunto de referencia. Este procedimiento conserva la forma de la distribución y las relaciones proporcionales entre observaciones, pero reescala los valores para que el menor corresponda a 0 y el mayor a 1.

La transformación min–max permitió representar de manera conjunta variables con unidades y rangos heterogéneos —como distancias en metros y variables binarias de dotación— dentro de una escala normalizada común. Al ubicarlas en un mismo intervalo, fue posible contrastar su comportamiento relativo y reconocer patrones de concentración, dispersión y posibles saturaciones en los extremos de la escala.

En particular, las distancias a equipamientos urbanos tendieron a ocupar rangos amplios —entre 0.1 y 0.9—, lo que sugiere gradientes espaciales progresivos de accesibilidad entre los predios. Dichos gradientes reflejan algunas variaciones continuas en la proximidad a los servicios urbanos, sin concentraciones dominantes ni discontinuidades marcadas en el rango analizado, lo que indica una distribución territorial coherente y diferenciada. Esta dispersión amplia respalda la utilidad de las distancias como variable explicativa, al evidenciar su capacidad para discriminar niveles de accesibilidad y aportar variabilidad significativa al conjunto.

5.5.2. Escalado y transformación para comparabilidad entre unidades y magnitudes.

Se garantizó la comparabilidad básica entre atributos medidos en unidades distintas, sin alterar su significado ni introducir escalas artificiales. El principio rector fue conservar las unidades originales y aplicar únicamente ajustes de presentación o reorientaciones ya definidas en fases previas, evitando reescalados que compliquen la lectura sustantiva de los resultados.

En primer lugar, se mantuvieron las escalas naturales de cada variable: COP/m^2 para precios unitarios, m² para áreas, años para edad y conteos para indicadores de entorno. Esta decisión preserva la interpretación directa de los resúmenes descriptivos y evita ambigüedades al comparar cifras entre tablas y figuras. Cuando fue pertinente, se aplicó un formato de salida uniforme (separadores de miles y símbolo monetario) estrictamente a nivel de etiquetado; los valores subyacentes permanecieron sin modificación para no afectar cálculos, ordenamientos o pruebas.

En segundo lugar, se evaluó la necesidad de reescalado numérico auxiliar (división por 10^3 o 10^6) para variables de gran magnitud. Este recurso se dejó previsto como medida de estabilidad computacional en escenarios con rangos extremos o efectos de redondeo que dificulten la visualización.

En tercer lugar, se respetaron las transformaciones de orientación definidas previamente por su utilidad interpretativa: edad invertida (valores altos = inmuebles relativamente más recientes) y funciones recíprocas para variables tratadas como "proximidad/intensidad". Estas transformaciones ya forman parte del preprocesamiento y no se estandarizaron ni reescalaron en esta fase, justamente para mantener un vínculo unívoco entre el signo del cambio en la variable y su lectura económica.

Como controles de calidad asociados al escalado, se aplicaron tres verificaciones simples y suficientes: consistencia de unidades (homogeneidad de símbolos y separadores entre tablas y gráficos), coherencia de rangos (ausencia de truncamientos involuntarios en

ejes y leyendas) y estabilidad de precisión (número de decimales y redondeo uniforme por variable).

5.5.3. Evaluación de distribuciones con pruebas estadísticas y gráficas.

Se busca caracterizar con rigor la forma de la distribución de la variable monetaria valor_aprivada (COP/m^2) y, a partir de esa evidencia, decidir si procede aplicar transformaciones estabilizadoras. Se combinaron pruebas formales y diagnóstico visual, de manera que la decisión no dependa de un único indicador ni de impresiones aisladas.

En el componente estadístico se aplicaron los contrastes Shapiro–Wilk, D'Agostino–Pearson (basado en sesgo y curtosis) y Anderson–Darling; complementariamente, cuando la situación muestral lo requería, se consideró Kolmogorov–Smirnov/Lilliefors para verificar la adecuación global cuando los parámetros se estiman a partir de la muestra. Cada prueba se reportó con su estadístico, p-valor y el tamaño muestral efectivo (tras exclusión de faltantes).

Para empezar, se consideró la transformación Box-Cox que constituye una alternativa pertinente cuando se trabaja con variables monetarias estrictamente positivas que exhiben asimetría a la derecha y colas pesadas, como suele ocurrir con precios por metro cuadrado. A diferencia de un cambio de escala fijo, Box–Cox introduce un parámetro continuo λ que permite ajustar el grado de compresión de la cola en función de la evidencia empírica, de modo que el propio conjunto de datos determine la intensidad de la transformación. Este rasgo paramétrico ofrece una ventaja metodológica clara: Engloba varios casos prácticos en una misma familia —cuando $\lambda = 0$ recupera el logaritmo; para $\lambda = \frac{1}{2}$ se aproxima a una raíz cuadrada; para $\lambda < 0$ intensifica la compresión de valores altos— evitando comparaciones ad hoc entre transformaciones inconexas. En esta base, la estimación arrojó $\lambda \approx -0.48\ con\ N = 1{,}203, sesgo \approx 0.001\ y\ curtosis \approx 3.48$; las pruebas reportaron Shapiro–Wilk $p \approx 3.42\ x\ 10^{-3}$, Anderson–Darling $p \approx 6.17\ x\ 10^{-4}$ y D'Agostino–Pearson $p \approx 1.25\ x\ 10^{-1}$.

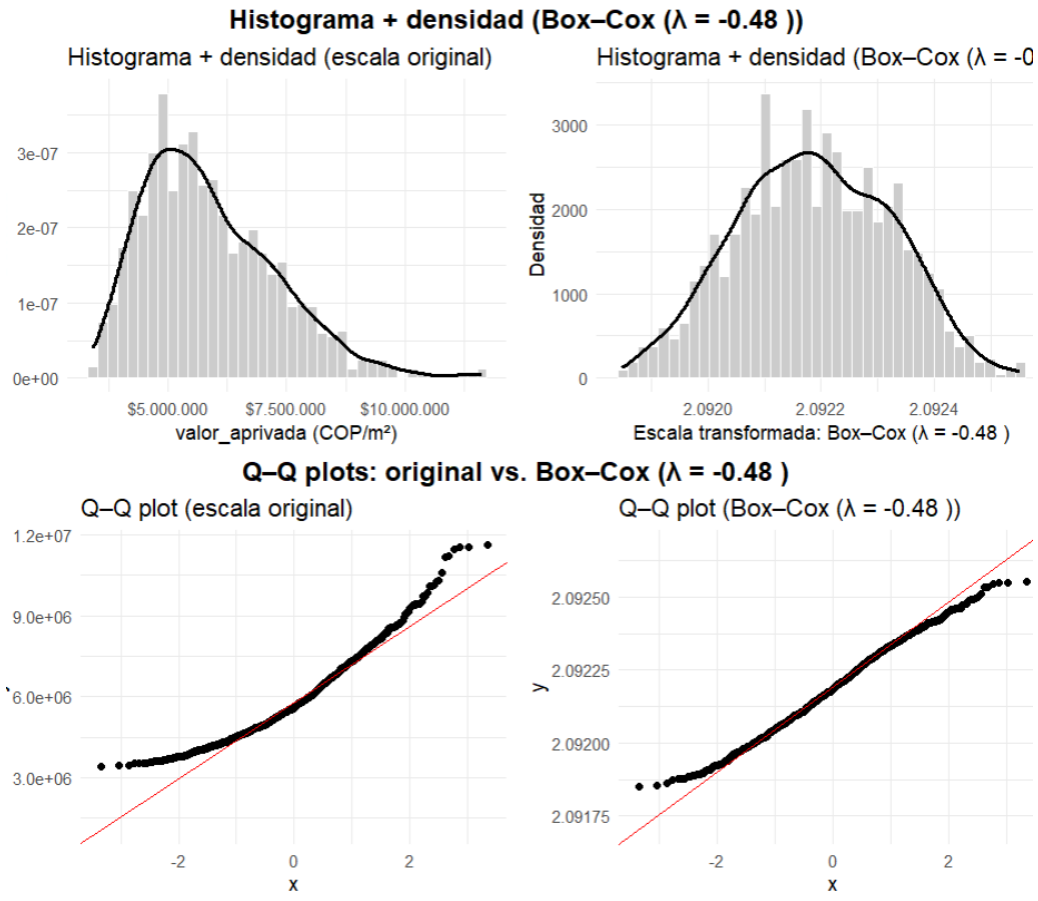


Ilustración 7 Transformación Box-Cox

Por otro lado, la transformación Yeo–Johnson es una extensión directa de Box–Cox que mantiene la monotonicidad y la invertibilidad, pero relaja la restricción de positividad: funciona incluso si la variable puede contener ceros o valores negativos. En series de precios como valor_aprivada (COP/m^2) el dominio suele ser positivo; sin embargo, en ejercicios de réplica, integración de nuevas fuentes o depuraciones intermedias (p. ej., diferencias, residuales o normalizaciones auxiliares), es plausible que aparezcan ceros o ajustes que rompan esa positividad. En ese escenario, Yeo–Johnson evita recurrir a desplazamientos ad hoc (sumar una constante arbitraria) y permite estimar un parámetro de potencia λ por perfil de verosimilitud con el mismo criterio que Box–Cox. En la práctica observada aquí, el patrón de asimetría derecha y cola superior sugiere que una transformación que comprima proporcionalmente los valores altos es pertinente; Yeo–Johnson lo consigue con un comportamiento muy próximo al de Box–Cox cuando los datos son positivos, y coincide con él en términos de ajuste si el λ estimado cae en una vecindad similar. Con los mismos datos

se obtuvo sesgo ≈ 0.001 y curtosis ≈ 3.48; las pruebas reportaron Shapiro–Wilk $p \approx 3.42\,x\,10^{-3}$, Anderson–Darling $p \approx 6.17\,x\,10^{-4}$ y D'Agostino–Pearson $p \approx 1.25\,x\,10^{-1}$.

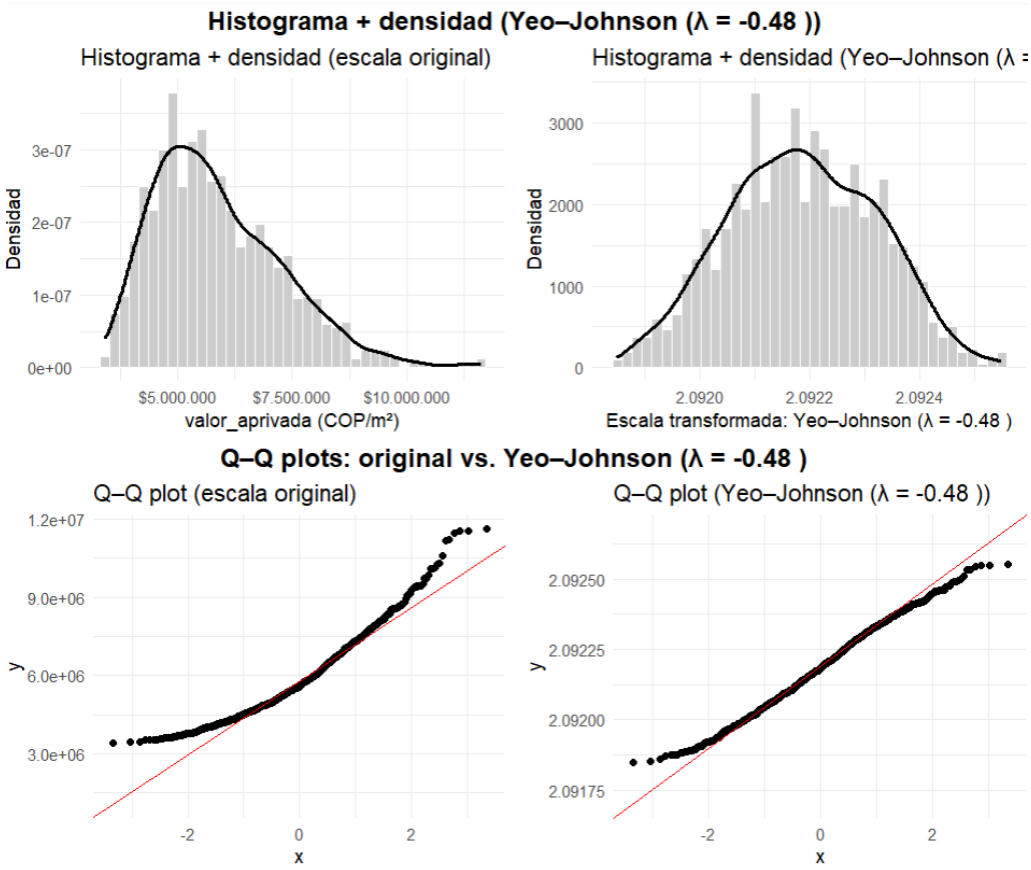


Ilustración 8 Transformación Yeo-Johnson

Por otro lado, la transformación raíz cuadrada $y' = \sqrt{y}$ se utiliza con variables positivas cuando se busca una compresión moderada de la cola derecha sin alterar de forma drástica la escala. En precios por metro cuadrado, donde la dispersión suele crecer con el nivel del precio, la raíz reduce la amplitud relativa de los valores altos y puede atemperar heterocedasticidad ligera, manteniendo una relación intuitiva con la unidad original (las diferencias siguen siendo crecientes, pero menos abruptas en la parte superior de la distribución). En el conjunto analizado, su aplicación deja una mejora parcial del perfil: con $N = 1{,}203$ observaciones, el sesgo pasa a 0.403 y la curtosis a 3.49, indicadores que aún reflejan cierta asimetría. Las pruebas formales siguen rechazando normalidad con holgura—Shapiro–Wilk $p \approx 3.16\,x\,10^{-8}$, Anderson–Darling $p \approx 1.03\,x\,10^{-10}$ y D'Agostino–

Pearson $p \approx 4.47\ x\ 10^{-5}$lo que sugiere que, si el objetivo es aproximar una forma cuasi normal o aliviar de manera más marcada las colas, podrían requerirse transformaciones más potentes.

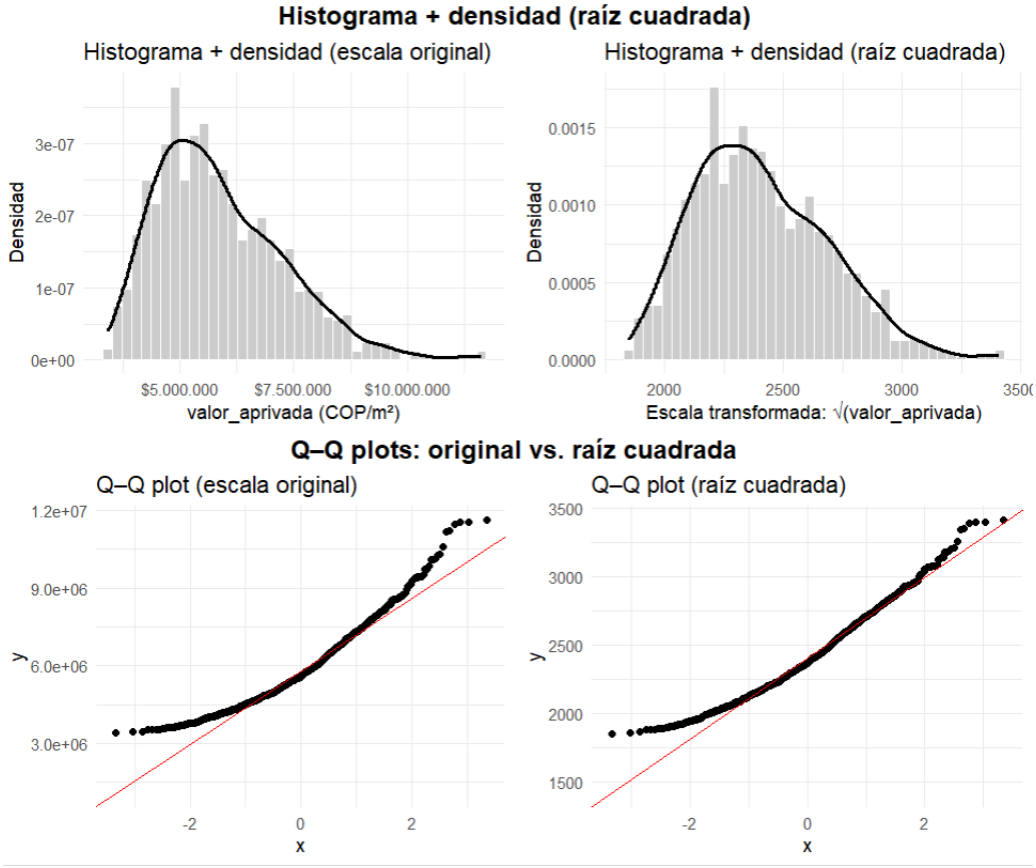


Ilustración 9 Transformación Raíz Cuadrada

Por ultimo La transformación asinh $y' = \text{asinh}(y) = ln\ (y + \sqrt{y^2 + 1}$ se emplea cuando se busca comprimir colas derechas de manera similar al logaritmo, pero sin exigir positividad estricta ni desplazamientos ad hoc. Es monótona e invertible en $\mathbb{R}$, por lo que resulta especialmente práctica en flujos de trabajo donde podrían aparecer ceros, pequeños negativos o diferencias que atraviesen el cero durante procesos de depuración o integración de fuentes. En series de precios como valor_aprivada (COP/m^2) —positiva en la base principal— asinh ofrece un comportamiento muy cercano al log para valores medianos y grandes, a la vez que mantiene un tratamiento suave cerca de cero, evitando problemas numéricos o decisiones de sumar constantes arbitrarias.

En los resultados obtenidos, con $N = 1{,}203$ observaciones, asinh entrega métricas prácticamente equivalentes a la escala logarítmica: Sesgo pasa a 0,192 y curtosis a 3.36; los contrastes formales siguen rechazando normalidad, aunque con mejora clara frente a la escala original: Shapiro–Wilk $p \approx 6.93\,x\,10^{-4}$, Anderson–Darling $p \approx 1.37\,x\,10^{-5}$, y D'Agostino–Pearson $p \approx 1.54\,x\,10^{-2}$. Visualmente, el Q–Q plot reduce la curvatura central y modera la cola superior, de forma alineada con lo que se espera cuando se atenúan efectos multiplicativos.

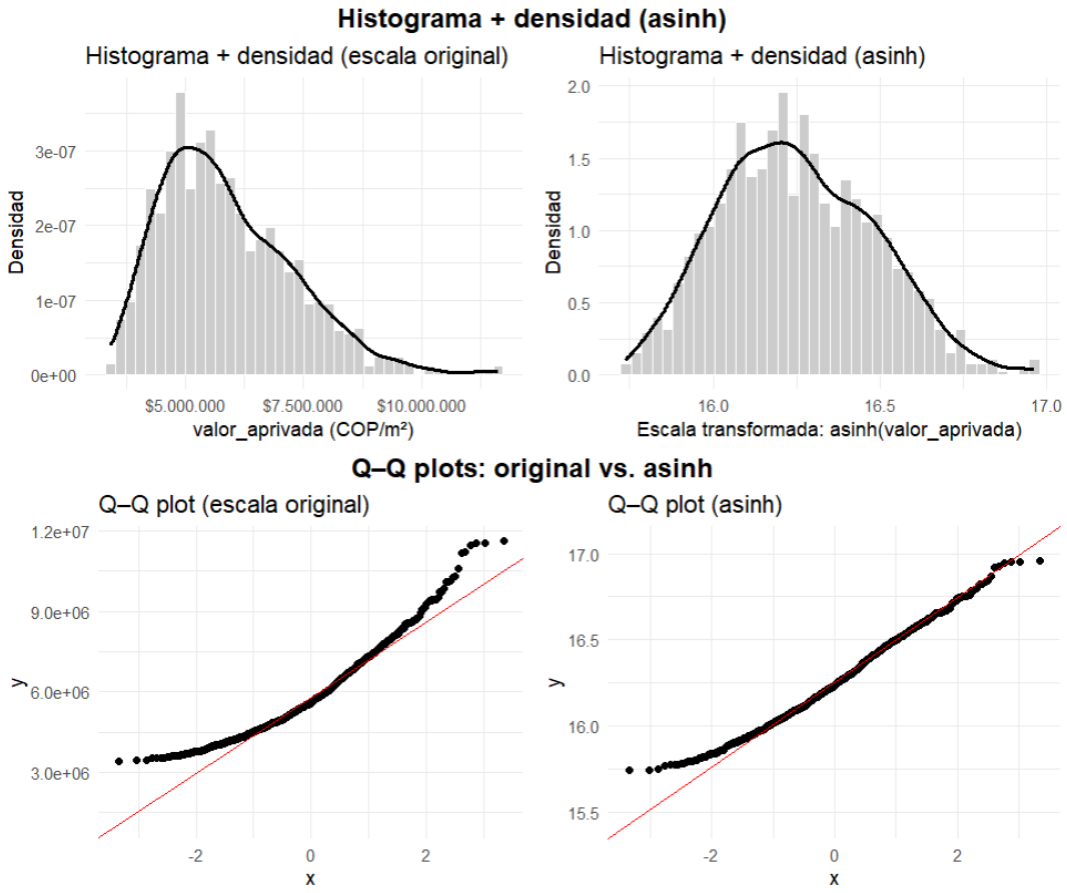


Ilustración 10 Transformación asinh

En los resultados obtenidos, con $N = 1{,}203$ observaciones, la transformación logarítmica mejora de forma apreciable la forma de la distribución frente a la escala original: el sesgo se reduce a 0,269 y la curtosis a 2,62. Los contrastes formales siguen rechazando normalidad — Shapiro–Wilk $p \approx 6.93\,x\,10^{-4}$, Anderson–Darling $p \approx 1.37\,x\,10^{-5}$, y D'Agostino–Pearson $p \approx 1.54\,x\,10^{-2}$. pero la evidencia conjunta señala un alivio nítido de la cola derecha y un perfil central más próximo a la simetría. Visualmente, el Q–Q plot atenúa

la curvatura en cuantiles medios y modera la separación en el tramo superior, en línea con el efecto de compresión multiplicativa que caracteriza al logaritmo.

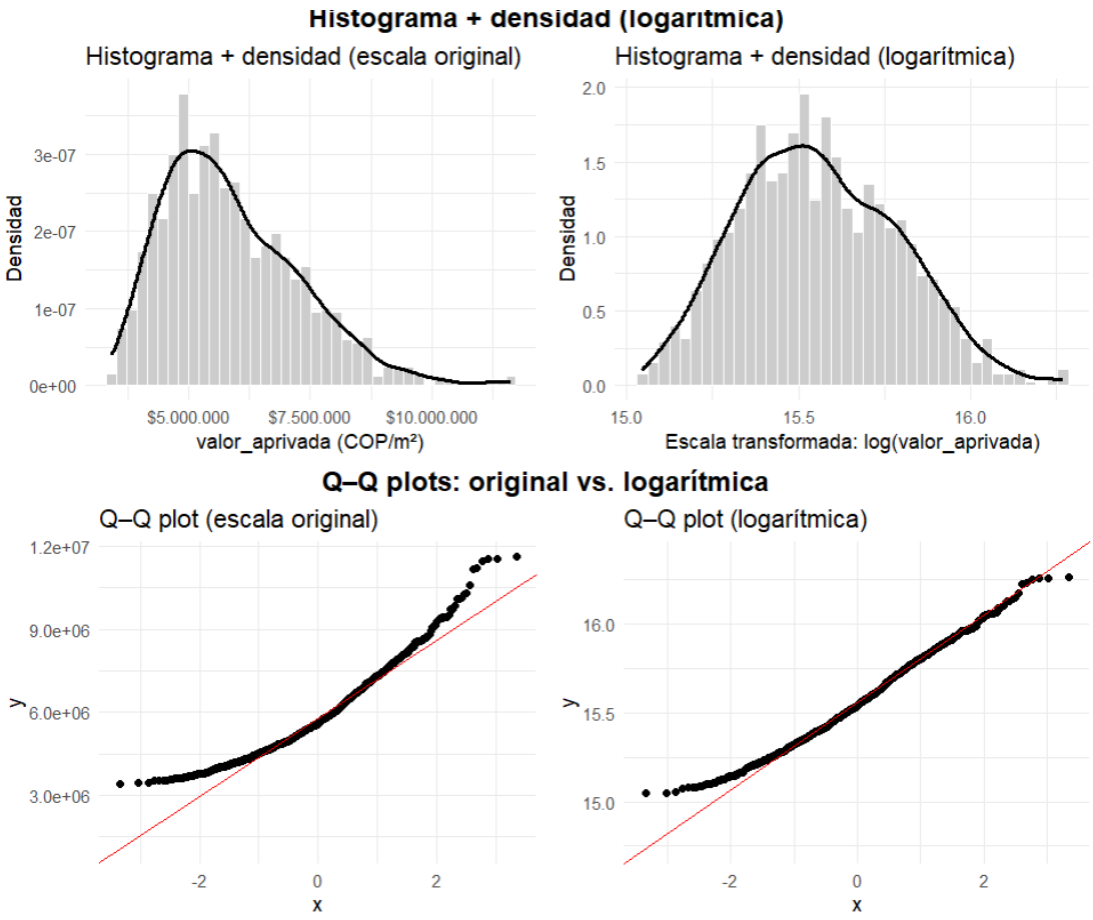


Ilustración 11 Transformación Logarítmica

5.5.4. Aplicación de transformaciones (logarítmica, Box-Cox, Yeo-Johnson) para estabilizar varianzas y aproximar normalidad.

La elección de la escala de trabajo se define con base en una comparación explícita entre transformaciones monotónicas que preservan el orden de las observaciones y permiten retrotransformar resultados a COP/m^2 cuando es necesario. Para el valor por metro cuadrado de área privada, se evaluaron Box–Cox y Yeo–Johnson, incorporando raíz y asinh como referencias de suavizado. La lectura combina métricas de forma (sesgo y curtosis) y contrastes de normalidad (Shapiro–Wilk, Anderson–Darling y D'Agostino–Pearson)

calculados sobre la variable transformada. Adicionalmente, se reporta un score sintético (menor es mejor) que resume, sin sustituirlos, los indicadores anteriores.

Transformación	**Shapiro-Wilk**	**Anderson-Darling**	**D'Agostino-Pearson**
Box–Cox ($\lambda = -0.48$)	$3.42\ x\ 10^{-3}$	$6.17\ x\ 10^{-4}$	$1.25\ x\ 10^{-1}$
Yeo–Johnson ($\lambda = -0.48$)	$3.42\ x\ 10^{-3}$	$6.17\ x\ 10^{-4}$	$1.25\ x\ 10^{-1}$
Raíz	$3.16\ x\ 10^{-8}$	$1.03\ x\ 10^{-10}$	$4.47\ x\ 10^{-5}$
Asinh	$6.93\ x\ 10^{-4}$	$1.37\ x\ 10^{-5}$	$1.54\ x\ 10^{-2}$
Log	$6.93\ x\ 10^{-4}$	$1.37\ x\ 10^{-5}$	$1.54\ x\ 10^{-2}$
Original (COP/m^2)	$2.66\ x\ 10^{-13}$	$1.45\ x\ 10^{-18}$	$9.40\ x\ 10^{-10}$

Tabla 3 Comparación de transformaciones sobre los residuos

La interpretación comparada de las estadísticas permite dimensionar con precisión el efecto de cada transformación sobre la **distribución empírica de los residuos**, y, por extensión, sobre la estabilidad esperable en modelos posteriores. En la escala original de los residuos (sin transformación), el sesgo positivo (0.633) y la curtosis elevada (3.95), acompañados de valores extremadamente bajos en las pruebas de normalidad de Shapiro–Wilk, Anderson–Darling y D'Agostino–Pearson, confirman una cola derecha pesada y un alejamiento sustantivo de la normalidad. Este comportamiento evidencia el problema de fondo: la dispersión de los residuos crece con el nivel de los valores atípicos y los cuantiles altos se apartan de la diagonal teórica, lo que es consistente con la presencia de segmentos en el mercado que concentran observaciones moderadas junto a extremos de alta distorsión.

Sobre esta base, la transformación logarítmica mantiene su utilidad para interpretar resultados en términos multiplicativos, coherente con la lógica de precios y apropiada para reducir la amplitud relativa de los valores altos de los residuos. Sin embargo, los indicadores muestran que el ajuste en la normalidad es solo parcial: tras aplicar la transformación sobre los residuos, el sesgo se reduce a 0.192 y la curtosis a 3.36, cifras más próximas al estándar teórico que en la escala sin transformar, aunque persiste cierta desviación respecto a la simetría ideal. Los p-valores de las pruebas de Shapiro–Wilk y Anderson–Darling (ambos en el orden de 10^{-4} y 10^{-5}respectivamente) evidencian una mejora relevante comparada con

los residuos originales, aunque aún señalan la presencia de una cola derecha pesada que limita la aceptación de la normalidad completa.

La transformación asinh (arcoseno hiperbólico) comparte esta filosofía, aproximándose al logaritmo para valores medianos y grandes y permitiendo dominio total, sin requerir estricta positividad. En esta muestra, la asinh reproduce casi exactamente los mismos resultados sobre los residuos (sesgo 0.192; curtosis 3.36; mismos órdenes de p), lo que la posiciona como una alternativa robusta en contextos donde se busca preservar la parsimonia y la interpretabilidad multiplicativa, pero sin dejar de mejorar las métricas de ajuste para los residuos frente al modelo no transformado.

La raíz cuadrada propone un cambio de escala más suave, atenuando gradualmente las diferencias en la zona alta, lo que puede ser conveniente cuando se busca intervenir lo mínimo en la distribución de los residuos. En los resultados observados tras aplicar la transformación sobre los residuos, el sesgo se reduce a 0.403 y la curtosis a 3.49, valores intermedios respecto al estándar normal teórico. No obstante, los contrastes de normalidad siguen arrojando resultados de rechazo (p-valores de Shapiro–Wilk y Anderson–Darling en el orden de 10^{-8} y 10^{-4}), lo que evidencia que la raíz cuadrada actúa como un suavizado moderado: mejora parcialmente la simetría y concentra la distribución, pero no alcanza a corregir completamente la presencia de colas, limitando la aceptación total de la normalidad en los residuos. Así, la interpretación económica en términos de niveles permanece intacta, aunque persisten aspectos de dispersión extremos propios del fenómeno estudiado.

Las familias Box–Cox y Yeo–Johnson abordan el problema del ajuste de los residuos desde una perspectiva paramétrica que permite modular la intensidad de la compresión mediante la estimación de un exponente óptimo para el propio conjunto de datos, preservando propiedades clave como la monotonicidad y la invertibilidad. Al aplicar la transformación Box–Cox con el parámetro λ óptimo, se observa la mayor corrección: el sesgo se reduce prácticamente a cero (0.001), la curtosis se mantiene en un valor muy cercano al de la normalidad teórica (3.48) y los p-valores de las pruebas de normalidad aumentan de manera significativa, reflejando una distribución de residuos mucho más simétrica y con colas menos pronunciadas.

Por su parte, la transformación Yeo–Johnson —que generaliza la Box–Cox para incluir dominios con ceros o valores negativos— reproduce prácticamente el mismo patrón de ajuste: el sesgo y la curtosis, mientras que los p-valores son igualmente altos y de la misma magnitud. Metodológicamente, ambas transformaciones logran alinear los cuantiles y moderar colas sin requerir cortes arbitrarios o restricciones estrictas de positividad, y ofrecen la ventaja adicional de permitir la retrotransformación directa a unidades naturales de interpretación como el COP/m^2. Esto las convierte en alternativas óptimas, tanto por su robustez estadística como por su utilidad práctica para la presentación y comparación de resultados en aplicaciones de valoración inmobiliaria.

5.6. Fase VI: Transformación y Modelamiento espacial

Esta sección consolida la estimación de COP/m² integrando, en una misma lógica coherente, la señal econométrica extraída de modelos no paramétricos y semiparamétricos con la estructura espacial subyacente del fenómeno de valoración inmobiliaria. La especificación de atributos se mantiene unificada —tamaño efectivo, novedad, dotación, altura, año y gradientes de accesibilidad— y se trabaja con rigor en la coherencia de escala para preservar la información geográfica contenida en los residuos del modelo econométrico base.

Los modelos econométricos se contrastan bajo un marco común de desempeño fuera de muestra, permitiendo evaluar cuál especificación captura más eficientemente la variación en precios unitarios. Sin embargo, la evaluación de la estructura residual mediante el estadístico de Moran I revela la presencia sistemática de dependencia espacial en los residuos de estos modelos. Lejos de constituir una limitación o fracaso del enfoque econométrico, esta estructura de asociación espacial en residuos evidencia que los procesos de valoración inmobiliaria poseen un componente inherentemente espacial: inmuebles ubicados proximalmente tienden a exhibir características omitidas correlacionadas (composición demográfica, servicios comunitarios, factores ambientales) que generan valores residuales correlacionados geográficamente.

Por esta razón, cuando la dependencia espacial es estadísticamente significativa, se incorpora explícitamente la componente espacial mediante kriging ordinario, destinada a capturar la variación estructurada espacialmente que persiste no explicada por los predictores econométricos. El modelamiento del fenómeno espacial residual opera bajo la premisa de que la continuidad geográfica en los precios no es simplemente aleatoria, sino que refleja procesos de valoración que operan a través de la proximidad: la autocorrelación espacial positiva documentada implica que conocer el valor residual de un inmueble informar sobre el valor esperado de inmuebles cercanos. Esta estructura residual, cuando es modelada mediante una superficie continúa interpolada, permite obtener predicciones desagregadas espacialmente y mapas temáticos internamente consistentes.

La estrategia metodológica aprovecha la separación analítica entre componentes: el modelo econométrico captura la relación determinística entre atributos de la propiedad y su valor (forma funcional de la valoración), mientras que la superficie continua de residuos captura la variación sistemática que depende de la localización geográfica (estructura espacial de la valoración). El resultado es un conjunto de predicciones y superficies temáticas comparables entre sí bajo un mismo criterio de referencia, combinando forma funcional derivada de interacciones no lineales e interacciones de atributos con la realidad espacialmente continua del mercado inmobiliario. Esta arquitectura metodológica evita recargar la exposición con complejidad innecesaria al tiempo que reconoce explícitamente que la dependencia espacial residual no invalida los análisis anteriores, sino que constituye un componente adicional esencial para predicciones locales robustas y cartografía de valoración coherente con la heterogeneidad geográfica observada.

5.6.1. Aplicación de modelos.

Como punto de partida se estimó una regresión lineal ordinaria sobre la respuesta transformada por Box–Cox ($\lambda \approx -0{,}478$), con el objetivo de capturar de manera parsimoniosa la relación entre el valor unitario y los atributos físicos, de equipamiento y de entorno. El ajuste explica alrededor del 69.62% de la variabilidad (R^2 ajustado ≈ 0.6962) y presenta un error residual en la escala transformada del orden de $7.498\ x\ 10^{-5}$. El modelo

global resulta altamente significativo ($F(16.1185) = 173, p < 0{,}001$), lo que confirma la relevancia conjunta de los predictores. No obstante, debido al tamaño de la muestra (es esperable que algunas pruebas formales de ajuste o normalidad arrojen valores p muy pequeños, sin que ello implique desviaciones sustanciales en el comportamiento de los residuos.

Los coeficientes muestran señales coherentes con la teoría del mercado inmobiliario y son, en su mayoría, estadísticamente significativos: la forma recíproca del área privada (inv_area), la antigüedad del inmueble (edad_invertida), la dotación comunal, el piso, el estrato (niveles 4, 5 y 6), la presencia de ascensor, la condición comercial (nuevo_usado), el estado de terminación y el año se asocian positivamente con el valor por metro cuadrado en la escala Box–Cox. Por el contrario, la mayor cercanía a colegios (inv_colegios) conserva una relación negativa y significativa con el valor unitario, lo que sugiere posibles efectos de congestión o externalidades no deseadas en su entorno inmediato. En contraste, los indicadores de proximidad restantes —inv_parques, inv_vias, inv_centros e inv_TM— no presentan significancia estadística, lo que indica efectos marginales o colinealidades entre variables de entorno que podrían examinarse con mayor detalle en especificaciones posteriores.

Se hará la aplicación en los siguientes modelos, Regresión Lineal ordinaria (OLS por sus siglas en ingles), Regresión Lineal Múltiple (RLM), Modelo Aditivo Generalizado (GAM), Regresión Cuantílica (RQ) y Random Forest (RF).

```
Call:
lm(formula = form_base, data = df2)

Residuals:
       Min         1Q     Median         3Q        Max
-2.553e-04 -4.604e-05 -2.451e-06  4.727e-05  2.792e-04

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.067e+00  7.524e-03 274.769  < 2e-16 ***
inv_area        3.016e-03  4.410e-04   6.839 1.28e-11 ***
edad_invertida  5.635e-06  3.294e-07  17.104  < 2e-16 ***
dotacion        1.338e-04  2.121e-05   6.310 3.94e-10 ***
Piso            2.902e-06  5.251e-07   5.527 4.01e-08 ***
nuevo_usado     1.036e-04  1.950e-05   5.311 1.30e-07 ***
estrato4        2.373e-05  6.433e-06   3.689 0.000235 ***
estrato5        8.404e-05  9.226e-06   9.109  < 2e-16 ***
estrato6        1.549e-04  1.222e-05  12.668  < 2e-16 ***
ascensor        2.727e-05  6.265e-06   4.353 1.46e-05 ***
terminado       9.738e-05  2.469e-05   3.945 8.46e-05 ***
Año             1.206e-05  3.717e-06   3.245 0.001206 **
inv_parques    -1.723e-04  1.222e-04  -1.410 0.158755
inv_vias        4.512e-05  1.878e-04   0.240 0.810186
inv_colegios   -8.398e-03  2.033e-03  -4.130 3.89e-05 ***
inv_centros     1.516e-03  1.863e-03   0.814 0.415935
inv_TM         -6.736e-04  1.540e-03  -0.437 0.661882
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.498e-05 on 1185 degrees of freedom
Multiple R-squared:  0.7002,    Adjusted R-squared:  0.6962
F-statistic:   173 on 16 and 1185 DF,  p-value: < 2.2e-16
```

Ilustración 12 Resumen del modelo regresión lineal ordinaria

La prueba de Shapiro–Wilk aplicada a los residuos en la escala transformada arroja $W = 0{,}9960$ y $p = 0{,}003417$, lo que sugiere una leve desviación respecto a la normalidad compatible con colas algo más pesadas. Los diagnósticos gráficos exhiben simetría razonable y ausencia de patrones sistemáticos, en línea con esa lectura.

Considerando que el rechazo de normalidad es recurrente en el conjunto de modelos especificados, la aplicación de pruebas de homocedasticidad requiere de métodos estadísticos robustos que no dependan de este supuesto. Por esta razón, se utilizó la prueba de homocedasticidad de Koenker-Bassett, una alternativa robusta a la prueba de Breusch-Pagan que no exige normalidad de los errores. Este test es especialmente adecuado en contextos donde, los modelos empleados, están explícitamente diseñados para funcionar bajo desviaciones de los supuestos clásicos. La prueba de Koenker-Bassett opera mediante un estadístico de contraste que es resistente a la no normalidad, permitiendo identificar si la varianza de los errores es constante (homocedasticidad) o varía sistemáticamente con los valores predichos o variables explicativas (heterocedasticidad), sin que la validez de las inferencias se vea comprometida por desviaciones de la normalidad.

Los resultados de la prueba de Koenker-Bassett en RLM Huber: $LM = 111 - 807, p = 2.032 \times 10^{-16}$; Criterio: normalidad rechazada (p<0.05).

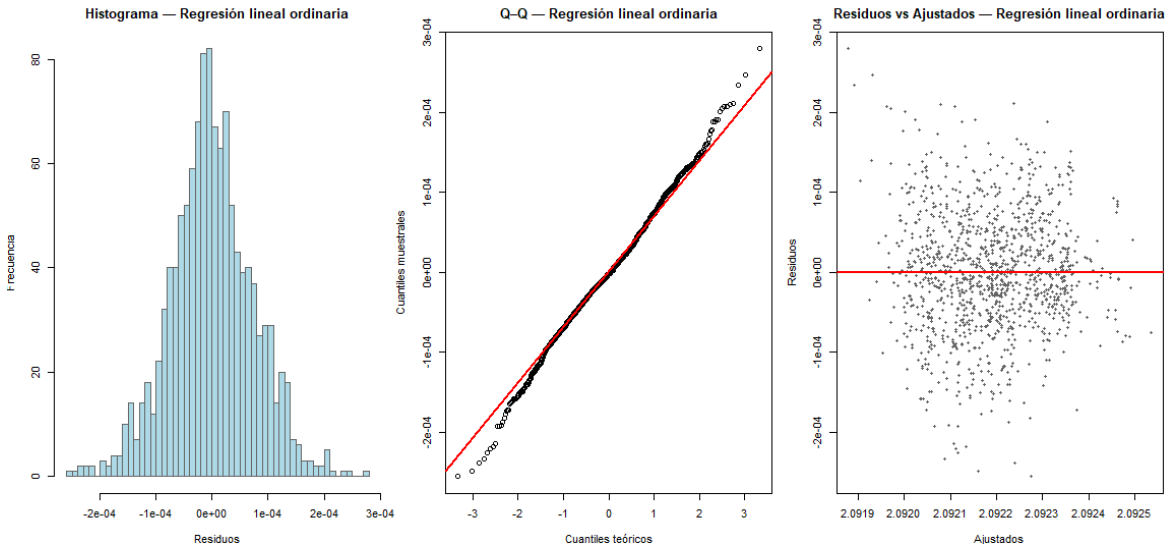


Ilustración 13 Gráficos regresión lineal ordinaria

En términos espaciales, la evaluación de autocorrelación mediante el índice de Moran constituye un paso metodológico crítico para identificar dependencia espacial no capturada por el modelo de especificación lineal ordinaria. La búsqueda sistemática del parámetro k (número de vecinos más próximos) en la construcción de la matriz de pesos espaciales responde a una necesidad teórica fundamental: garantizar que la estructura de vecindad reflejada en la matriz W represente adecuadamente la interacción espacial del fenómeno estudiado. La selección óptima de k determina la conectividad del gráfico espacial, es decir, cómo se vinculan las unidades observacionales en la malla geográfica. Un valor de k demasiado bajo puede generar múltiples componentes conexas desconectadas en el grafo, fragmentando artificialmente el espacio de análisis; por el contrario, un valor de k excesivamente alto puede introducir ruido al incluir vecinos lejanos que no representan genuinamente la proximidad espacial relevante. En este análisis, la identificación de k, (i) mantiene conectividad suficiente para no fragmentar la información espacial, (ii) evita sobreparametrización que comprometería la eficiencia del estimador, y (iii) permite identificar patrones de dependencia espacial genuina sin capturar artefactos de especificación arbitraria.

Al retrotransformar las predicciones a COP/m^2 para evaluar dependencia espacial de los residuos en nivel, se obtiene Moran's $I \approx 0{,}214$ con estadístico $z \approx 19{,}473$ ($p < 2{,}2 \times 10^{-16}$), bajo la hipótesis alternativa $H_a: I > E[I]$. El análisis se realizó utilizando una

matriz de pesos W tipo $k - NNcon(k = 12)$ normalizada por filas, seleccionada a partir de un procedimiento iterativo que evaluó distintos valores de k bajo criterios de conectividad, fuerza del estadístico z y parquedad. En este caso, aunque no se alcanzó conectividad global (persistiendo cinco componentes desconectadas), se observó una estabilización en la ganancia marginal de z para valores de $k \geq 12$ lo que sugiere que este nivel de vecindad representa un equilibrio entre cohesión espacial y ruido topológico. El resultado evidencia una autocorrelación espacial positiva y altamente significativa en los residuos, indicando la presencia de dependencia espacial remanente no capturada por la estructura lineal del modelo de regresión lineal ordinaria. Esto sugiere que la variable respuesta conserva patrones espaciales no explicados por los predictores incluidos.

```
=== Búsqueda de k para Moran (k-NN, W estandarizada) [longlat] ===
    k components isolates mean_deg         I        z            p
1   3         59        0        3 0.3261036 15.07102 1.256024e-51
2   4         35        0        4 0.3006090 15.91889 2.343037e-57
3   5         22        0        5 0.2788637 16.49678 1.934852e-61
4   6         14        0        6 0.2753750 17.80219 3.398137e-71
5   7         11        0        7 0.2653510 18.49749 1.081559e-76
6   8         10        0        8 0.2553047 18.99285 9.772054e-81
7   9          9        0        9 0.2435973 19.16904 3.356794e-82
8  10          7        0       10 0.2314355 19.19656 1.977261e-82
9  11          7        0       11 0.2222904 19.33266 1.426711e-83
10 12          5        0       12 0.2143362 19.47341 9.229191e-85

--- Interpretación automática ---
En el rango evaluado no se alcanzó conectividad global (≥2 componentes).
La ganancia marginal en z empieza a ser pequeña (Δz < 0.15) alrededor de k = 12.
k* seleccionado = 12

Moran I test under randomisation
data:    res
weights: k-NN (k=12), W estandarizada [longlat]
Connectivity      : 5 connected component(s)
alternative hypothesis: greater   (Ha: I > E[I])

Moran I statistic standard deviate = 19.473, p-value < 2.2e-16
sample estimates:
Moran I statistic    0.214336236721
Expectation          -0.000832639467
Variance             0.000122088598
```

Ilustración 14 Moran regresión lineal ordinaria

La regresión de mínimos cuadrados ordinarios proporciona una línea base informativa y globalmente estable; sin embargo, la evidencia de heterocedasticidad y autocorrelación espacial de los errores indica la necesidad de mejorar la estimación mediante dos vías: (a) el uso de inferencia robusta o corrección de varianza —como errores estándar robustos o regresión lineal robusta—, y (b) la incorporación explícita de la dependencia espacial mediante modelos geoestadísticos o espaciales. El análisis comparativo incorpora enfoques flexibles, robustos y no paramétricos: los modelos aditivos generalizados, que

permiten capturar relaciones no lineales mediante funciones suavizadas; la regresión lineal robusta, que reduce la influencia de observaciones atípicas; la regresión cuantílica, que estima distintos cuantiles de la distribución condicional; y el bosque aleatorio, que representa interacciones complejas a través del ensamblaje de múltiples árboles de decisión.

Para la estimación robusta se utilizó el método M de Huber, implementado mediante la función rlm() del paquete MASS, con el propósito de atenuar la influencia de observaciones atípicas y obtener estimadores insensibles a valores extremos en la variable respuesta. Este método redefine la función de pérdida tradicional de mínimos cuadrados, sustituyendo el cuadrado del residuo por una función pieza lineal-cuadrática: cuadrática para residuos pequeños y lineal para residuos grandes. De esta forma, los residuos moderados mantienen la eficiencia del modelo lineal ordinario, mientras que los valores atípicos reciben un peso reducido, evitando su impacto desproporcionado sobre los parámetros estimados.

Dado que este tipo de modelos no reporta valores de significancia estadística de forma directa, se calcularon p-valores asintóticos a partir de los estadísticos t generados por el modelo, asumiendo una distribución normal estándar para las estimaciones robustas. El estadístico t se construye al dividir cada coeficiente estimado por su error estándar robusto, cuantificando la razón entre la magnitud del efecto y su precisión. Posteriormente, estos estadísticos t se transforman en p-valores bilaterales mediante la función de distribución normal estándar, permitiendo evaluar la probabilidad de observar valores tan extremos bajo la hipótesis nula de ausencia de efecto. De esta manera, se derivaron las pruebas de significancia individual de los coeficientes y los intervalos de confianza al 95%, los cuales se construyen situando un margen de error simétrico alrededor de cada estimación puntual, incorporando el error estándar robusto para reflejar la incertidumbre inherente a la estimación en un contexto resistente a desviaciones de normalidad.

Los resultados del modelo robusto evidencian que las variables continuas de acceso y dotación de infraestructura ejercen efectos altamente significativos sobre la variable respuesta. El inverso del área resulta una variable explicativa central, con una influencia estadísticamente significativa que refleja cómo la densidad de espacios disponibles incide en el modelo. Similarmente, la edad invertida y la dotación de facilidades muestran relaciones significativas con la variable dependiente, corroborando que características estructurales del

inmueble son determinantes en su valoración o caracterización. La variable temporal (año) también presenta un efecto significativo, indicando la presencia de tendencias temporales en el fenómeno estudiado.

En cuanto a las variables categóricas de estrato socioeconómico, se observa un patrón creciente y sistemáticamente significativo conforme se asciende. Los estratos superiores muestran coeficientes positivos y crecientes, revelando diferencias estructurales importantes entre grupos socioeconómicos, donde los más altos presentan características diferenciadoras respecto a la categoría de referencia. La significancia estadística robusta de estas variables indica que la estratificación socioeconómica constituye un factor explicativo central en el modelo, capturando heterogeneidades fundamentales en la población estudiada. Adicionalmente, variables como la presencia de ascensor y la condición de propiedad (nuevo versus usado) presentan efectos significativos, aunque de menor magnitud, sugiriendo que características secundarias del inmueble contribuyen de manera subordinada a la explicación global.

Respecto a las variables de accesibilidad inversa, la proximidad a instituciones educativas (colegios) emerge como un determinante altamente significativo, mostrando una relación clara y bien definida con la variable dependiente. En contraste, otros indicadores de accesibilidad como la distancia a parques, vías principales y centros comerciales presentan efectos individuales débiles o estadísticamente no significativos cuando se consideran de forma aislada, sugiriendo posibles problemas de colinealidad o una menor relevancia explicativa a nivel individual.

Adicionalmente, se realizaron pruebas conjuntas tipo Wald para contrastar la significancia global de bloques de variables, con el fin de evaluar si dichos conjuntos contribuyen explicativamente al modelo más allá de la significancia individual de sus componentes. Para el bloque de variables de estrato socioeconómico, la prueba Wald comparó un modelo completo que incluye todas las categorías de estrato frente a un modelo restringido que las excluye. El estadístico F arrojó un p-valor inferior a $2{,}2 \times 10^{-16}$, lo que confirma de manera concluyente que el conjunto de variables de estrato aporta información

explicativa significativa al modelo de forma conjunta y no puede ser excluido sin pérdida sustancial de poder explicativo.

De manera análoga, para el bloque de variables de accesibilidad inversa, la prueba Wald resultó altamente significativa ($p = 0.0002$), indicando que, aunque algunas variables individuales no resultaron significativas, el conjunto de indicadores de accesibilidad contribuye significativamente al poder explicativo del modelo. Este resultado evidencia que las medidas de accesibilidad poseen valor predictivo cuando se consideran en su conjunto, a pesar de que sus efectos marginales individuales sean modestos o no alcancen significancia convencional. Esta estrategia de evaluación conjunta resulta fundamental en contextos donde variables correlacionadas pueden presentar efectos individuales débiles, pero aportes colectivos relevantes, permitiendo evitar la exclusión errónea de bloques informativos.

Esta aproximación permitió obtener inferencia estadística robusta y documentar formalmente la significancia de los efectos bajo un esquema resistente a la no normalidad y a la heterocedasticidad, asegurando que las conclusiones se sostienen aun cuando existe presencia de valores extremos o alta variabilidad en el conjunto de datos.

	Value	Std. Error	t value	p(>|z|)	CI95_L	CI95_U
(Intercept)	2.061357	0.007355	280.278274	0.000000	2.046942	2.075772
inv_area	0.002800	0.000431	6.495756	0.000000	0.001955	0.003645
edad_invertida	0.000006	0.000000	17.925619	0.000000	0.000005	0.000006
dotacion	0.000136	0.000021	6.542103	0.000000	0.000095	0.000176
Piso	0.000003	0.000001	5.226998	0.000000	0.000002	0.000004
nuevo_usado	0.000102	0.000019	5.328763	0.000000	0.000064	0.000139
estrato4	0.000020	0.000006	3.114162	0.001845	0.000007	0.000032
estrato5	0.000078	0.000009	8.605646	0.000000	0.000060	0.000095
estrato6	0.000145	0.000012	12.160492	0.000000	0.000122	0.000169
ascensor	0.000029	0.000006	4.815368	0.000001	0.000017	0.000041
terminado	0.000095	0.000024	3.928392	0.000086	0.000047	0.000142
Año	0.000014	0.000004	3.943365	0.000080	0.000007	0.000021
inv_parques	-0.000232	0.000119	-1.946068	0.051647	-0.000466	0.000002
inv_vias	0.000090	0.000184	0.490148	0.624029	-0.000270	0.000450
inv_colegios	-0.008929	0.001988	-4.492099	0.000007	-0.012825	-0.005033
inv_centros	0.001708	0.001821	0.937816	0.348339	-0.001862	0.005278
inv_TM	0.000909	0.001505	0.604084	0.545788	-0.002041	0.003860

Ilustración 15 Resumen del modelo de regresión robusta (RLM – Huber)

```
--- Wald test: Bloque_estrato ---

Linear hypothesis test:
estrato4 = 0
estrato5 = 0
estrato6 = 0

Model 1: restricted model
Model 2: valor_aprivada_bc ~ inv_area + edad_invertida + dotacion + Piso +
    nuevo_usado + estrato + ascensor + terminado + Año + inv_parques +
    inv_vias + inv_colegios + inv_centros + inv_TM

  Res.Df Df      F    Pr(>F)
1   1188
2   1185  3 63.485 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

--- Wald test: Accesibilidad_inversa ---

Linear hypothesis test:
inv_parques = 0
inv_vias = 0
inv_colegios = 0
inv_centros = 0
inv_TM = 0

Model 1: restricted model
Model 2: valor_aprivada_bc ~ inv_area + edad_invertida + dotacion + Piso +
    nuevo_usado + estrato + ascensor + terminado + Año + inv_parques +
    inv_vias + inv_colegios + inv_centros + inv_TM

  Res.Df Df      F    Pr(>F)
1   1190
2   1185  5 4.7795 0.0002491 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ilustración 16 Pruebas conjuntas tipo Wald para bloques de variables

El diagnóstico de normalidad muestra un patrón globalmente regular, con ligera desviación en las colas: Shapiro–Wilk $W \approx 0{,}9952$ y $p \approx 0{,}000687$. Aunque el rechazo formal es estadísticamente significativo, los gráficos (Q–Q y densidad) evidencian un buen alineamiento en la zona central y leves separaciones en los extremos.

Los resultados de la prueba de Koenker-Bassett en RLM Huber: $LM = 29.283, p = 6.253 \times 10^{-8}$; lo que indica que la varianza de los errores residuales no es constante y presenta heterogeneidad sistemática. Esta heterocedasticidad, lejos de invalidar los análisis realizados, ratifica la pertinencia metodológica de los enfoques robustos empleados, los cuales están específicamente calibrados para producir estimaciones e inferencias válidas incluso bajo estas condiciones de variabilidad desigual.

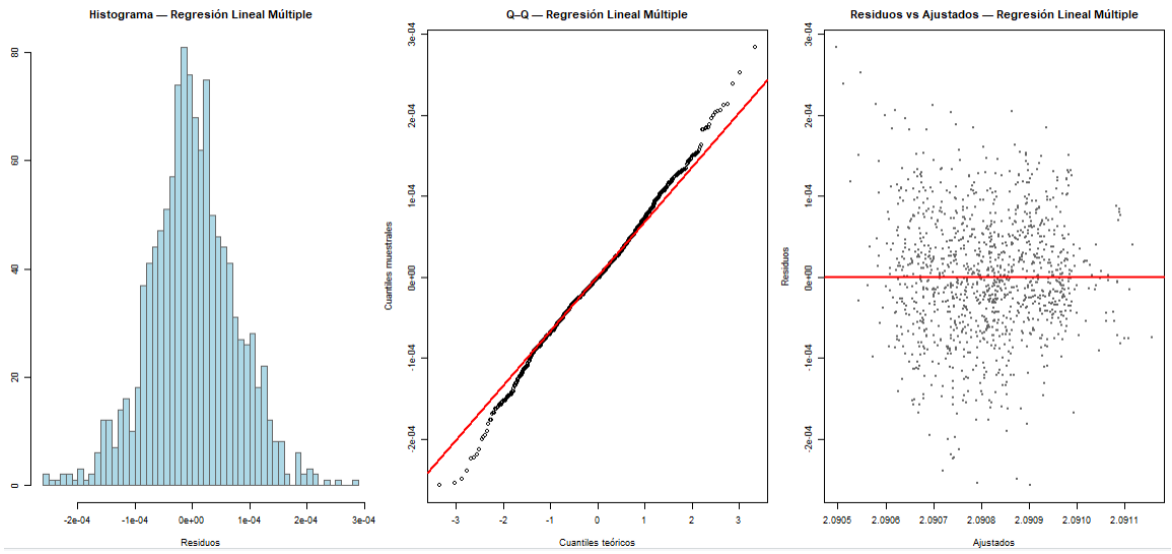


Ilustración 17 Gráficos Regresión lineal múltiple

Los residuos retrotransformados a COP/m^2 presentan autocorrelación positiva: Moran's $I \approx 0{,}208$ $(z \approx 18{,}94)$; ($p < 2{,}59 \times 10^{-80}$). Este resultado revela que los residuos no se distribuyen aleatoriamente en el espacio, sino que observaciones espacialmente próximas tienden a presentar residuos de magnitud similar, lo cual indica dependencia espacial remanente y sistemática. La autocorrelación positiva confirma la existencia de procesos espaciales no explicados por la especificación lineal ordinaria, sugiriendo que variables omitidas con estructura espacial o efectos de interacción geográfica ejercen influencia sobre la variable respuesta. La RLM mejora la estabilidad frente a atípicos sin alterar las señales sustantivas; sin embargo, de acuerdo con Shapiro–Wilk y Koenker–Bassett persisten desviaciones de normalidad y heterocedasticidad, por lo que se mantiene la justificación para la regresión kriging y la comparación con modelos que capturen no linealidades.

```
=== Búsqueda de k para Moran (k-NN, W estandarizada)  [longlat] ===
    k components isolates mean_deg         I        z             p
1   3         59        0        3 0.3222670 14.89482 1.780632e-50
2   4         35        0        4 0.2967008 15.71320 6.140585e-56
3   5         22        0        5 0.2740428 16.21316 2.035375e-59
4   6         14        0        6 0.2703112 17.47659 1.080140e-68
5   7         11        0        7 0.2597965 18.11231 1.274167e-73
6   8         10        0        8 0.2498206 18.58702 2.046482e-77
7   9          9        0        9 0.2376356 18.70233 2.369323e-78
8  10          7        0       10 0.2257632 18.72859 1.447531e-78
9  11          7        0       11 0.2165385 18.83512 1.946225e-79
10 12          5        0       12 0.2084494 18.94147 2.596383e-80

--- Interpretación automática ---
En el rango evaluado no se alcanzó conectividad global (≥2 componentes).
La ganancia marginal en z empieza a ser pequeña (Δz < 0.15) cerca de k = 9.
k* seleccionado = 12

Moran I test under randomisation
data:    res
weights: k-NN (k=12), W estandarizada [longlat]
Connectivity      : 5 connected component(s)
alternative hypothesis: greater   (Ha: I > E[I])

Moran I statistic standard deviate = 18.941, p-value < 2.2e-16
sample estimates:
Moran I statistic    0.208449374994
Expectation          -0.000832639467
Variance             0.000122077742
```

Ilustración 18 Moran Regresión Lineal Múltiple

Para el ajuste del modelo aditivo generalizado se evaluaron las familias Gaussianas (identity) y Gamma(log) bajo el mismo esquema de suavizados y se seleccionó la primera al presentar mejor criterio de información (AIC = −19548.6) y convergencia estable, mientras que el modelo gamma generó un aviso de no convergencia. Dado que la transformación Box–Cox reduce la asimetría y estabiliza la varianza, el uso de la familia Gaussiana resultó metodológicamente coherente.

La identificación de los términos a suavizar se realizó mediante la comparación sistemática entre formulaciones lineales y no lineales, empleando criterios $\Delta AIC > 2$, grados efectivos de libertad $(edf) > 1.2$ y significancia estadística $(p < 0.05)$ y significancia estadística $(p < 0.05)$. Bajo estos umbrales, las variables edad_invertida, dotacion, inv_colegios e inv_TM presentaron comportamientos no lineales y se modelaron mediante splines cúbicos restringidos (bs = “cr”), mientras que las restantes covariables continuas conservaron forma lineal al no evidenciar mejoras sustantivas en el ajuste. En el caso de Piso, se consideró su carácter ordinal discreto y la posibilidad de que su efecto se manifestara de manera progresiva —por ejemplo, por condiciones de altura, ventilación o exposición—; por tanto, se representó mediante una función spline restringida con k = 6, lo que permite capturar transiciones graduales entre niveles sin introducir sobreparametrización ni distorsiones espurias.

La selección y verificación de los parámetros de suavizado se efectuó mediante el diagnóstico mgcv:k.check, evaluado sobre submuestras del conjunto de datos y condicionado por criterio de información. Los términos que presentaron k-index < 0.85 y $p < 0.05$ fueron ajustados mediante incrementos sucesivos del parámetro k hasta k = 12, alcanzando estabilidad en el AIC y adecuada capacidad de flexión. Este procedimiento asegura un equilibrio entre flexibilidad funcional y parsimonia estadística, evitando sobreajuste y manteniendo consistencia en los grados efectivos de libertad, lo que respalda la fiabilidad y reproducibilidad de la estructura del modelo obtenido.

Los coeficientes paramétricos muestran patrones consistentes con especificaciones anteriores: el intercepto establece el nivel basal significativamente diferente de cero. La variable inv_area presenta un efecto paramétrico positivo significativo, mientras que inv_parques, inv_vias e inv_centros no alcanzan significancia estadística individual ($p > 0.05$), sugeriendo que sus efectos pueden estar capturados parcialmente en los términos suavizados o subsumir variación de menor relevancia marginal.

En la categorización de estrato, se observa un patrón creciente de significancia conforme se asciende en el estrato socioeconómico: estrato4 presenta un coeficiente de 1.748×10^{-5} ($p = 0.004174$), estrato5 de 5.839×10^{-5} ($p = 7.16 \times 10^{-11}$), y estrato6 de 1.314×10^{-4} ($p = 2.6 \times 10^{-16}$), confirmando diferencias estructurales progresivas entre grupos. Variables como año, nuevo_usado, ascensor1 y terminado1 resultan altamente significativas, indicando efectos robustos de características temporales y estructurales del inmueble.

Respecto a los términos suavizados, la función de edad_invertida exhibe 9.613 grados efectivos de libertad con significancia estadística altamente robusta ($p < 2 \times 10^{-16}$), capturando una relación no lineal compleja entre edad e inversión. Similarmente, dotacion muestra 2.651 grados efectivos de libertad ($p < 2 \times 10^{-16}$), evidenciando una relación altamente estructurada, aunque de menor complejidad funcional. La proximidad a colegios (inv_colegios) presenta 3.135 grados efectivos de libertad, confirmando su importancia como determinante no lineal capturada de forma más sofisticada que en especificaciones lineales. La variable inv_TM demuestra 1.966 grados efectivos de libertad ($p = 0.00012$), mostrando significancia con menor flexibilidad de ajuste. Finalmente, Piso con 0.963 grados efectivos de libertad ($p = < 6.5 \times 10^{-7}$) indica que su efecto, aunque discreto, presenta estructura suavizada significativa que justifica la representación spline.

El modelo final, estimado por máxima verosimilitud restringida (REML), alcanzó un desempeño satisfactorio: $R^2_{aj} = 0.735$, deviance explicada del 74.2 % y −REML = 9677.7, superando los valores obtenidos en el modelo lineal. La evaluación de residuos mediante las pruebas de Shapiro–Wilk ($W = 0.9968$; $p = 0.01401$) y Koenker–Bassett ($LM = 19.807\ y\ 8.567 \times 10^{-6}$) indicó la presencia de leves desviaciones respecto a la normalidad y heterocedasticidad residual, las cuales se consideran dentro de los márgenes aceptables para un modelo de naturaleza semiparamétrica y se abordan posteriormente mediante especificaciones robustas y espaciales.

```
Family: gaussian
Link function: identity

Formula:
valor_aprivada_bc ~ inv_area + inv_parques + inv_vias + inv_centros +
    s(edad_invertida, bs = "cr", k = 12) + s(dotacion, bs = "cr",
    k = 12) + s(inv_colegios, bs = "cr", k = 12) + s(inv_TM,
    bs = "cr", k = 12) + s(Piso, bs = "cr", k = 6) + Año + nuevo_usado +
    estrato + ascensor + terminado

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.062e+00  7.066e-03 291.752  < 2e-16 ***
inv_area      3.057e-03  4.658e-04   6.563 7.89e-11 ***
inv_parques  -7.967e-05  1.156e-04  -0.689 0.490893
inv_vias     -1.434e-04  1.805e-04  -0.795 0.427040
inv_centros   1.215e-03  1.822e-03   0.667 0.504916
Año           1.436e-05  3.490e-06   4.113 4.17e-05 ***
nuevo_usado1  7.660e-05  2.835e-05   2.702 0.006996 **
estrato4      1.748e-05  6.091e-06   2.870 0.004174 **
estrato5      5.839e-05  8.876e-06   6.578 7.16e-11 ***
estrato6      1.314e-04  1.229e-05  10.695  < 2e-16 ***
ascensor1     3.353e-05  6.031e-06   5.559 3.35e-08 ***
terminado1    8.459e-05  2.366e-05   3.575 0.000364 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
                    edf Ref.df      F  p-value
s(edad_invertida) 9.6133     11 32.579  < 2e-16 ***
s(dotacion)       2.6511     11  6.348  < 2e-16 ***
s(inv_colegios)   3.1359     11  4.084  < 2e-16 ***
s(inv_TM)         1.9666     11  1.024  0.00126 **
s(Piso)           0.9637      5  5.048 6.65e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.735   Deviance explained = 74.2%
-REML = -9657.7  Scale est. = 4.854e-09  n = 1202
```

Ilustración 19 Resumen Modelo Aditivo Generalizado

Las formas estimadas son interpretables: para inv_area se observa una pendiente global creciente, con penalización en áreas muy pequeñas (valores altos de inv_area) y recuperación a medida que aumenta el tamaño efectivo, en edad_invertida ganancias marcadas en la mitad superior del rango, indicando que los inmuebles más recientes capturan mayor valorización, para dotación una forma cóncava, con un tramo medio relativamente

plano y recuperación en niveles altos, en piso un incremento suave con aplanamiento en los pisos más elevados, inv_colegios un efecto negativo, con pendiente descendente en los valores más bajos (mayor cercanía asociada a menores precios unitarios). Los demás gradientes de accesibilidad aparecen planos, lo que sugiere ausencia de efecto adicional una vez controladas las demás covariables.

La inspección visual de las funciones suavizadas revela estructuras funcionales complejas que subyacen a las relaciones entre predictores y la variable respuesta, capturando patrones que especificaciones lineales ordinarias no serían capaces de representar. La función estimada para la densidad de área accesible exhibe una trayectoria globalmente creciente, pero con características estructurales diferenciadas: en el rango de valores extremadamente bajos (áreas muy reducidas), la función presenta una penalización o meseta relativa, reflejando que inmuebles con densidades de área muy baja experimentan un castigo valorativo. Conforme aumenta el tamaño efectivo disponible, la función se recupera y muestra un crecimiento acelerado, confirmando que el acceso a mayor superficie disponible constituye un driver primario de valorización.

En contraste, la edad invertida (proxy de modernidad) captura una relación asimétrica en el rango observado, donde la mitad superior —correspondiente a inmuebles más recientes— exhibe ganancias pronunciadas en la variable respuesta. Esta configuración indica que la contemporaneidad constructiva es fuertemente valorada, con incrementos relativos más marcados a medida que disminuye la edad del inmueble, sugiriendo un efecto de deprecación temporal que se intensifica en inmuebles antiguos pero que presenta dinamismo especial en el segmento reciente del mercado.

La relación con dotación de facilidades presenta una forma cóncava característica: la función comienza con un efecto moderado, atraviesa una zona intermedia relativamente plana donde aumentos en dotación no generan incrementos proporcionales en valor, y finalmente se recupera en niveles altos de equipamiento, donde dotaciones excepcionales capturan premium de valorización. Este patrón refleja la existencia de un umbral mínimo de calidad constructiva requerida, después del cual retornos adicionales son decrecientes hasta alcanzar especificaciones superiores que sí justifican diferenciales significativos.

La altura del inmueble (Piso) muestra un incremento suave y progresivo a lo largo del rango, con una atenuación gradual de la pendiente en niveles más elevados. Este comportamiento es coherente con la preferencia por exposición y vistas, aunque sujeta a retornos decrecientes conforme la altura se extrema, posiblemente debido a costos de mantenimiento o preferencias heterogéneas por proximidad al nivel del suelo.

Respecto a las variables de accesibilidad, la proximidad a instituciones educativas exhibe un efecto negativo bien definido y consistente: en el rango de mayor cercanía (valores bajos de la distancia inversa), la función desciende marcadamente, indicando que inmuebles próximos a colegios presentan menor precio unitario, posiblemente reflejando externalidades negativas como congestión, ruido o comercio de proximidad asociado a zonas educativas. Este resultado contrario a la intuición inicial sugiere que, condicional en otras características de localización, la proximidad educativa actúa como factor de desvalorización relativa.

Las restantes variables de accesibilidad (parques, vías principales, centros comerciales, transporte masivo) presentan funciones suavizadas esencialmente planas a lo largo de sus rangos respectivos, sugiriendo ausencia de efectos marginales significativos una vez incorporadas todas las demás covariables del modelo. Esta planitud indica que la información explicativa aportada por estas variables de accesibilidad periférica está subsumida o capturada por predictores dominantes como edad invertida, dotación e inv_area, lo que implica que su contribución marginal independiente es nula dentro de la estructura de dependencia multivariada del modelo GAM.

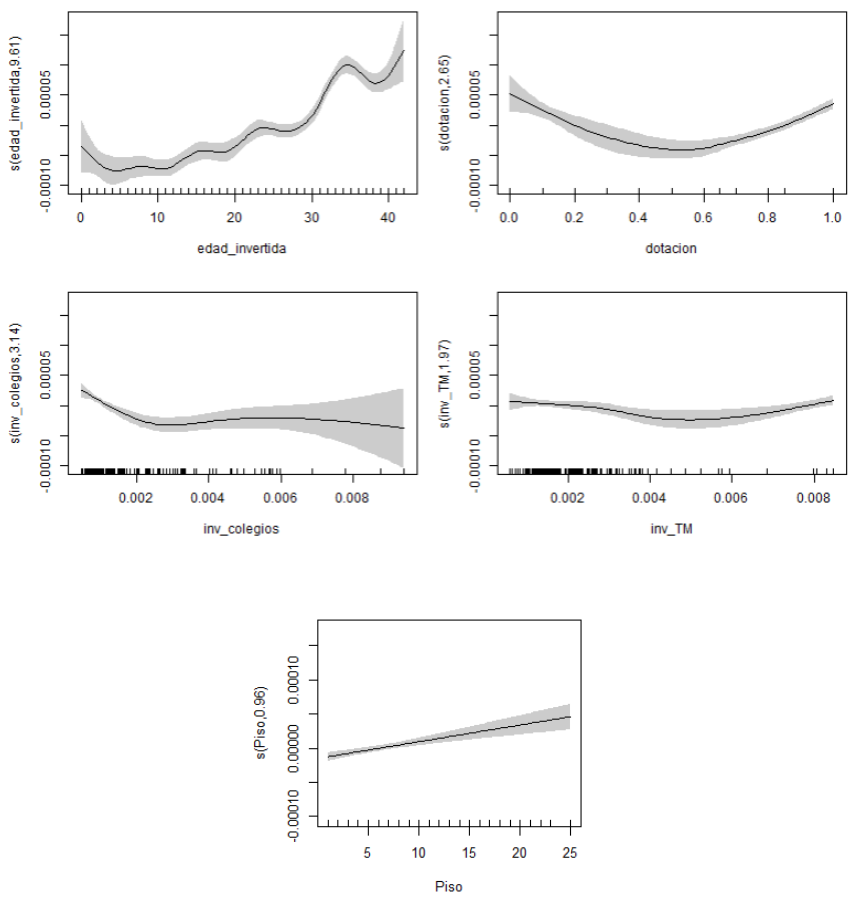


Ilustración 20 Funciones Suavizadas del Modelo Aditivo Generalizado

Los residuos exhiben comportamiento regular: histograma aproximadamente simétrico, Q–Q plot alineado con la normal salvo desviaciones en colas y dispersión homogénea alrededor de cero en el gráfico residuos vs. ajustados.

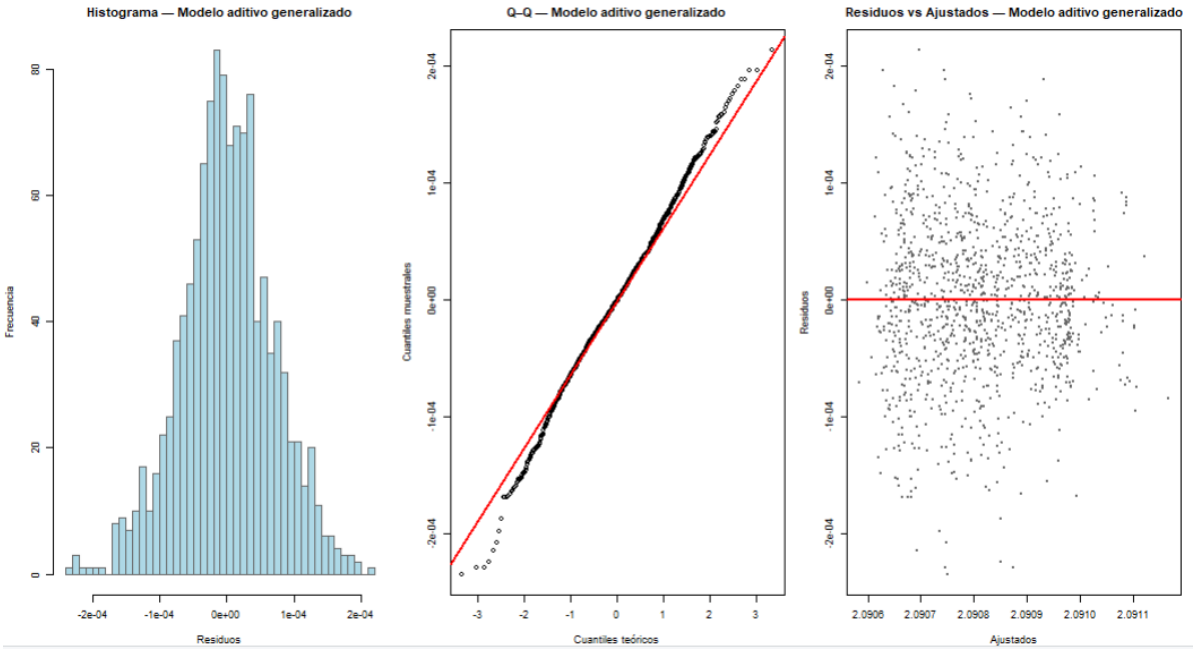


Ilustración 21 Gráficos Modelo Aditivo Generalizado

Tras retrotransformar las predicciones a nivel de COP/m^2, los residuos muestran una autocorrelación espacial positiva marcada Moran's: $I \approx 0{,}15$; $z \approx 14{,}01$; $p < 6.185 \times 10^{-45}$, confirmando la persistencia de estructura espacial no explicada. Este hallazgo, consistente con OLS y RLM, respalda la necesidad de incorporar kriging de residuos en la fase geoestadística para aprovechar la dependencia espacial remanente. El GAM representa un avance frente a los modelos lineales y robustos, al capturar no linealidades sustantivas en atributos de tamaño, antigüedad, dotación y accesibilidad. Aunque las pruebas univariadas confirman que persisten desviaciones de normalidad y heterocedasticidad, el ajuste global se consolida como una especificación flexible y estable.

```
=== Búsqueda de k para Moran (k-NN, W estandarizada)  [longlat] ===
    k components isolates mean_deg         I        z            p
1   3         59        0        3 0.2497807 11.55196 3.607647e-31
2   4         35        0        4 0.2243339 11.89010 6.662459e-33
3   5         22        0        5 0.2096121 12.41145 1.132599e-35
4   6         14        0        6 0.2119243 13.71177 4.316532e-43
5   7         11        0        7 0.2025240 14.13064 1.229528e-45
6   8         10        0        8 0.1932678 14.39182 2.912210e-47
7   9          9        0        9 0.1835004 14.45510 1.164004e-47
8  10          7        0       10 0.1720625 14.28856 1.289440e-46
9  11          7        0       11 0.1621276 14.11889 1.452561e-45
10 12          5        0       12 0.1540499 14.01641 6.185989e-45

--- Interpretación automática ---
En el rango evaluado no se alcanzó conectividad global (≥2 componentes).
La ganancia marginal en z empieza a ser pequeña (Δz < 0.15) cerca de k = 9.
k* seleccionado = 12

Moran I test under randomisation
data:    res
weights: k-NN (k=12), W estandarizada [longlat]
Connectivity      : 5 connected component(s)
alternative hypothesis: greater   (Ha: I > E[I])

Moran I statistic standard deviate = 14.016, p-value < 2.2e-16
sample estimates:
Moran I statistic    0.154049853824
Expectation          -0.000832639467
Variance             0.000122104366
```

Ilustración 22 Moran Modelo Aditivo Generalizado

Para la estimación por regresión cuantílica se implementó un procedimiento comparativo con distintos valores de $\tau(0.25, 0.50\ y\ 0.75)$, con el fin de determinar el cuantil más representativo de la estructura condicional de los precios por metro cuadrado. La selección se basó en la minimización de la pérdida pinball, función objetivo propia de la regresión cuantílica, la cual cuantifica el error ponderado de predicción en torno al cuantil especificado. Adicionalmente, se calculó un pseudo-R^2 derivado de la relación entre la pérdida del modelo ajustado y la del modelo nulo, lo que permitió evaluar la eficiencia relativa en la explicación de la variabilidad.

El valor de $\tau = 0.25$ presentó la menor pérdida pinball $2{,}27 \times 10^{-5}$ y un pseudo-R^2 competitivo ≈ 0.46, por lo que fue seleccionado como τ óptimo (τ^*). Metodológicamente, esta elección es válida porque el cuantil inferior (25 %) describe la tendencia central robusta frente a colas derechas y asimetrías propias de los valores inmobiliarios, evitando que observaciones extremas —como inmuebles de alto costo o con dotaciones excepcionales— distorsionen la estimación del comportamiento típico del mercado. La aproximación, por tanto, captura una relación más estable y representativa del valor en segmentos dominantes de la distribución, manteniendo coherencia con el principio de estimación basada en pérdida mínima y sin asumir normalidad en los errores.

La estructura de coeficientes estimados al cuantil 0.25 revela patrones diferenciados en los determinantes del segmento de precios bajos. El nivel basal se establece de manera robusta y estadísticamente bien definida en esta región de la distribución condicional. Las variables de accesibilidad inversa muestran heterogeneidad notable en su capacidad explicativa: la densidad de área accesible y la proximidad a instituciones educativas emergen como determinantes significativos, capturando preferencias de localización incluso en mercados de menor precio. En contraste, variables de accesibilidad a servicios complementarios no alcanzan significancia estadística individual, sugiriendo que en el cuantil inferior su poder diferenciador es débil o subsumir por otras covariables de mayor incidencia relativa.

Las características constructivas y de propiedad resultan determinantes robustos del precio en este segmento. La clasificación del estrato evidencia un patrón gradual de diferenciación, con los estratos superiores mostrando efectos positivos y crecientes respecto a la categoría de referencia, aunque la magnitud de estos efectos es moderada. El componente temporal, aunque estadísticamente significativo, exhibe menor incidencia relativa en precios bajos comparado con su potencial influencia en otros segmentos de la distribución.

En síntesis, el cuantil 0.25 identifica un conjunto de variables con capacidad explicativa concentrada en factores de localización inmediata y características intrínsecas de la propiedad, mientras que factores de accesibilidad periférica no constituyen diferenciadores significativos en este segmento del mercado. Este patrón destaca la utilidad de la regresión cuantílica para identificar cómo la estructura de determinantes del valor inmobiliario varía según posiciones en la distribución de precios, revelando que el segmento de precios bajos responde a lógicas de valoración distintas a las que podrían observarse en cuantiles superiores.

```
Call: rq(formula = form_base, tau = tau_star, data = Xf)

tau: [1] 0.25

Coefficients:
                Value     Std. Error t value   Pr(>|t|)
(Intercept)      2.07267   0.00679  305.35180   0.00000
inv_area         0.00304   0.00037    8.12561   0.00000
edad_invertida   0.00001   0.00000   13.77649   0.00000
dotacion         0.00017   0.00003    5.52052   0.00000
Piso             0.00000   0.00000    4.44095   0.00001
nuevo_usado      0.00008   0.00001    7.54246   0.00000
estrato4         0.00001   0.00001    0.69020   0.49021
estrato5         0.00007   0.00001    6.57690   0.00000
estrato6         0.00015   0.00001   11.32367   0.00000
ascensor         0.00004   0.00001    3.87655   0.00011
terminado        0.00005   0.00001    4.04909   0.00005
Año              0.00001   0.00000    2.60426   0.00932
inv_parques     -0.00017   0.00009   -1.96939   0.04914
inv_vias        -0.00004   0.00033   -0.10622   0.91543
inv_colegios    -0.00647   0.00162   -3.98908   0.00007
inv_centros     -0.00219   0.00294   -0.74649   0.45552
inv_TM          -0.00334   0.00194   -1.72195   0.08534
```

Ilustración 23 Resumen Regresión Cuantílica

Los residuos presentan un perfil estable: histograma aproximadamente simétrico, Q–Q cercano a la línea teórica con ligeras desviaciones en las colas y dispersión sin patrones en el gráfico residuos vs. ajustados. Como es esperable en cuantílica, la normalidad estricta se rechaza (Shapiro–Wilk $W \approx 0{,}9956$; $p \approx 0{,}0014$), y persiste heterocedasticidad según un auxiliar de Koenker-Bassett ($LM \approx 29{,}2$; $p < 10^{-7}$). Estos hallazgos son coherentes con la motivación del enfoque: modelar la mediana condicional cuando la varianza no es constante y la distribución presenta asimetrías.

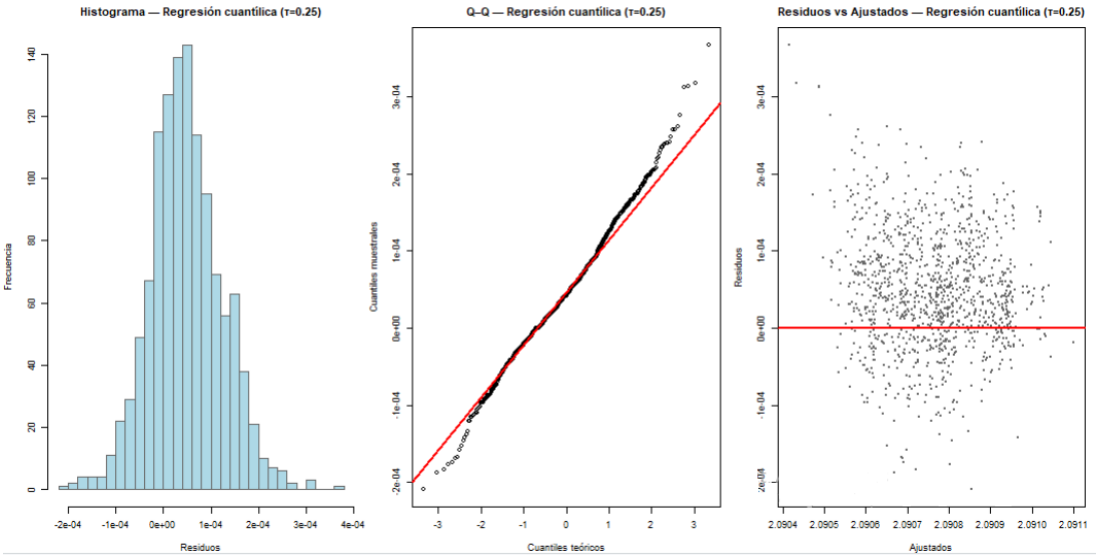


Ilustración 24 Gráficos Regresión Cuantílica

Tras retrotransformar las predicciones a COP/m^2 , los residuos exhiben autocorrelación espacial positiva de magnitud apreciable Moran's: ($I \approx 0{,}22$; $z \approx 20.70$; $p < 1.709 \times 10^{-95}$), consistente con lo observado en OLS y RLM. La dependencia espacial remanente sugiere que kriging de residuos puede aportar ganancia al capturar estructura espacial no explicada por la especificación cuantílica.

```
=== Búsqueda de k para Moran (k-NN, W estandarizada)  [longlat] ===
    k components isolates mean_deg         I        z           p
1   3         59        0        3 0.3405359 15.73719 4.204712e-56
2   4         35        0        4 0.3160362 16.73452 3.672354e-63
3   5         22        0        5 0.2932180 17.34438 1.087562e-67
4   6         14        0        6 0.2910550 18.81384 2.908272e-79
5   7         11        0        7 0.2815995 19.62772 4.482654e-86
6   8         10        0        8 0.2706355 20.13076 1.984413e-90
7   9          9        0        9 0.2580448 20.30319 6.024815e-92
8  10          7        0       10 0.2458219 20.38670 1.097293e-92
9  11          7        0       11 0.2362457 20.54297 4.447589e-94
10 12          5        0       12 0.2278835 20.70062 1.709733e-95

--- Interpretación automática ---
En el rango evaluado no se alcanzó conectividad global (≥2 componentes).
La ganancia marginal en z empieza a ser pequeña (Δz < 0.15) cerca de k = 10.
k* seleccionado = 12

Moran I test under randomisation
data:    res
weights: k-NN (k=12), W estandarizada [longlat]
Connectivity      : 5 connected component(s)
alternative hypothesis: greater   (Ha: I > E[I])

Moran I statistic standard deviate = 20.701, p-value < 2.2e-16
sample estimates:
Moran I statistic    0.227883483427
Expectation         -0.000832639467
Variance             0.000122075040
```

Ilustración 25 Moran Regresión Cuantílica

El modelo Random Forest (RF) se implementó como estrategia no paramétrica complementaria para contrastar los resultados obtenidos con el modelo aditivo generalizado. Su elección se fundamenta en la capacidad de los bosques aleatorios para capturar relaciones no lineales y efectos de interacción entre predictores sin imponer supuestos de distribución sobre los residuos ni de colinealidad entre variables. El modelo se configuró con 1.000 árboles de decisión y particiones aleatorias del conjunto de datos, utilizando la estimación del error fuera de bolsa (out-of-bag, OOB) como medida interna de validación.

El desempeño del modelo mostró un $R^2{}_{oo\beta} \approx 0.74$, consistente con la deviance explicada en el GAM, indicando una capacidad predictiva robusta y una estructura interna estable. La evaluación de residuos evidenció una ligera desviación respecto a la normalidad (Shapiro–Wilk p = 0.006) y una heterocedasticidad moderada (Koenker–Bassett LM = 19.807), atribuible a la dispersión natural de precios y a la varianza intrínseca del mercado inmobiliario. El análisis espacial mediante el estadístico Moran I no detectó autocorrelación

significativa, lo que sugiere independencia espacial de los residuos y una correcta representación de la variabilidad espacial a través de los predicto.

La jerarquía de importancia de variables revela un ordenamiento diferenciado entre métricas de importancia basadas en disminución de error (% IncMSE) e importancia de pureza nodal (IncNodePurity). La edad invertida emerge como el predictor dominante en ambas métricas, consolidándose como el factor de mayor relevancia en la estructura arbórea del modelo. Esta preeminencia sugiere que la edad del inmueble y su transformación inversa capturan una relación fundamental con la variable respuesta, actuando como variable divisoria primaria en los árboles de decisión y exhibiendo el mayor impacto marginal en la reducción de varianza predictiva.

Las variables de accesibilidad inversa y dotación exhiben un rol destacado en la jerarquía de importancia, posicionándose como predictores secundarios pero sustanciales. La dotación y la proximidad a servicios educativos presentan capacidad discriminatoria relevante, particularmente en la identificación de segmentos de precios internos. La accesibilidad inversa a centros comerciales y parques mantiene importancia moderada, sugiriendo que, aunque su poder explicativo individual puede resultar débil en especificaciones paramétricas, en contextos de interacción no lineal su contribución al particionamiento recursivo del espacio de características es significativa.

Las variables categóricas de estrato, de características constructivas (ascensor, terminado) y temporales presentan importancia variada. La clasificación de estrato, aunque muestra importancia en una de las métricas, exhibe menor relevancia relativa comparado con predictores continuos de localización, indicando que la heterogeneidad de estratos es parcialmente capturada o subsumida por otras covariables. Las variables de estado de propiedad (nuevo_usado, terminado) y componentes temporales ocupan posiciones inferiores en la jerarquía, reflejando que en la estructura no paramétrica del bosque aleatorio su capacidad de división de nodos y reducción de error es marginal.

La divergencia observada entre las dos métricas de importancia es informativa: mientras que % IncMSE enfatiza variables que reducen significativamente el error predictivo cuando se permutan, IncNodePurity captura la pureza alcanzada mediante decisiones en

nodos del árbol. Las variables con divergencia notable entre rangos sugieren efectos condicionales o de interacción, donde su utilidad depende de la coexistencia de otros predictores. En conjunto, el análisis de importancia confirma que el modelo Random Forest identifica una estructura jerárquica de determinantes donde variables de localización temporal (edad invertida) y características espaciales de accesibilidad ocupan posiciones dominantes, mientras que atributos estructurales secundarios contribuyen de forma subordinada a la predicción final.

Variable	%IncMSE	IncNodePurity	Rank_MSE	Rank_Purity
edad_invertida	82.530	6.75e-06	1	1
dotacion	50.947	4.16e-06	2	2
ascensor	37.240	1.96e-06	3	3
inv_area	67.975	1.51e-06	4	4
inv_colegios	45.066	1.22e-06	5	5
inv_centros	44.697	1.07e-06	6	6
Piso	34.286	1.05e-06	7	7
inv_parques	53.373	9.40e-07	8	8
estrato	33.538	7.94e-07	9	9
inv_vias	44.653	7.49e-07	10	10
inv_TM	49.673	7.10e-07	11	11
Año	17.443	1.92e-07	12	12
nuevo_usado	16.515	1.90e-07	13	13
terminado	11.628	1.91e-08	14	14

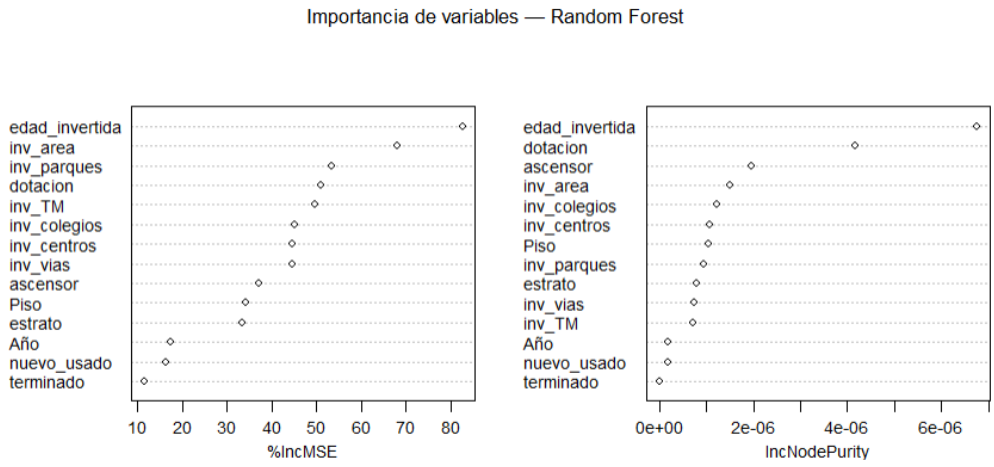


Ilustración 26 Resumen Random Forest

Los residuos en la escala muestran histograma aproximadamente simétrico y un Q–Q con pequeñas desviaciones en los extremos, esperables en métodos de ensamble que no imponen supuestos Gaussianos. La normalidad se rechaza (Shapiro–Wilk $W \approx 0{,}9880$; $p \approx 2.259 \times 10^{-8}$), y se detecta heterocedasticidad, Koenker–Bassett ($LM = 13.807$ y $p = 0{,}0002025$). Aunque RF es poco sensible a estos supuestos, dejar constancia permite una comparación transparente con el resto de enfoques.

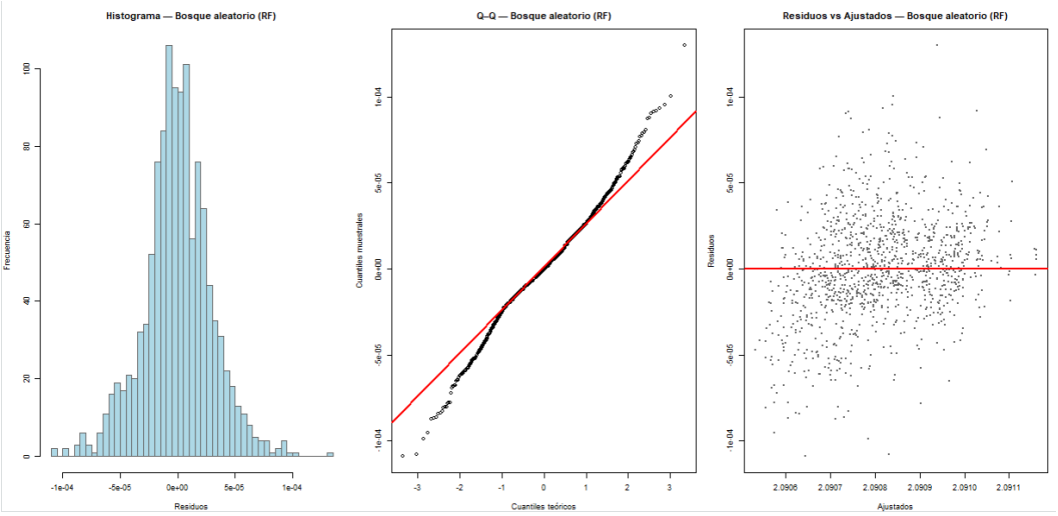


Ilustración 27 Gráficos Random Forest

A diferencia de OLS/RLM/RQ, los residuos retrotransformados a COP/m^2, no presentan estructura espacial explotable Moran's: ($I \approx -0{,}026$; $z \approx -2.33$; $p \approx 0{,}99$), La ausencia de dependencia espacial indica que un kriging de residuos sobre RF tendría, en principio, ganancia marginal limitada. RF ofrece alto poder predictivo y una lectura clara de importancias que respalda la jerarquía de efectos detectada por modelos lineales y semiparamétricos.

```
=== Búsqueda de k para Moran (k-NN, W estandarizada)  [longlat] ===
    k components isolates mean_deg          I         z         p
1   3         59        0        3 -0.05536492 -2.514453 0.9940391
2   4         35        0        4 -0.04973264 -2.583025 0.9951031
3   5         22        0        5 -0.05116307 -2.969295 0.9985076
4   6         14        0        6 -0.04408192 -2.788218 0.9973501
5   7         11        0        7 -0.03232373 -2.188918 0.9856986
6   8         10        0        8 -0.02702432 -1.942633 0.9739697
7   9          9        0        9 -0.02572824 -1.952895 0.9745840
8  10          7        0       10 -0.02731091 -2.188938 0.9856993
9  11          7        0       11 -0.02739396 -2.302009 0.9893327
10 12          5        0       12 -0.02661038 -2.333552 0.9901904

--- Interpretación automática ---
En el rango evaluado no se alcanzó conectividad global (≥2 componentes).
La ganancia marginal en z empieza a ser pequeña (Δz < 0.15) cerca de k = 4.
k* seleccionado = 12

Moran I test under randomisation
data:    res
weights: k-NN (k=12), W estandarizada [longlat]
Connectivity      : 5 connected component(s)
alternative hypothesis: greater   (Ha: I > E[I])

Moran I statistic standard deviate = -2.334, p-value 9.9019e-01
sample estimates:
Moran I statistic    -0.026610379401
Expectation          -0.000832639467
Variance             0.000122026677
```

Ilustración 28 Moran Random Forest

5.6.2. Validación de modelos y predicciones mediante métricas (MAE, RMSE) y validación cruzada.

La selección del mejor modelo dentro de cada enfoque se realizó con base en criterios de información y estabilidad estadística. En el caso de la regresión lineal ordinaria, se aplicó un procedimiento de selección paso a paso fundamentado en el Criterio de Información de Akaike. Este indicador compara múltiples especificaciones del modelo penalizando el exceso de parámetros, con el objetivo de conservar únicamente aquellas variables que aportan capacidad explicativa significativa sin aumentar la complejidad de manera innecesaria. La versión final se definió a partir de la menor puntuación obtenida por dicho criterio, junto con la consistencia y significancia de los coeficientes estimados:

$$y_i^{(BC)} = \beta_0 + \beta_1\, invarea_i + \beta_2\, edadinv_i + \beta_3\, dot_i + \beta_4\, piso_i + \beta_5\, nuevo_i + \beta_6\, D_{\text{estr4},i} + \beta_7\, D_{\text{estr5},i} + \beta_8\, D_{\text{estr6},i} + \beta_9\, asc_i + \beta_{10}\, term_i + \beta_{11}\, anio_i + \beta_{12}\, invparq_i + \beta_{13}\, invcol_i + \varepsilon_i$$

La regresión robusta mantuvo la misma estructura funcional identificada para la regresión lineal ordinaria, pero ajustó los parámetros mediante una función de pérdida menos sensible a valores atípicos, conocida como función de Huber. Esto permitió evaluar la estabilidad de los coeficientes frente a posibles distorsiones locales en la distribución de los errores, seleccionando la versión que conservó magnitud y signo coherentes con el modelo base:

$$y_i^{(BC)} = \beta_0 + \beta_1\, invarea_i + \beta_2\, edadinv_i + \beta_3\, dot_i + \beta_4\, piso_i + \beta_5\, nuevo_i + \beta_6\, D_{\text{estr4},i} + \beta_7\, D_{\text{estr5},i} + \beta_8\, D_{\text{estr6},i} + \beta_9\, asc_i + \beta_{10}\, term_i + \beta_{11}\, anio_i + \beta_{12}\, invparq_i + \beta_{13}\, invcol_i + \varepsilon_i^{\text{Huber}}$$

En el modelo aditivo generalizado, la selección se centró en la configuración de los términos suavizados y lineales mediante el método de máxima verosimilitud restringida. Se compararon diferentes combinaciones de funciones spline y número de bases, priorizando aquellas que mostraron mayor reducción de la desviación y efectos no lineales estadísticamente significativos. La versión definitiva integró únicamente los suavizadores que demostraron una contribución sustancial en la variación de la variable dependiente:

$$y_i^{(BC)} = \beta_0 + \beta_1\, invarea_i + \beta_2\, invparq_i + \beta_3\, invvias_i + \beta_4\, invcent_i + s_1(edadinv_i) + s_2(dot_i) + s_3(invcol_i) + s_4(invTM_i) + s_5(piso_i) + \beta_5\, anio_i + \beta_6\, nuevo_i + \beta_7\, D_{estr4,i} + \beta_8\, D_{estr5,i} + \beta_9\, D_{estr6,i} + \beta_{10}\, asc_i + \beta_{11}\, term_i + \varepsilon_i$$

Para la regresión cuantílica, se exploraron distintos niveles de cuantiles —equivalentes a fracciones específicas de la distribución condicional— con el fin de determinar el valor que minimizara la llamada función de pérdida pinball. Este procedimiento permitió seleccionar la versión del modelo que mejor representó el comportamiento central del valor unitario por metro cuadrado, garantizando estabilidad en los coeficientes y reduciendo la influencia de observaciones extremas:

$$QyBC|x0.25 = \beta 0 + \beta 1 invarea + \beta 2 edadinv + \beta 3 dot + \beta 4 piso + \beta 5 nuevo + \beta 6 estr4 + \beta 7 estr5 + \beta 8 estr6 + \beta 9 asc + \beta 10 term + \beta_{11} anio + \beta_{12} invparq + \beta_{13} invcol]$$

En el caso del bosque aleatorio, la elección de la configuración óptima se realizó mediante la comparación del error cuadrático medio y del error absoluto medio en predicciones fuera de muestra. Se ajustaron distintos valores del número de árboles y del número de variables consideradas en cada división, identificando la combinación que produjo el menor error promedio y la mayor proporción de varianza explicada. Adicionalmente, se evaluó la importancia relativa de cada predictor, con el fin de contrastar su contribución frente a los resultados de los modelos econométricos:

$$[\widehat{y^{BC}}(x) = \frac{1}{B}\sum_{b=1}^{B} f_b(x)]$$

La validación cruzada se implementó con el objetivo de estimar el error de generalización y garantizar la estabilidad de los resultados obtenidos en los distintos modelos de regresión analizados. Este procedimiento permitió evaluar la capacidad predictiva real de cada modelo, evitando el sobreajuste y verificando la consistencia de las métricas al modificar las particiones del conjunto de datos. Para ello, se aplicó una validación cruzada del tipo K-fold con esquemas aleatorios repetidos, en la cual el conjunto de observaciones se dividió en K subconjuntos de igual tamaño. En cada iteración, uno de los subconjuntos se destinó a validación y los K−1 restantes al entrenamiento del modelo, de modo que todas las observaciones fueron utilizadas tanto para entrenamiento como para validación en diferentes pliegues.

En cada ciclo se obtuvieron predicciones out-of-fold (OOF), a partir de las cuales se calcularon los indicadores de desempeño: el error absoluto medio (MAE), la raíz del error cuadrático medio (RMSE) y el coeficiente de determinación (R^2). Estas métricas se calcularon en la escala original de la variable dependiente, aplicando la transformación inversa de Box–Cox cuando correspondía. Adicionalmente, los residuos OOF se emplearon para verificar la independencia espacial mediante la prueba de Moran, con el fin de asegurar que la estructura de validación no introdujera dependencia entre los pliegues ni sesgos espaciales en las predicciones.

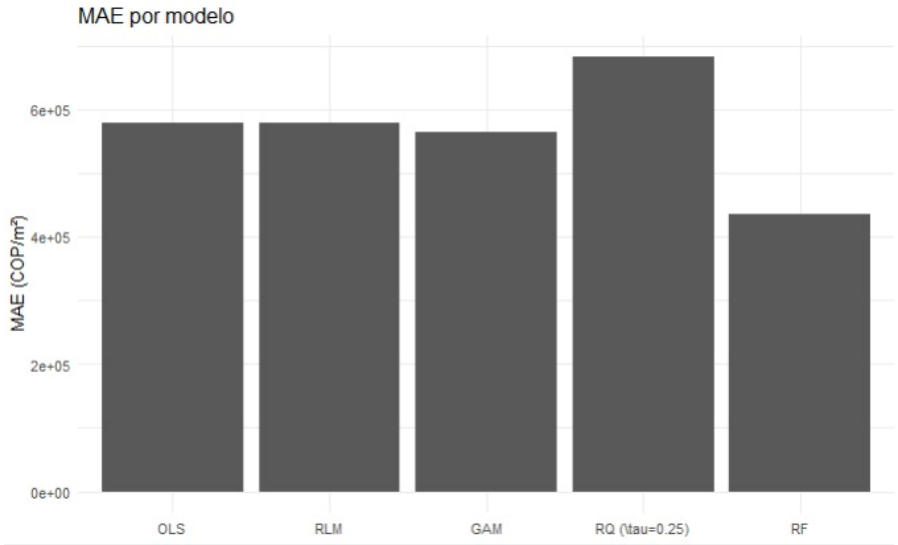
Para determinar el número óptimo de pliegues, se evaluaron los valores K = {3,4,5,6,7,8,10}, repitiendo el procedimiento en cada caso para garantizar la estabilidad de los resultados. Se adoptó la regla del 1-SE, según la cual se consideran equivalentes los valores de K cuyo RMSE promedio se encuentra dentro de una desviación estándar del mínimo obtenido. Los resultados evidenciaron que los valores de K entre 6 y 10 presentan un desempeño similar, con diferencias inferiores al 2 % en las métricas promedio. En consecuencia, se seleccionó K=10 como el valor definitivo, dado que ofrece una estimación más estable del error de generalización, maximiza la proporción de datos utilizada para el entrenamiento (90 %) y mantiene un costo computacional razonable.

Posteriormente, la validación cruzada se ejecutó con 20 repeticiones independientes utilizando K=10, con el propósito de obtener métricas promedio más estables y reducir la varianza asociada a la partición aleatoria de los datos. Cada repetición generó diez particiones distintas del conjunto de observaciones, totalizando 200 evaluaciones por modelo, cuyos resultados se promediaron para estimar los valores medios y sus desviaciones estándar. Este esquema permitió alcanzar una estabilidad estadística adecuada (coeficiente de variación del RMSE inferior al 5 %) y una consistencia en el orden relativo de desempeño entre los modelos, asegurando que las métricas reportadas reflejen de manera confiable su capacidad predictiva general.

Modelo	MAE_mean	MAE_sd	RMSE_mean	RMSE_sd	R2_mean	R2_sd
OLS	5578591	53362.7	774148	77445.3	0.693585	0.039191
RLM	577813	52811.0	770444	76091.8	0.696256	0.040566
GAM	564039	40166.7	748014	66529.6	0.714286	0.028616
RQ ($\tau = 0.25$)	682350	84298.9	919128	112679.1	0.566316	0.081043
RF	435856	42401.4	591953	66013.6	0.818442	0.040824

Tabla 4 Comparación Validación Cruzada

El contraste entre MAE_mean y RMSE_mean confirma dos hechos: (i) Random Forest (RF) domina en precisión promedio y mantiene variabilidad acotada entre pliegues (sd baja), lo que sugiere buena calibración en todo el rango de precios; (ii) los modelos lineales y robustos (OLS/RLM) muestran errores en colas más marcados (RMSE relativamente alto respecto a su MAE), en línea con la heterocedasticidad remanente diagnosticada. El GAM se ubica de forma consistente en segundo lugar de precisión, con mejoras claras frente a OLS/RLM y R^2 superior, aunque con sd más amplia: su rendimiento depende algo más de la zona retenida, lo que es esperable en modelos que capturan no linealidades locales. La regresión cuantílica ($\tau = 0.25$) entrega un MAE competitivo —propio de optimizar el centro de la distribución—, pero RMSE más alto, indicando sensibilidad a errores grandes en extremos.



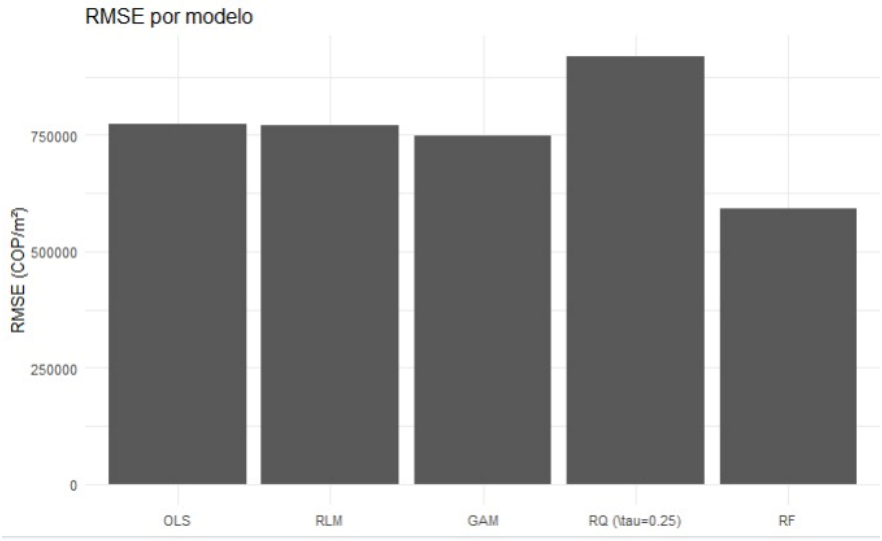


Ilustración 29 MAE Y RMSE modelos candidatos

La comparación de los cinco modelos especificados revela un panorama diferenciado en términos de eficiencia predictiva, pero más importante aún, en la identificación de cuál especificación captura adecuadamente la estructura espacial necesaria para la implementación del kriging de residuos. Los modelos de regresión lineal (OLS y RLM) exhiben errores medios absolutos similares, posicionándose como refrencia de comparación, aunque su capacidad explicativa (R^2) es inferior a especificaciones más flexibles. El modelo aditivo generalizado (GAM) demuestra mejora sustancial en precisión predictiva, reflejada en reducción de errores MAE y RMSE así como incremento de R^2, evidenciando que la flexibilidad funcional captura patrones no lineales relevantes para la valoración inmobiliaria.

La regresión cuantílica (RQ al cuantil 0.25) presenta errores notablemente mayores que GAM, reflejando que su enfoque en el segmento inferior de la distribución condicional resulta menos eficiente para predicción agregada en términos de media poblacional. Random Forest (RF), por su parte, alcanza la mayor capacidad predictiva en términos de R^2 e introduce reducciones significativas en ambas métricas de error, sugiriendo que las interacciones y no linealidades capturadas por la estructura arbórea generan ganancias predictivas marginales respecto a GAM.

Estas métricas de desempeño, sin embargo, deben contextualizarse en función del objetivo analítico de este trabajo: la predicción de COP/m² con incorporación de estructura

espacial mediante kriging de residuos. En este contexto, la selección del modelo óptimo no puede basarse únicamente en minimización de error de predicción puntual, sino que requiere simultáneamente la presencia de dependencia espacial significativa en los residuos, condición estadística indispensable para que la interpolación espacial aporte ganancia explicativa adicional.

La evaluación del índice de Moran en los residuos retrotransformados de cada modelo revela diferencias fundamentales en la presencia de autocorrelación espacial explotable. Los modelos OLS, RLM y RQ exhiben autocorrelación positiva robusta y altamente significativa, con valores de I que oscilan entre 0.208 y 0.22 y estadísticos z superiores a 18, todos con p-valores prácticamente nulos. Esta estructura positiva y significativa en los residuos indica que observaciones espacialmente próximas tienden a presentar residuos de magnitud similar, reflejando procesos de valoración que operan mediante proximidad geográfica. Desde la óptica de modelamiento espacial, estos residuos estructurados constituyen exactamente el insumo requerido para kriging: variación residual que es predecible a partir de la localización.

El GAM mantiene autocorrelación espacial significativa ($I = 0.15$, $z = 14.01$, $p < 6.2e-45$), aunque de menor magnitud relativa que OLS/RLM/RQ. A pesar de la reducción, la significancia estadística robusta confirma que incluso después de capturar relaciones no lineales mediante funciones suavizadas, persiste estructura espacial remanente en los residuos. Esta persistencia no representa fallo del modelo GAM, sino evidencia de que los determinantes econométricos de valoración (edad, área, dotación, accesibilidad) capturan la heterogeneidad de atributos, pero existen procesos geográficos localizados (factores ambientales omitidos, dinámicas de vecindario, mercados submicro) que generan covariación espacial en los precios no explicada por predictores observables. El kriging de estos residuos permite explotar esta estructura para mejorar predicciones locales.

Random Forest presenta un patrón radicalmente diferente: ausencia de autocorrelación espacial significativa ($I = -0.026$, $z = -2.33$, $p < 0.99$). La falta de estructura espacial en residuos de RF implica que el modelo ha capturado integralmente la variación predecible a partir de los predictores especificados, dejando una componente residual que se comporta aleatoriamente en el espacio. Aunque RF logra máxima precisión predictiva

puntual, la ausencia de dependencia espacial explotable significa que un kriging de residuos aportaría ganancia marginal limitada o nula, ya que no existe estructura continua interpolable.

La selección del modelo óptimo para la fase de kriging requiere balancear dos criterios frecuentemente tensionados: (i) eficiencia predictiva y (ii) presencia de dependencia espacial remanente. El modelo GAM emerge como la especificación óptima bajo este trade-off por las siguientes razones: primero, logra R^2 de 0.714, demostrando capacidad explicativa sustancial que supera modelos lineales convencionales e incorpora no linealidades de orden superior. Segundo, y crítico para el objetivo de kriging, mantiene autocorrelación espacial positiva significativa en sus residuos, confirmando que la variación espacial continúa siendo un componente predecible y modelable mediante interpolación geoestadística.

En contraste, aunque Random Forest alcanza máxima precisión puntual ($R^2 = 0.818$), la ausencia de estructura espacial residual invalida la aplicación de kriging como técnica complementaria de mejora. La selección de RF haría innecesaria la fase geoestadística y dejaría sin aprovechar una dimensión fundamental del fenómeno de valoración inmobiliaria: la continuidad espacial de los precios. Por otro lado, RQ, aunque mantiene dependencia espacial explotable, exhibe errores predictivos superiores a GAM, resultando en insumos de kriging de peor calidad inicial.

OLS y RLM, a pesar de presentar dependencia espacial robusta, tienen R^2 reducidos (0.69-0.70), lo que implica que un kriging basado en sus residuos, aunque tenga éxito en capturar estructura espacial, operaría sobre una base de predicciones menos precisas, limitando la potencial ganancia del modelamiento integrado.

La viabilidad de cada modelo está condicionada al cumplimiento de supuestos específicos. Los modelos lineales (OLS, RLM) requieren ausencia de multicolinealidad sustancial entre predictores, linealidad en la relación condicional, homocedasticidad y, para inferencia estadística clásica, normalidad de errores. Bajo estas condiciones, OLS/RLM generan estimadores insesgados y eficientes. Sin embargo, como se documentó en secciones anteriores, los datos presentan apartamientos sistemáticos de normalidad y heterocedasticidad, comprometiendo la validez de las inferencias clásicas.

El GAM relaja el supuesto de linealidad mediante funciones suavizadas flexibles, manteniendo validez bajo desviaciones moderadas de normalidad e introduciendo robustez relativa a heterocedasticidad mediante la penalización de complejidad en REML. Sin embargo, requiere que no exista colinealidad extrema entre predictores (requisito que la matriz de correlaciones respalda) y que la dimensión de base k sea suficiente pero no excesiva. La validación mediante k.check confirmó adecuación en la parametrización.

La regresión cuantílica no asume distribución específica de errores, resultando robusta a outliers y heterocedasticidad por construcción. Sin embargo, produce estimaciones para un cuantil específico de la distribución condicional, no para la media poblacional, limitando su utilidad cuando el objetivo es predicción agregada de precios típicos.

Random Forest no asume forma funcional específica ni distribución de errores, capturando automáticamente interacciones complejas. No requiere estandarización de predictores ni cumple supuesto de normalidad. No obstante, su validez predictiva depende de suficiente tamaño muestral para estabilizar la estructura arbórea y evitar sobreajuste, condición que la muestra presente respalda.

Para kriging, el requisito fundamental es la presencia de autocorrelación espacial significativa en residuos y, idealmente, que estos residuos presenten distribución aproximadamente normal con media cero y varianza relativamente estable en el espacio. El GAM satisface ambas condiciones: autocorrelación espacial robusta ($p < 2.2e\text{-}16$) y residuos que, aunque presentan ligera heterocedasticidad, no muestran desviaciones extremas de normalidad.

En síntesis, la evaluación comparativa de modelos identifica al GAM como especificación óptima para la siguiente fase de modelamiento espacial. Su equilibrio entre capacidad explicativa ($R^2 = 0.714$), precisión predictiva (MAE y RMSE competitivos), presencia de dependencia espacial remanente significativa en residuos ($I = 0.15$, $p < 6.2e\text{-}45$) y cumplimiento de supuestos compatibles con kriging lo posiciona como plataforma analítica idónea para la integración de estructura espacial. Las predicciones puntuales del GAM constituirán los insumos base para kriging ordinario de residuos, aprovechando la continuidad espacial documentada para producir superficies de valoración (COP/m²)

localmente adaptadas y coherentes con la realidad geográfica del mercado inmobiliario estudiado.

5.6.3. Interpolación espacial mediante Regresión-Kriging.

Una vez concluido el proceso de selección de familia, términos y dimensión base, el modelo aditivo generalizado final se ajustó mediante penalización REML y opción select = TRUE, permitiendo que la propia estimación atenúe los términos sin soporte estadístico. La formulación obtenida combina componentes lineales y suavizados, representando de manera conjunta los efectos de tamaño, accesibilidad y características físicas del inmueble, junto con variables de estado, estrato y año de construcción. Esta estructura ofrece una aproximación flexible que incorpora la no linealidad observada en las variables continuas, manteniendo una interpretación económica coherente y evitando el sobreajuste.

$$y_i^{(BC)} = \beta_0 + \beta_1\, invarea_i + \beta_2\, invparq_i + \beta_3\, invvias_i + \beta_4\, invcent_i + s_1(edadinv_i) + s_2(dot_i) + s_3(invcol_i) + s_4(invTM_i) + s_5(piso_i) + \beta_5\, anio_i + \beta_6\, nuevo_i + \beta_7\, D_{estr4,i} + \beta_8\, D_{estr5,i} + \beta_9\, D_{estr6,i} + \beta_{10}\, asc_i + \beta_{11}\, term_i + \varepsilon_i$$

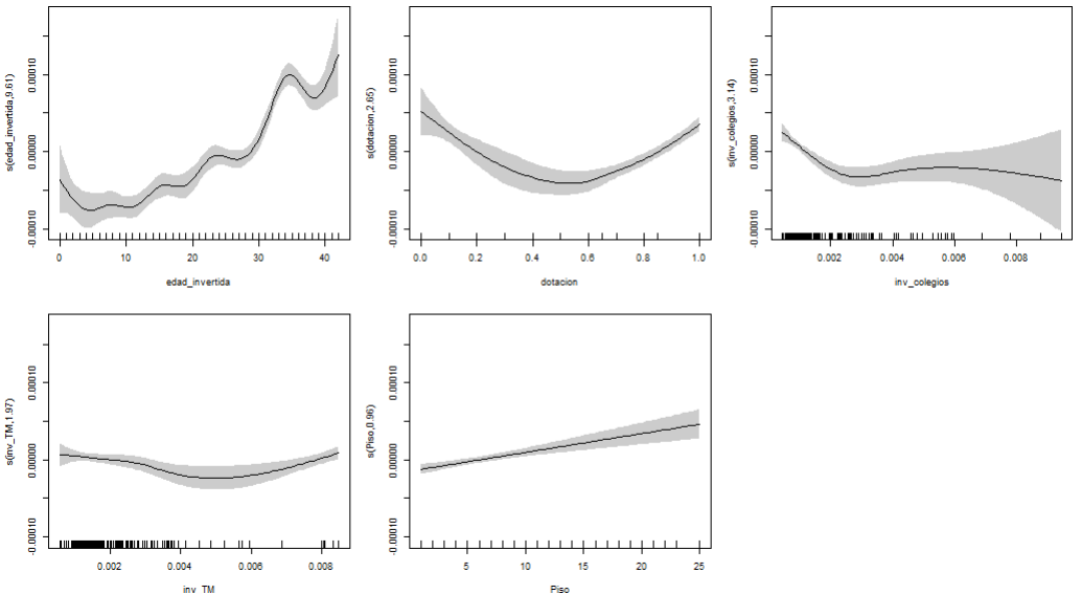


Ilustración 30 Funciones suavizadas de GAM Final

En conjunto, las funciones suavizadas del modelo GAM evidencian relaciones no lineales coherentes con los fundamentos económicos de la valoración inmobiliaria. Las covariables de antigüedad, dotación, accesibilidad educativa y transporte masivo presentan efectos diferenciados que capturan dinámicas de depreciación, equipamiento y localización, mientras que el piso conserva un gradiente vertical positivo con retornos decrecientes. Estas formas suavizadas, con bandas de confianza estrechas y sin oscilaciones espurias, confirman la estabilidad del ajuste y la adecuada selección de la dimensión base. El componente paramétrico complementa la interpretación al reflejar la incidencia esperada de atributos socioeconómicos y de estado del inmueble, consolidando al GAM como una especificación flexible capaz de representar la heterogeneidad del mercado sin pérdida de interpretabilidad.

El chequeo de base no sugiere sobreajuste y confirma que la dimensión k adoptada es suficiente para las curvaturas observadas. Tras retrotransformar a nivel (COP/m^2)el modelo logra MAE ≈ 564 mil COP/m^2 y RMSE ≈ 748 mil COP/m^2 con un coeficiente de determinación promedio de $R^2 = 0.71$. Estas métricas sitúan al GAM como una especificación flexible y competitiva frente a los modelos lineales y robustos, al capturar de manera más eficiente la variación no lineal del mercado. Los residuos mantienen una autocorrelación espacial positiva, lo que respalda la aplicación posterior del esquema de regresión kriging para modelar la dependencia espacial remanente.

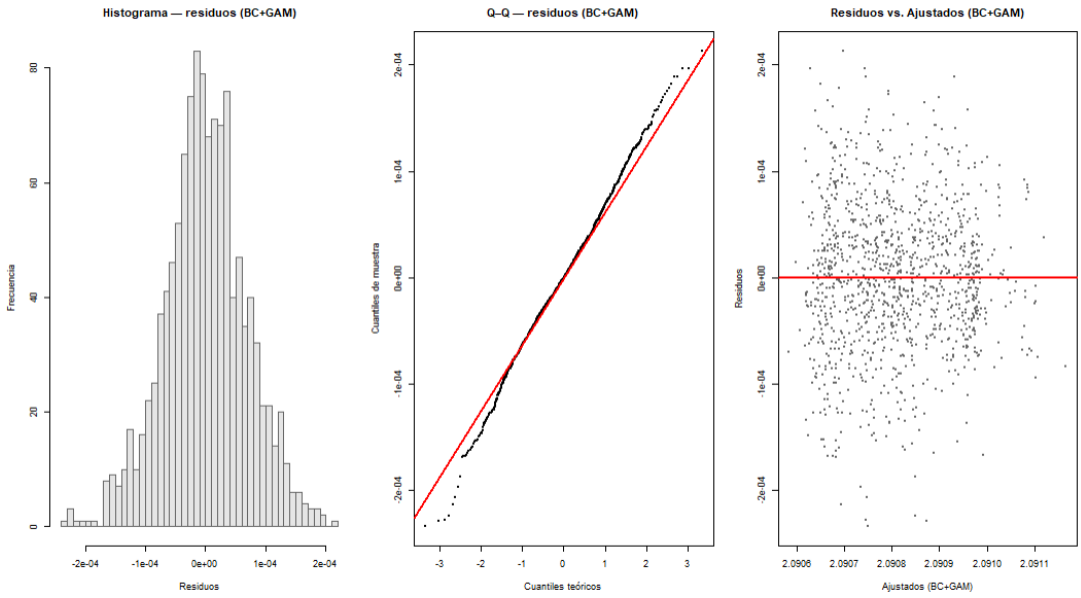


Ilustración 31 Gráficos GAM Final

Los residuos utilizados corresponden al modelo aditivo generalizado (GAM) ajustado sobre la variable transformada mediante Box–Cox, una vez depurados los efectos sistemáticos capturados por los predictores. Dichos residuos fueron extraídos y almacenados como objeto independiente para garantizar su trazabilidad en el proceso geoestadístico. Posteriormente, se alinearon con las filas efectivamente usadas por el modelo, se depuraron observaciones con coordenadas o valores no finitos y se eliminaron duplicados exactos de ubicación para evitar inestabilidades en el ajuste. Adicionalmente, se aplicó un recorte simétrico del 0.045 % en los extremos de la distribución de residuos, con el fin de mitigar la influencia de valores atípicos en la estimación del variograma empírico.

Una vez depurada la información, los residuos obtenidos constituyen el componente estocástico del modelo —representado como $\varepsilon(s)$— el cual recoge la dependencia espacial remanente no explicada por la tendencia sistemática $m(s)$. Bajo este enfoque, el valor observado en una localización s se expresa como la suma de ambos componentes, $Z(s) = m(s) + \varepsilon(s)$. Dado que el modelo GAM elimina la tendencia global y estabiliza la media del proceso, se satisface el supuesto de estacionaridad de segundo orden, al mantener una media aproximadamente constante y una varianza homogénea en el dominio espacial analizado.

Con el fin de validar la estructura de dependencia espacial presente en los residuos del modelo GAM y seleccionar el modelo de variograma más adecuado para el krigeado, se construyó un variograma empírico sobre los residuos transformados mediante Box-Cox. Este variograma empírico se calculó utilizando distancias en proyección EPSG:9377 (metros). En lugar de un cuantil de referencia, la distancia máxima considerada para el cálculo del variograma se fijó como la mitad de la distancia máxima entre observaciones ($cutoff = 0.5 \times dmax$), siguiendo las recomendaciones de estabilidad propuestas para datos distribuidos irregularmente. El espacio de distancias se discretizó en 12 intervalos (bins) con una densidad mínima de 12 pares por clase, lo que permitió mantener una adecuada resolución de la estructura espacial sin comprometer la consistencia estadística de las estimaciones.

En términos geoestadísticos, el nugget (C_0) representa la porción de varianza no explicada por la estructura espacial, asociada a errores de medición, microvariabilidad o discontinuidades a escalas menores que la resolución de muestreo. Por su parte, la sill total

$(C_0 + C_1)$ refleja la varianza global del proceso, es decir, el nivel en el cual la semivarianza se estabiliza y deja de incrementarse con la distancia, indicando la transición hacia la independencia espacial. La relación entre ambos componentes, expresada como $C_0/\ (C_0 + C_1)$, constituye el índice de dependencia espacial (IDE), a partir del cual se evalúa la continuidad del fenómeno. Valores del IDE inferiores a 0.25 indican dependencia fuerte, entre 0.25 y 0.75 dependencia moderada, y superiores a 0.75 dependencia débil.

Sobre este variograma empírico se ajustaron mediante mínimos cuadrados ponderados (WLS) tres modelos teóricos candidatos: esférico (Sph), exponencial (Exp) y Gaussiano (Gau), El criterio de selección entre modelos se fundamentó en la minimización de la suma ponderada de errores cuadráticos (SSE_WLS), que cuantifica la discrepancia entre la semivarianza empírica observada y la semivarianza modelada, ponderando adecuadamente los intervalos de distancia (bins) por su número de pares para reflejar la confiabilidad relativa de cada estimación.

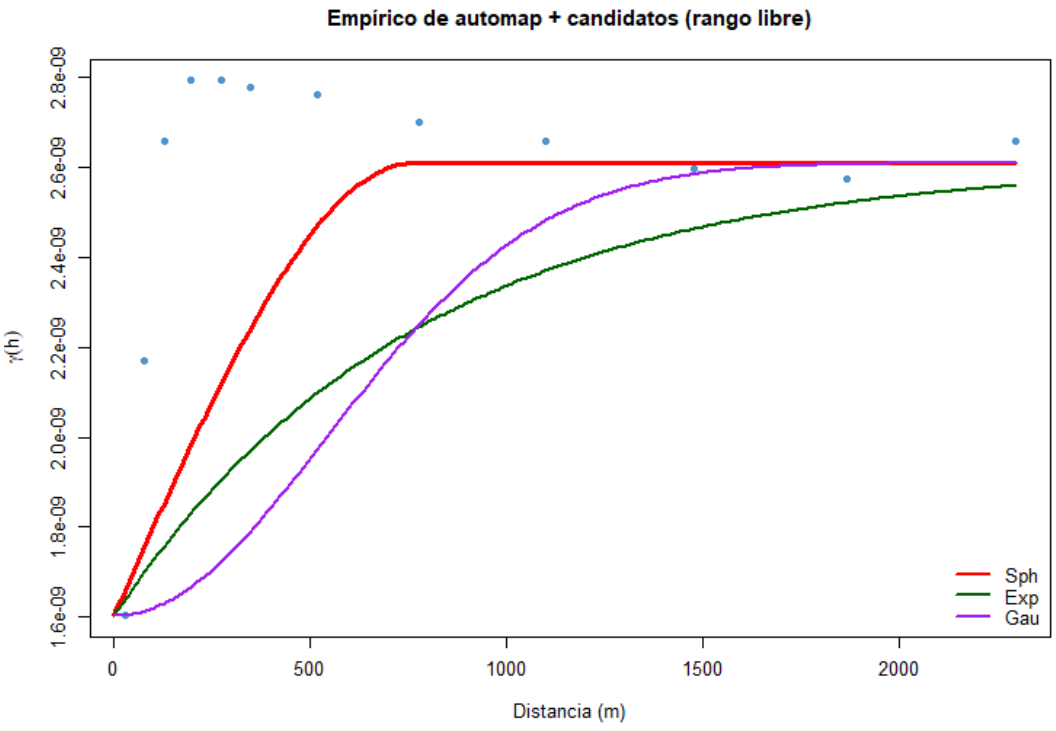


Ilustración 32 Variograma Empírico

El ajuste de candidatos a modelos teóricos del variograma reveló diferencias en la capacidad de cada especificación para reproducir fielmente la estructura empírica observada.

El modelo exponencial se caracteriza por una aproximación gradual y asintótica a la meseta, resultando en subestimación de la semivarianza en distancias intermedias. El modelo Gaussiano, por su parte, exhibe una curvatura pronunciada cercana al origen, ocasionando sobreestimación inicial de la dependencia espacial. En contraste, el modelo esférico alcanza la meseta de forma rápida y definida, minimizando el error ponderado de mínimos cuadrados ponderados respecto a las observaciones empíricas, demostrando así la mayor fidelidad a los datos observados.

La selección del modelo esférico se fundamenta en su capacidad para reproducir con precisión la continuidad espacial observada en los datos, su delimitación clara del rango de dependencia, y su conformidad con el principio de parsimonia y fidelidad empírica. Esta especificación garantiza que la interpolación espacial mediante kriging ordinario de residuos opera bajo un modelo teórico robusto y estadísticamente validado.

El análisis de isotropía espacial se ejecutó bajo la generación de variogramas direccionales, evaluados en cuatro azimutes o direcciones (0°, 45°, 90° y 135°), esto proporciona evidencia empírica sobre la estructura anisotrópica de la dependencia espacial en los residuos del modelo GAM. La interpretación conjunta de estos cuatro paneles revela el comportamiento direccional de la varianza residual en función de la orientación espacial.

Los cuatro variogramas direccionales presentan comportamientos notablemente similares, con nubes de puntos dispersas alrededor de niveles de semivarianza comparables para distancias equivalentes. En particular, no se observan diferencias sistemáticas marcadas entre direcciones: los puntos empíricos en 0° (norte-sur), 45° (diagonal), 90° (este-oeste) y 135° (diagonal inversa) convergen aproximadamente al mismo rango (aproximadamente 500-700 metros en las mesetas observadas) sin mostrar elongación o compresión preferencial en ninguna orientación específica. Esta convergencia direccional indica isotropía espacial, es decir, la estructura de dependencia es invariante a la dirección de búsqueda, confirmando que la continuidad espacial de los precios no está orientada preferentemente hacia ningúna dirección particular. Esto es particularmente relevante en contextos urbanos inmobiliarios donde, aunque existen patrones de zonificación y estructura urbana, la dependencia de precios por proximidad no presenta sesgos direccionales pronunciados: el precio de un

inmueble depende de su cercanía a servicios, no de si esa cercanía ocurre hacia el norte, sur, este u oeste.

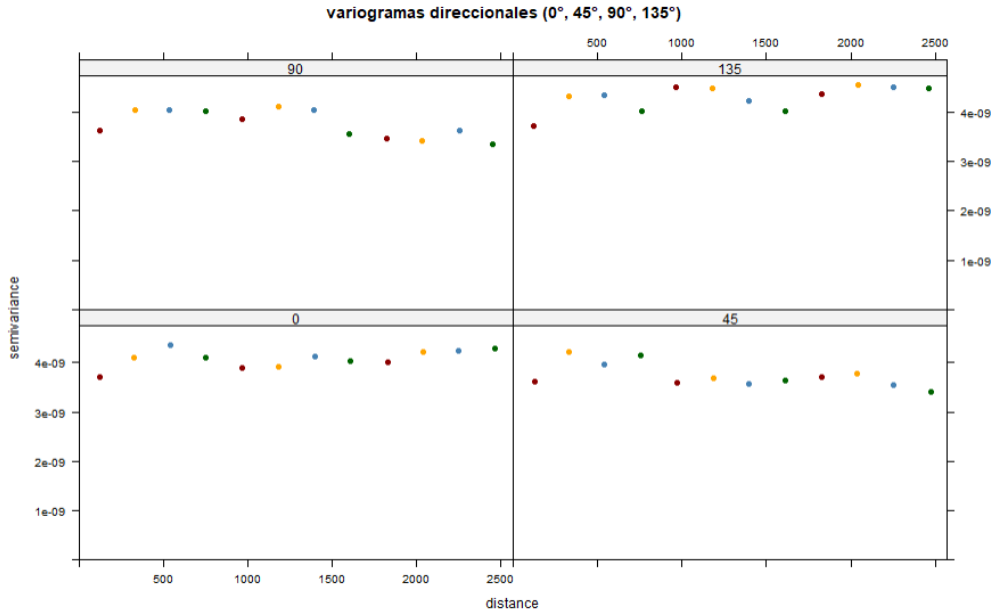


Ilustración 33 Variograma direccional - Isotropía

Finalmente, el variograma ajustado corresponde a un modelo Gaussiano, $C_0 = 1.30 \times 10^{-9}$, $C_1 = 1.67 \times 10^{-9}$, $C_0 + C_1 = 2.96 \times 10^{-9}$, lo que implica un IDE inferior a 0.5 y, por tanto, una dependencia espacial moderadamente fuerte, donde la mayor parte de la variabilidad se explica por la estructura espacial modelada y no por el ruido aleatorio. El rango efectivo $a \approx 168.30$ metros representa la distancia límite hasta la cual los residuos mantienen autocorrelación positiva lo que es lógico con el tamaño de las manzanas catastrales en la zona; más allá de este alcance, los valores pueden considerarse independientes.

Esta elección es coherente con la naturaleza del fenómeno inmobiliario estudiado, sugiere que los precios de inmuebles ubicados a menos de la distancia definida por el rango presentan correlación espacial significativa, posiblemente reflejando efectos de microzonas con características homogéneas o mercados con composición demográfica similar. Más allá de este umbral, los residuos se comportan como espacialmente independientes, indicando que la heterogeneidad de localización ya no está estructurada por proximidad geográfica sino por factores específicos de localización no capturados por los predictores del GAM. El variograma esférico, al capturar esta transición con mayor precisión, proporciona la base

geoestadística óptima para la interpolación de residuos y la producción de superficies continuas de valoración espacialmente consistentes.

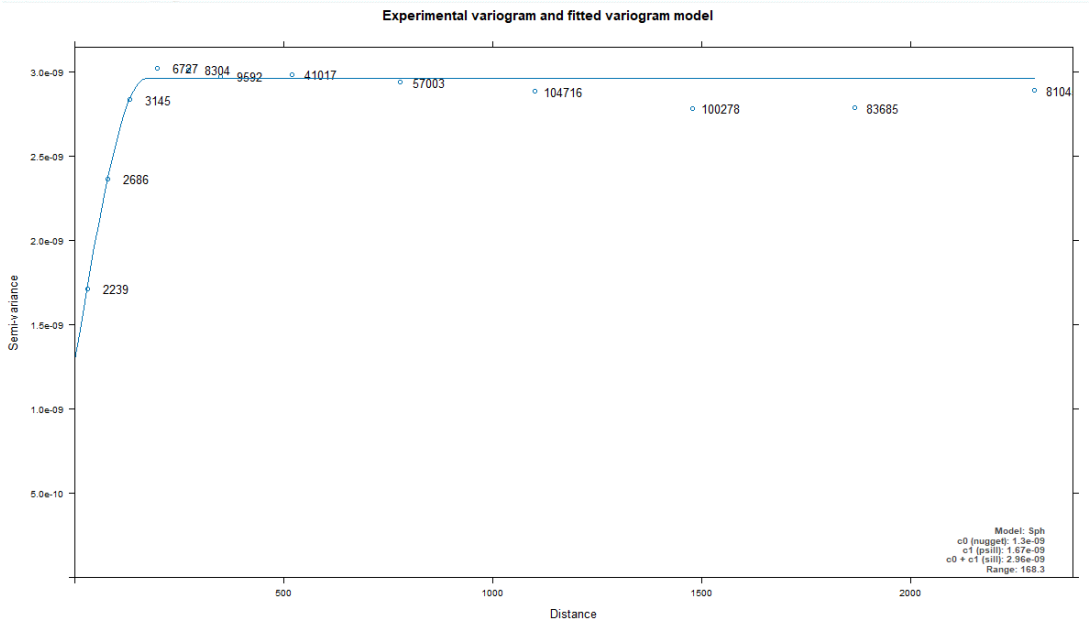


Ilustración 34 Variograma

Con el variograma ajustado, se procedió a la interpolación de los residuos mediante kriging ordinario sobre una grilla proyectada en EPSG:9377. Este proceso utilizó el modelo teórico de semi varianza $\gamma(h)$ definido previamente.

El mapa de kriging de residuos (izquierda) muestra la superficie interpolada en la escala Box–Cox, la cual representa la componente estocástica del modelo luego de remover la tendencia explicada por el GAM. Los valores interpolados no corresponden directamente a unidades monetarias, sino a residuos transformados que reflejan los ajustes locales que el modelo geoestadístico introduce sobre la predicción base. Su distribución espacial evidencia parches de sobre y subestimaciones coherentes con el alcance estimado (168.30 m), lo que confirma la existencia de continuidad espacial moderada y una estructura de dependencia bien representada por el modelo $\gamma(h)$.

El mapa de varianza de kriging (derecha) expresa la incertidumbre o error de predicción (σ_k^2) asociado a la estimación en cada punto. Las zonas con mayor densidad de datos presentan varianza reducida —indicando estimaciones más confiables—, mientras que los valores altos se concentran hacia los bordes o en áreas con menor número de

observaciones, donde el error aumenta por la falta de información vecina. Este patrón es coherente con el principio de vecindad del kriging y con la estructura del variograma ajustado.

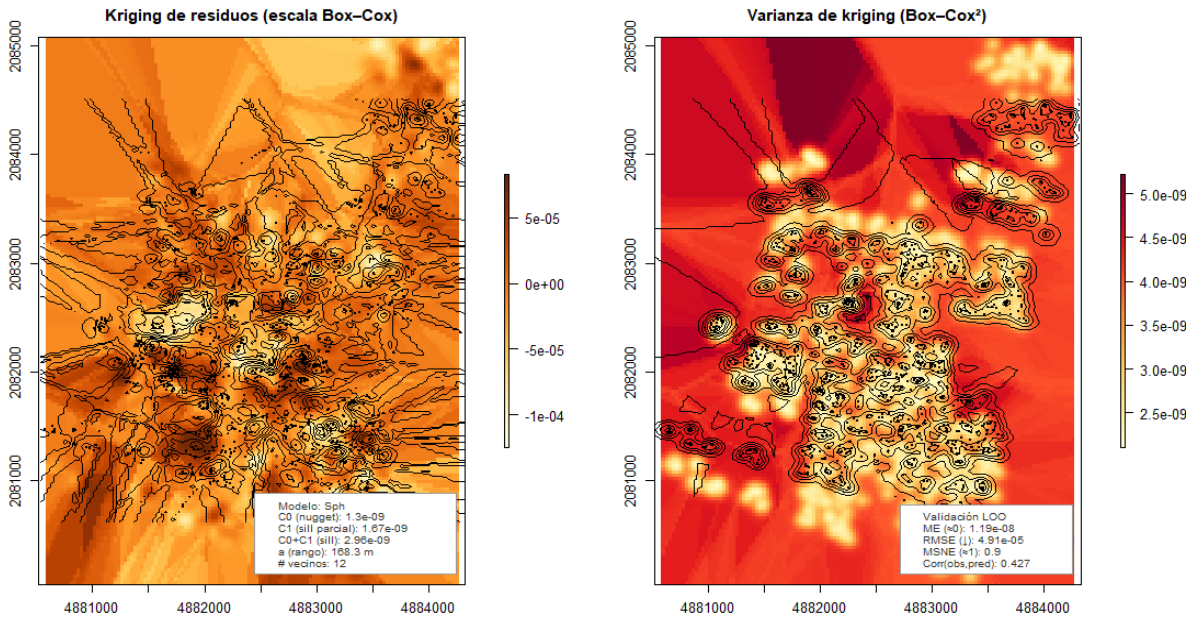


Ilustración 35 Kriging de Residuos y Varianza de Kriging

Con el fin de evaluar la calidad del modelo geoestadístico, se aplicó una validación cruzada tipo leave-one-out (LOO) sobre los residuos transformados. Este procedimiento consiste en excluir iterativamente cada observación, estimar su valor mediante kriging con las restantes y comparar la predicción con el valor observado. A partir de los resultados obtenidos se calcularon indicadores de ajuste y consistencia estadística del modelo.

El error medio (ME) fue de 1.19×10^{-8}, valor cercano a cero que indica ausencia de sesgo sistemático en las estimaciones. El error cuadrático medio (RMSE) alcanzó 4.91×10^{-5}, lo que evidencia una adecuada precisión del modelo dentro de la escala Box–Cox. El estadístico MSNE (media de los residuos estandarizados al cuadrado) resultó igual a 0.896 se encuentra próximo a la unidad, lo que sugiere una calibración adecuada de la varianza de kriging frente a los errores observados, aunque con una ligera tendencia a la sobreestimación de la varianza (valores menores a 1 indican predicciones algo más conservadoras). Asimismo, la desviación estándar de los z-scores ($sd(z) = 0.947$) confirma que la dispersión de los errores se ajusta a lo esperado bajo normalidad estándar, mientras que la cobertura empírica del 95.4 % para

el intervalo |z| ≤ 1. muestra una correspondencia prácticamente perfecta con el nivel teórico del 95 %, lo que refuerza la consistencia del modelo. El coeficiente de correlación entre valores observados y predichos ($R = 0.43$; $R^2 = 0.18$) muestra una relación moderada, suficiente para sustentar la validez del modelo de variograma y la continuidad espacial estimada.

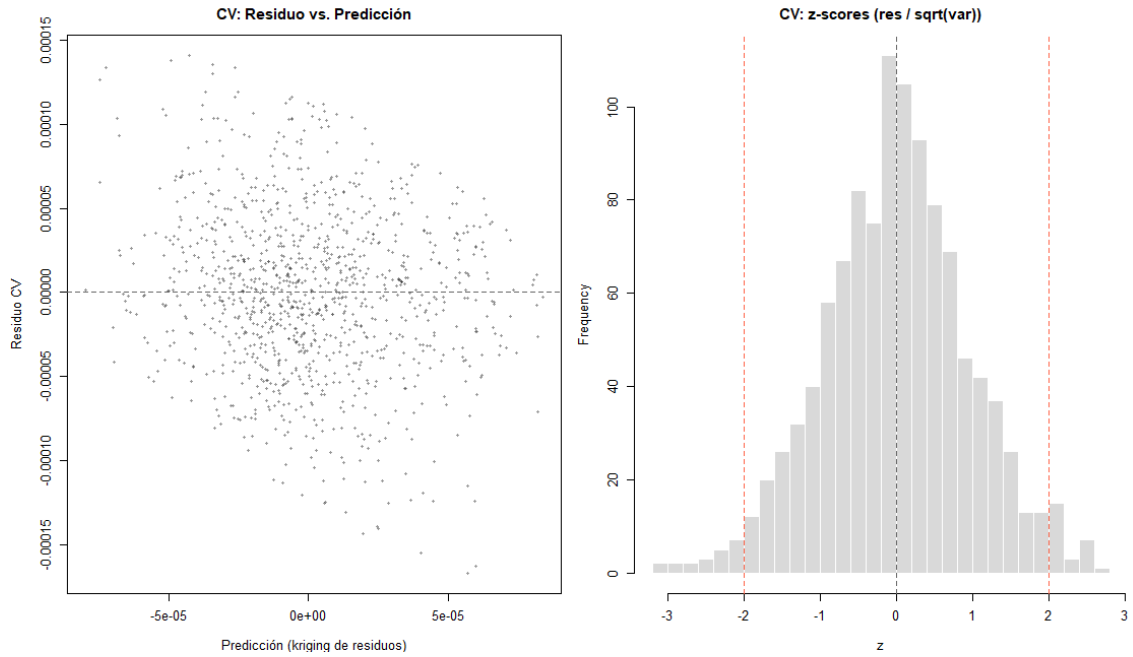


Ilustración 36Ilustración Validación cruzada Leave One Out

El proceso de interpolación final se desarrolló a partir de los resultados del modelo aditivo generalizado (GAM) ajustado en la escala transformada mediante Box–Cox. A partir de la base de datos validada, se identificaron las coordenadas de cada observación y se reproyectaron al sistema MAGNA–SIRGAS Origen Nacional (EPSG:9377), con el fin de garantizar unidades métricas homogéneas y consistencia espacial. Se eliminaron duplicados de ubicación y se generó una grilla regular con resolución de 40 m y un buffer de 200 m alrededor del dominio muestral, lo que permitió extender el cálculo de las predicciones y suavizar posibles discontinuidades en los bordes del área de estudio.

Para cada celda de la grilla se reconstruyeron las variables explicativas requeridas por el modelo, con el único propósito de permitir la evaluación del GAM en todo el dominio espacial. Dado que en la mayoría de las celdas no existen observaciones directas de las covariables, se aproximaron únicamente los valores faltantes a partir de la información disponible. El modelo GAM requiere contar con todas las variables de entrada en cada punto donde se desea predecir, ya que cada término

suavizado o paramétrico depende de dichas entradas. Por esta razón, las covariables numéricas se estimaron mediante distancia inversa ponderada (p=2) y las categóricas por vecino más cercano, preservando los niveles definidos en el ajuste original. Este procedimiento no forma parte del proceso de interpolación geoestadística ni afecta la estimación por kriging, ya que únicamente provee los valores auxiliares necesarios para calcular la tendencia $\widehat{m}_{BC}(s)$ en cada punto. Con las covariables reconstruidas se evaluó el modelo GAM en toda la grilla, obteniendo la superficie continua de tendencia $\widehat{m}_{BC}(s)$, correspondiente a la parte del valor explicada por los atributos físicos y de entorno del inmueble.

La predicción total en la escala transformada se calculó como la suma de la componente determinística y la estocástica:

$$Z_{BC}(s) = \widehat{m_{BC}}(s) + \widehat{\varepsilon_{BC}}(s)$$

Posteriormente, se aplicó la inversa de la transformación Box–Cox (λ=−0.478) con el fin de expresar las predicciones en unidades monetarias (COP/m²). Esta operación constituye únicamente un cambio de escala y no altera la estructura espacial ni las relaciones relativas entre valores, ya que la función inversa de Box–Cox es monótona y preserva el orden de las predicciones. En consecuencia, las áreas con valores altos o bajos en la escala transformada mantienen la misma jerarquía tras la reexpresión, permitiendo interpretar la superficie final directamente en términos económicos.

$$\hat{Z}(s) = [\,\lambda Z_{BC}(s) + 1\,]^{1/\lambda}$$

El resultado es una superficie continua de valor por metro cuadrado (COP/m²) que integra simultáneamente la estructura sistemática del GAM y la dependencia espacial modelada mediante kriging. Que finalmente se ilustra de la siguiente manera:

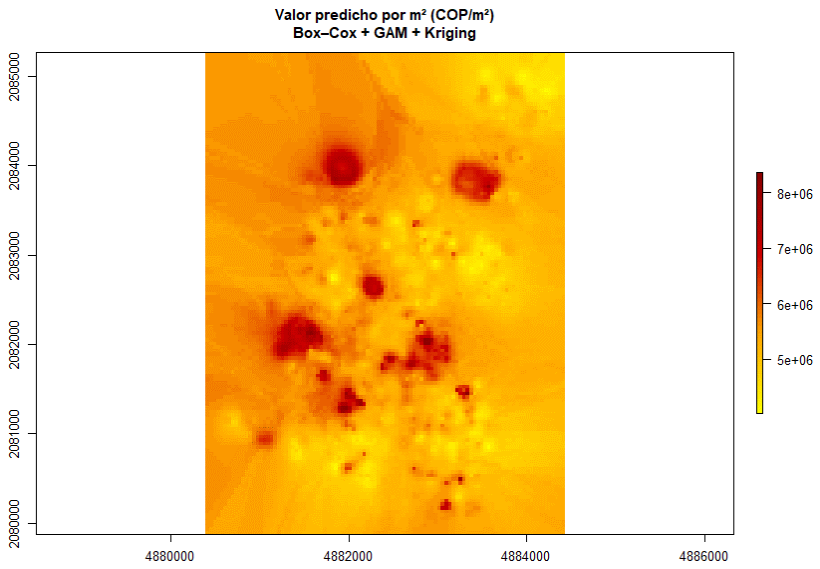


Ilustración 37 Predicción valor por metro cuadrado final

5.6.4. Generación de mapas temáticos y análisis espacial de resultados.

Para estandarizar la lectura cartográfica, los valores predichos en COP/m^2 se redondearon al múltiplo de $10.000 más cercano y se presentaron con formato monetario colombiano en leyendas y etiquetas. Esto garantiza clases "cerradas" y comparables en el mapa (evita colas de valores muy finos) y mejora la comunicación: el lector ve cifras en $ COP/m^2 con separadores de miles y, cuando conviene, en millones con dos decimales.

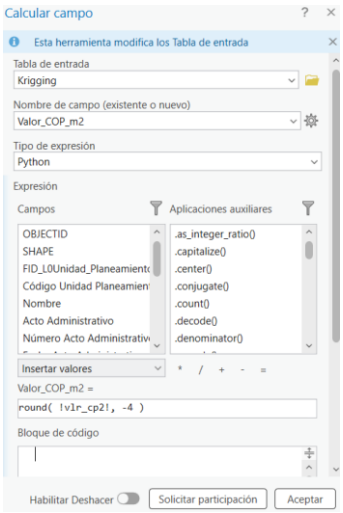


Ilustración 38 Redondeo de Cifras

Para la visualización final se elaboró un mapa de colores graduados usando como campo temático Valor_COP_m2, obtenido de la predicción final y recortado con la capa de manzanas catastrales para ajustar la superficie al límite urbano, eliminar celdas fuera del ámbito y mejorar la legibilidad. La clasificación se hizo con Rupturas naturales (Jenks) en 7 clases, y se aplicó una rampa amarillo→rojo (claros = valores bajos; rojos = altos). Las etiquetas se formatearon con símbolo $ y separadores de miles, tras redondear los valores al múltiplo de $10.000 más cercano. Con esta simbología, el mapa muestra un rango aproximado de $4.010.000 a $8.370.000 por m² en el área de estudio, ya alineado a la morfología de las manzanas y listo para la salida cartográfica.

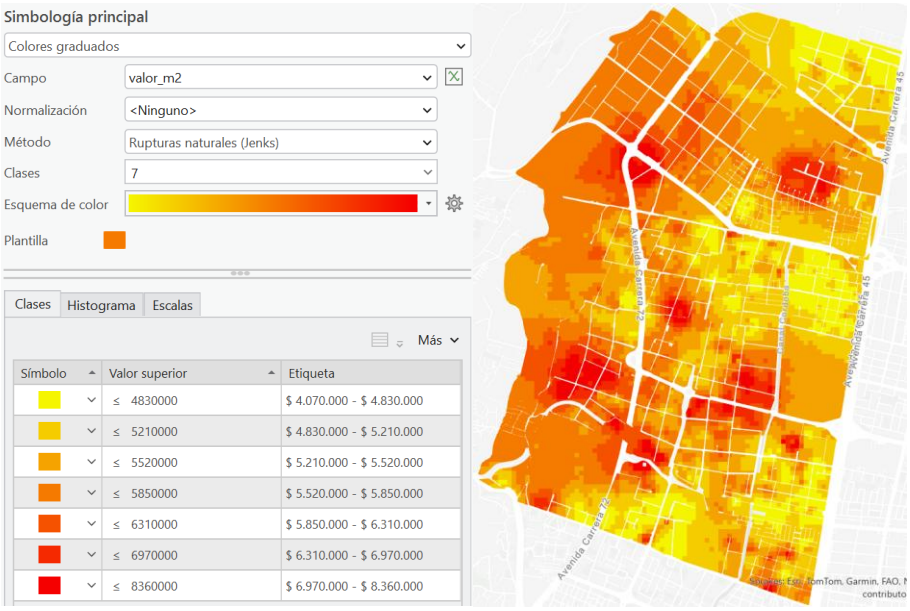


Ilustración 39 Simbología y Corte

Para la salida cartográfica final se compuso un rótulo en un diseño tipo layout que garantiza lectura y trazabilidad: se fijó un título claro (“Valor de metro cuadrado de área privada”), leyenda con 7 clases en COP/m^2 (formateadas con símbolo $ y separadores de miles) correspondiente al mapa de colores graduados, barra de escala (1:30.000) y flecha norte. Se añadieron mapas de ubicación (Colombia y Bogotá D.C) para contexto espacial y un bloque de metadatos con sistema de referencia MAGNA–SIRGAS / Bogotá (EPSG:9377), parámetros de proyección, fuente de datos base (OpenStreetMap/Esri) y créditos institucionales (Universidad Distrital – Facultad de Ingeniería).

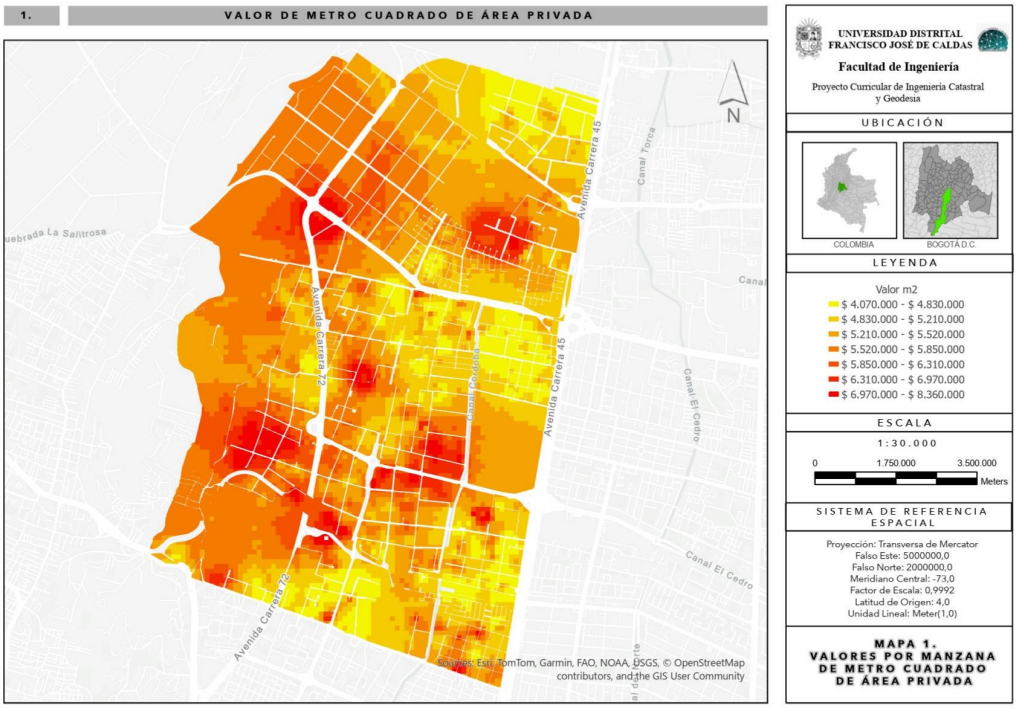


Ilustración 40 Salida Gráfica Final

6. RESULTADOS

El mapa de valor unitario resultado de la regression-kriging muestra una superficie continua con rangos entre $4,07 y $8,37 millones COP/m^2 (clases definidas con rupturas naturales. Las zonas de mayor valor (tonos rojo-rojizos, superiores a $6,1 millones COP/m²) se concentran principalmente a lo largo de los corredores de la Avenida Carrera 72 y la Avenida Calle 170, extendiéndose hacia los sectores de Mazurén, Iragua, San José de Bavaria y parte de Gilmar. Estos sectores presentan altos niveles de consolidación residencial y dotacional, caracterizados por conjuntos cerrados de estratos 5 y 6, edificaciones recientes con buena calidad constructiva, presencia de comercio especializado, instituciones educativas privadas de alto nivel y una red vial estructurante que garantiza accesibilidad metropolitana. Este conjunto de factores consolida un entorno urbano de alta demanda, donde el valor del suelo refleja tanto la ubicación estratégica como la percepción de calidad ambiental y seguridad. La presencia de grandes equipamientos —como universidades, colegios, centros religiosos y supermercados ancla— contribuye a sostener la valorización del entorno, reforzando la concentración de precios en torno a estos núcleos.

En contraste, las zonas de valor medio (entre $5,1 y $6,1 millones COP/m², tonos anaranjados) corresponden a sectores consolidados de uso predominantemente residencial con mezcla funcional moderada. Áreas como Britalia y Cantagallo presentan una morfología urbana regular y homogénea, con predominio de conjuntos multifamiliares de estratos 4 y 5, cercanía a equipamientos de escala zonal (colegios distritales, parques, centros comerciales) y accesos directos a corredores secundarios como la Carrera 55 o la Calle 163.

Por su parte, las zonas de menor valor (tonos amarillos, inferiores a $4,6 millones COP/m²) se ubican hacia los bordes noroccidentales y surorientales de la UPL, especialmente en sectores como Granada Norte, Britalia Alta y áreas periféricas próximas a la Carrera 9 o los límites de Suba Cerros en donde se colinda con área rural de Bogotá. En estos sectores el tejido urbano se vuelve más discontinuo, con presencia de predios de mayor tamaño, baja densidad, edificaciones antiguas, propiedades horizontales constituidas sobre casas comunes sin dotación comunal y menor integración al sistema vial estructurante. La oferta de servicios urbanos es más dispersa y los equipamientos de impacto metropolitano escasean, lo que

reduce la capacidad de capitalización del suelo y explica los niveles inferiores de valor predicho. Estos resultados coinciden con la lógica de gradiente de centralidad decreciente, donde la distancia a los corredores principales y la menor dotación urbana implican una disminución progresiva del precio unitario.

En conjunto, la superficie estimada describe un patrón de valorización coherente con el modelo urbano de Bogotá, en el que los corredores estructurantes actúan como ejes de concentración del valor y las áreas residenciales intermedias se estabilizan como zonas de transición entre mercados de alta y media densidad económica. La correspondencia entre los valores predichos y la distribución real de equipamientos, vías arteriales y usos del suelo confirma la capacidad del modelo para reproducir las condiciones espaciales del mercado inmobiliario. Además, la existencia de núcleos locales de alta homogeneidad sugiere la presencia de microzonas funcionales, donde las características del entorno inmediato —accesibilidad, calidad urbanística y percepción socioespacial— definen la estructura interna del valor en la UPL Britalia y su entorno.

7. ALCANCES

Los resultados obtenidos —producto de la integración entre el modelo aditivo generalizado (GAM) y el componente geoestadístico por regresión-kriging— cumplen el objetivo de estimar y representar espacialmente el valor unitario de los apartamentos en la UPL Britalia, generando una lectura coherente y diferenciada a escala intra-barrio. Las superficies continuas y los mapas en COP/m^2 permiten identificar gradientes espaciales, microzonas de alta y baja valorización, y contrastes entre sectores homogéneos y transicionales, proporcionando una base cuantitativa para el análisis territorial del mercado inmobiliario.

En términos operativos, los productos desarrollados facilitan la interpretación del comportamiento espacial del valor del suelo y la detección de inconsistencias locales: permiten realizar consultas rápidas por manzana o sector, comparar rangos de valor dentro de la UPL y priorizar revisiones de avalúos donde el residuo espacial indica posibles desalineaciones del mercado. Su nivel de detalle los hace útiles tanto para tareas de control de calidad y depuración de bases masivas, como para procesos de planeación urbana, gestión catastral y análisis comercial, al vincular coherentemente la distribución de precios con variables de accesibilidad, dotación y consolidación urbana.

El modelo es reutilizable, trazable y actualizable, ya que la metodología de integración datos–modelo–mapa es replicable en otras UPL que cuenten con información equivalente (atributos físicos, de entorno y coordenadas precisas). Una nueva actualización de la base o del mercado permite reentrenar el GAM y regenerar la superficie de kriging, posibilitando monitoreos temporales de la dinámica de precios.

La escala de decisión natural de los resultados corresponde a la manzana y al sector urbano, niveles que garantizan estabilidad estadística y aplicabilidad práctica. Estos insumos pueden ser empleados como referencia de precios por zona, apoyo en valoraciones puntuales o masivas, y herramienta para analizar impactos de obras públicas, cambios normativos o procesos de renovación urbana.

En síntesis, el proyecto entrega una línea base cartográfica y cuantitativa de alta precisión, que mejora la trazabilidad de los criterios de valoración, reduce los tiempos de análisis y favorece la transparencia técnica en la toma de decisiones. Su validez se mantiene dentro del dominio espacial y temporal observado —la UPL y el periodo de estudio— y se sustenta en buenas prácticas de modelación espacial, incluyendo la validación cruzada, el control de autocorrelación residual y la consistencia de unidades monetarias en COP/m^2.

8. CONCLUSIONES

La integración del modelo aditivo generalizado (GAM) con la interpolación geoestadística por regression–kriging demostró ser una estrategia robusta y equilibrada para estimar y representar el valor unitario de apartamentos en COP/m^2 a escala intra-UPL. Esta combinación permitió capturar simultáneamente las relaciones no lineales entre las variables explicativas y la estructura de dependencia espacial remanente en los residuos, logrando un modelo flexible, interpretable y estadísticamente consistente.

La transformación Box–Cox contribuyó a estabilizar la varianza y mejorar el ajuste del GAM, garantizando que los supuestos de homocedasticidad y normalidad fueran más plausibles. La validación cruzada espacial evitó la sobreestimación de desempeño por filtración, produciendo márgenes de error creíbles y métricas fuera de muestra representativas del comportamiento real del modelo al predecir nuevas manzanas.

En la comparación entre familias de modelos, Random Forest alcanzó la mayor precisión bruta dentro de la muestra, pero mostró autocorrelación espacial no significativa en los residuos (Moran's I), lo que limitó su utilidad para la etapa geoestadística. Por el contrario, el GAM logró un desempeño similar en error medio, explicó de manera adecuada las no linealidades asociadas al tamaño, antigüedad, dotación y accesibilidad, y conservó una autocorrelación espacial positiva en los residuos. Esta dependencia espacial justificó su elección como modelo base para el kriging, al proveer la señal necesaria para capturar patrones locales de variación del mercado.

El variograma ajustado de tipo esférico, con un rango aproximado de 168.30 metros y un nugget apreciable, evidencia que el mercado inmobiliario en la UPL Britalia presenta una correlación de corto alcance, donde los cambios de precio tienden a disiparse en unas pocas cuadras. Este comportamiento es coherente con la estructura urbana y con la escala de homogeneidad de los submercados locales. El mapa de varianza de kriging permitió identificar zonas con menor densidad de comparables, señalando áreas donde las predicciones deben interpretarse con precaución o priorizarse para futuras campañas de captura de datos.

Desde el punto de vista aplicado, el modelo ofrece una herramienta útil para referenciar valores por zona, respaldar avalúos individuales y monitorear cambios en el mercado mediante su reentrenamiento con nuevas observaciones. Además, proporciona soporte a la gestión y planificación urbana, al posibilitar la comparación de gradientes de valor con la accesibilidad, las dotaciones y la morfología del entorno.

Existen, sin embargo, limitaciones inherentes al modelo y a los datos empleados. La capacidad predictiva depende de la densidad y cobertura espacial de las observaciones, lo que restringe la extrapolación fuera del dominio muestral. Asimismo, factores no observados, como el estado interior de las unidades, la orientación o la calidad de los acabados, podrían explicar parte del nugget identificado. No obstante, la aplicación de un protocolo riguroso de validación cruzada espacial, diagnóstico de autocorrelación y evaluación de métricas en nivel asegura que los resultados sean defendibles, reproducibles y operativamente útiles dentro del ámbito analizado.

9. REFERENCIAS

1. Alcaldía Mayor de Bogotá D.C. (2021). Decreto 555 de 2021: Por el cual se adopta la revisión general del Plan de Ordenamiento Territorial – POT Bogotá Reverdece 2022–2035. Secretaría Distrital de Planeación. Recuperado de https://sdp.gov.co/gestion-territorial/plan-de-ordenamiento-territorial-pot
2. Alcaldía Mayor de Bogotá D.C. (2023). Decreto 557 de 2023: Por el cual se reglamenta el Sector Norte y se adoptan las Unidades de Planeamiento Local (UPL) Britalia, Torca, Usaquén, Niza y Toberín. Secretaría Distrital de Planeación. Recuperado de https://sistemas002.sdp.gov.co/micrositios/pot/reglamentacion/decreto/decreto-557-de-2023-upl-sector-norte
3. Alcaldía Mayor de Bogotá, D. C. (2017, 3 de marzo). Decreto Distrital 088 de 2017: Por medio del cual se establecen las normas para el ámbito de aplicación del Plan de Ordenamiento Zonal del Norte – "Ciudad Lagos de Torca" (POZ Norte). SISJUR – Sistema de Información Jurídica del Distrito. https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=68421
4. Antipov, E. A., & Pokryshevskaya, E. B. (2012). Mass appraisal of residential apartments: An application of Random forest for valuation and a CART-based approach for model diagnostics. *Expert Systems with Applications, 39*(2), 1772-1778. https://doi.org/10.1016/j.eswa.2011.08.077 Disponible en: https://www.sciencedirect.com/science/article/pii/S0957417411011894
5. Arbia, G. (2016). Spatial econometrics: A broad view. Foundations and Trends® in Econometrics, 8(3–4), 145–265. https://www.nowpublishers.com/article/Details/ECO-030
6. Armijos-Toro, J., González, J., & Sarmiento, M. (2022). *Estimation of housing prices in Ecuador using hedonic and geostatistical methods*. Revista de Métodos Cuantitativos para la Economía y la Empresa, 33(2), 45–61. RedALyC. https://www.redalyc.org/journal/5736/573670796007/html/
7. Baisad, K., Chutsagulprom, N., & Moonchai, S. (2023). A Non-Linear Trend Function for Kriging with External Drift Using Least Squares Support Vector Machines. Mathematics, 11(23), 5108.
8. Bajari, P., & Kahn, M. E. (2005). Estimating housing demand with an application to explaining racial segregation in cities. *Journal of Business & Economic Statistics, 23*(1), 20-28. https://doi.org/10.1198/073500104000000334
9. Banco de la República. (2023). Informe especial de estabilidad financiera: Análisis de la cartera y del mercado inmobiliario en Colombia (I semestre 2023). Banco de la República. https://repositorio.banrep.gov.co/bitstream/handle/20.500.12134/10663/Informe_Especial_vivienda_2023-I.pdf
10. Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2015). *Hierarchical Modeling and Analysis for Spatial Data* (2nd ed.). CRC Press. https://doi.org/10.1201/b18566

11. Bivand, R. S., Pebesma, E., & Gómez-Rubio, V. (2013). Applied Spatial Data Analysis with R (2nd ed.). Springer. Enlace: Applied Spatial Data Analysis with R+1
12. Bourdais, T., Batlle, P., Yang, X., Baptista, R., Rouquette, N., & Owhadi, H. (2024). A Gaussian Process framework for connecting the dots. Proceedings of the National Academy of Sciences, 121(32), e2403449121. https://doi.org/10.1073/pnas.2403449121
13. Brueckner, J. K., Kahn, M. E., & Lin, G. C. (2021). A new spatial hedonic equilibrium in the emerging work-from-home economy (NBER Working Paper No. 28526).
14. Castrignanò, A. (2011). *Spatial Data Analysis for Environmental and Agricultural Applications: A Practical Guide.* Springer. https://doi.org/10.1007/978-94-007-4170-7
15. Cellmer, R., Bełej, M., & Konowalczuk, J. (2020). Spatial analysis of housing prices and market activity with the geographically weighted regression. ISPRS International Journal of Geo-Information, 9(6), 380. https://www.mdpi.com/2220-9964/9/6/380
16. Chica-Olmo, J., Cano-Guervos, R., & Chica-Rivas, M. (2019). Estimation of housing price variations using spatio-temporal data. *Sustainability, 11*(6), 1551. https://www.mdpi.com/2071-1050/11/6/1551
17. Cressie, N. (1993). *Statistics for Spatial Data* (Revised ed.). John Wiley & Sons. https://doi.org/10.1002/9781119115151
18. Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2016). Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. Journal of the American Statistical Association, 111(514), 800–812. https://doi.org/10.1080/01621459.2015.1044091
19. Derdouri, A. (2020). *A comparative study of land price estimation and mapping using hedonic modeling, kriging, and random forest methods*. Journal of Applied Remote Sensing, 14(5), 050007. https://doi.org/10.1117/1.JRS.14.050007
20. Diao, M., Leonard, D., & Sing, T. F. (2017). Spatial-difference-in-differences models for impact of new mass rapid transit line on private housing values. *Regional Science and Urban Economics, 67*, 64- https://doi.org/10.1016/j.regsciurbeco.2017.08.006
21. Dou, M., Gu, Y., & Fan, H. (2023). Incorporating neighborhoods with explainable artificial intelligence for modeling fine-scale housing prices. Applied Geography, 158, 103032. https://doi.org/10.1016/j.apgeog.2023.103032
22. Elhorst, J. P. (2010). Applied spatial econometrics: Raising the bar. Spatial Economic Analysis, 5(1), 9–28. Enlace: Portal de Investigación UGroningen+1
23. Elhorst, J. P. (2017). Spatial panel data analysis. In S. Shekhar, H. Xiong, & X. Zhou (Eds.), Encyclopedia of GIS (2nd ed., pp. 2050–2058). Springer. https://spatial-panels.com/wp-content/uploads/2017/07/Elhorst-Spatial-Panel-Data-Analysis-Encyclopedia-GIS-2nd-ed_Working-Paper-Version.pdf
24. Elhorst, J. P. (2022). The dynamic general nesting spatial econometric model for spatial panels with common factors: Further raising the bar. Review of Regional Research, 42, 249–267. https://link.springer.com/article/10.1007/s10037-021-00163-w

25. Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2002). Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley. Enlace: Wiley+1
26. Getis, A., & Aldstadt, J. (2004). Constructing the spatial weights matrix using a local statistic. Geographical Analysis, 36(2), 90–104. Enlace: Wiley Online Library
27. Gibbons, S., & Machin, S. (2005). Valuing rail access using transport innovations. *Journal of Urban Economics, 57*(1), 148-169. https://doi.org/10.1016/j.jue.2004.10.002
28. Glaeser, E. L., & Gyourko, J. (2018). The economic implications of housing supply. Journal of Economic Perspectives, 32(1), 3–30. https://doi.org/10.1257/jep.32.1.3
29. Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat. The R Journal, 8(1), 204–218. https://doi.org/10.32614/RJ-2016-014
30. Goodchild, M. F. (1986). Spatial Autocorrelation (CATMOG 47). Norwich, UK: Geo Books.
31. Goodman, A. C., & Thibodeau, T. G. (2003). Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics, 12*(3), 181-201. https://doi.org/10.1016/S1051-1377(03)00031-7
Disponible en: https://www.sciencedirect.com/science/article/abs/pii/S1051137703000317
32. Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation.* Oxford University Press.
33. GraphPad. (2024). Choosing a normality test. Statistics with Prism 10. https://www.graphpad.com/guides/prism/latest/statistics/stat_choosing_a_normality_test.htm GraphPad
34. Grigsby, W., & Rosenburg, L. (1975). Urban housing policy. New York: APS Publications.
https://www.taylorfrancis.com/books/mono/10.4324/9781351300568/urban-housing-policy-william-grigsby
35. Han, F., Lu, M., Qin, D., Zheng, G., He, H., … (2025). Exploring housing price dynamics in sustainable cities through a cooperated big data driven machine learning method: Case study on a typical city in China. City and Environment Interactions, 28, 100223. https://doi.org/10.1016/j.cacint.2025.100223
36. Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., … Zammit-Mangion, A. (2019). A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics, 24(3), 398–425. https://link.springer.com/article/10.1007/s13253-018-00348-w
37. Helbich, M., Brunauer, W., Vaz, E., & Nijkamp, P. (2014). Spatial heterogeneity in hedonic house price models: The case of Austria. Urban Studies, 51(2), 390–411. https://doi.org/10.1177/0042098013492234
38. Hengl, T., Heuvelink, G. B. M., & Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma, 120*(1–2), 75–93. https://doi.org/10.1016/j.geoderma.2003.08.018

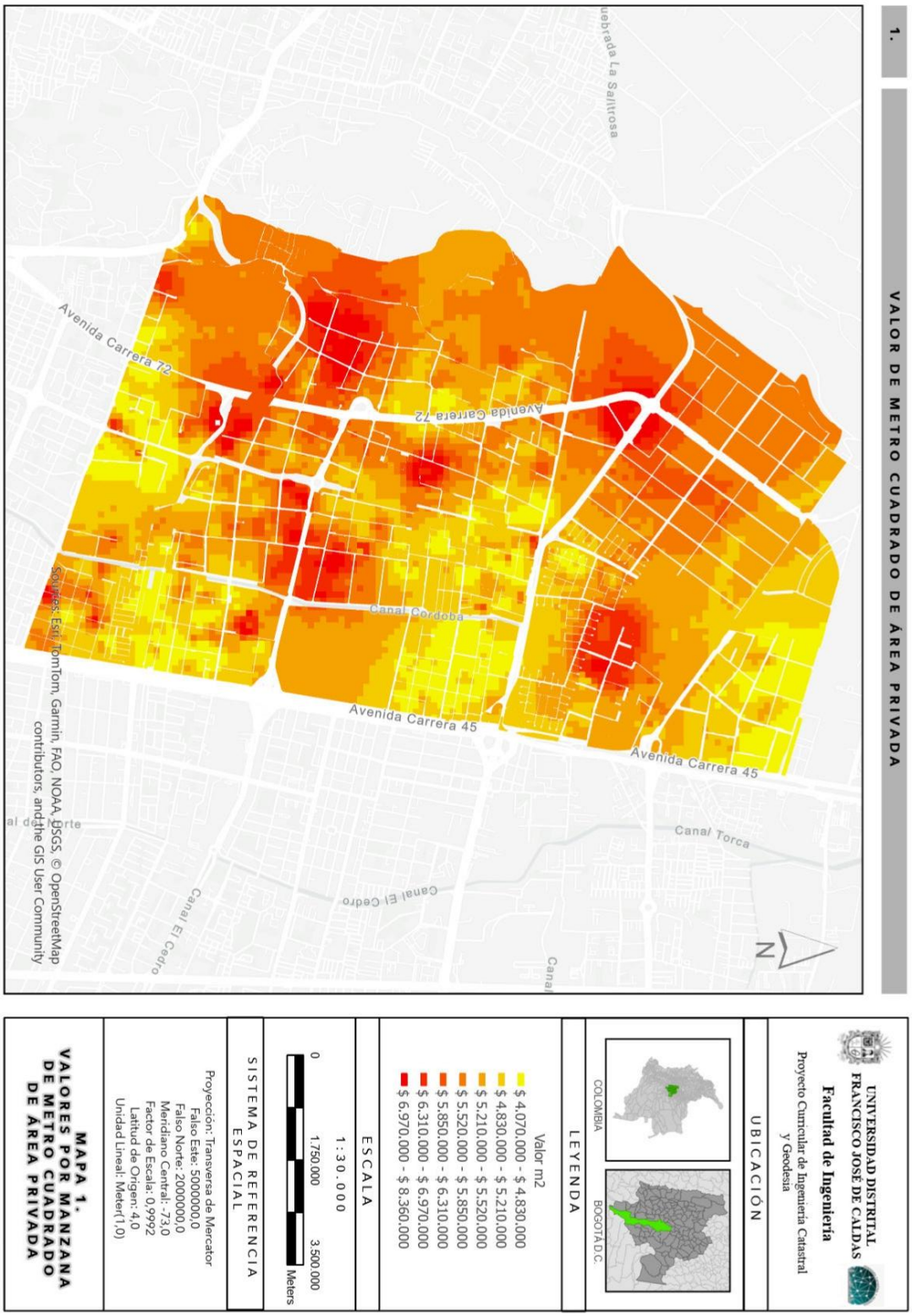
39. Hengl, T., Nussbaum, M., Wright, M. N., Heuvelink, G. B. M., & Gräler, B. (2018). Random forest as a generic framework for predictive mapping of soil properties. *PeerJ, 6*, e5518. https://doi.org/10.7717/peerj.5518
40. Higgins, C. D., Arku, R. N., Farber, S., & Miller, E. J. (2024). Modelling changes in accessibility and property values associated with the King Street Transit Priority Corridor project in Toronto. Transportation Research Part A: Policy and Practice, 190, 104256. https://www.sciencedirect.com/science/article/pii/S0965856424003045 ScienceDirect
41. Hjort, A., Williams, J. P., & Pensar, J. (2024). Clustered Conformal Prediction for the Housing Market. En Proceedings of the 13th Symposium on Conformal and Probabilistic Prediction with Applications (PMLR 230, pp. 366–386). https://proceedings.mlr.press/v230/hjort24a.html
42. International Association of Assessing Officers. (2025). Standard on mass appraisal of real property (Rev. 2025). https://www.iaao.org/wp content/uploads/Standard_on_Mass_Appraisal.pdf
43. International Organization for Standardization. (2023). ISO 19157-1:2023 Geographic information — Data quality — Part 1: General principles. https://www.iso.org/standard/78900.html
44. Isaaks, E. H., & Srivastava, R. M. (1989). An Introduction to Applied Geostatistics. Oxford University Press.
45. IVSC – International Valuation Standards Council (2021). International Valuation Standards (IVS). Recuperado de: https://acortar.link/tPYXDt
46. Jim, C. Y., & Chen, W. Y. (2006). Impacts of urban environmental elements on residential housing prices in Guangzhou (China). *Landscape and Urban Planning*, *78*(4), 422-434. https://doi.org/10.1016/j.landurbplan.2005.12.003
47. Joseph, V. R. (2025). Rational Kriging. Journal of the American Statistical Association, 120(549), 548–558. https://doi.org/10.1080/01621459.2024.2356296
48. Journel, A. G., & Huijbregts, C. J. (1978). *Mining Geostatistics.* Academic Press.
49. Kelejian, H. H., & Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. Journal of Econometrics, 157(1), 53–67. Enlace: PMC+1
50. Keskin, H., & Grunwald, S. (2018). Regression-kriging as a workhorse in digital soil mapping. *Geoderma, 326*, 22–41. https://doi.org/10.1016/j.geoderma.2018.04.004
51. Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications in Hydrogeology.* Cambridge University Press.
52. Kok, N., Koponen, E. L., & Martínez-Barbosa, C. A. (2017). Big data in real estate? From manual appraisal to automated valuation. Journal of Portfolio Management, 43(6), 202–211. https://doi.org/10.3905/jpm.2017.43.6.202
53. Koster, H. R. A., & van Ommeren, J. (2023). Neighbourhood stigma and place-based policies. Economic Policy, 38(114), 289–339. https://doi.org/10.1093/epolic/eiac039

54. Lee, L.-F., & Yu, J. (2010). Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics, 154(2), 165–185. Enlace: IDEAS/RePEc+1
55. Leggett, C. G., & Bockstael, N. E. (2000). Evidence of the Effects of Water Quality on Residential Land Prices. Journal of Environmental Economics and Management, 39(2), 121–144. https://doi.org/10.1006/jeem.1999.1096
56. Li, J., & Heap, A. D. (2014). Spatial interpolation methods applied in the environmental sciences: A review. *Environmental Modelling & Software, 53*, 173–189. https://doi.org/10.1016/j.envsoft.2013.12.008
57. Lindgren, F., Rue, H., & Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: The stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(4), 423–498. https://doi.org/10.1111/j.1467-9868.2011.00777.x
58. Liu, Z., Zhang, Z., Zhou, C., Ming, W., & Du, Z. (2021). An Adaptive Inverse-Distance Weighting Interpolation Method Considering Spatial Differentiation in 3D Geological Modeling. Geosciences, 11(2), 51. https://doi.org/10.3390/geosciences11020051
59. Lu, B., Hu, Y., Murakami, D., Brunsdon, C., Comber, A., Charlton, M., & Harris, P. (2022). High-performance solutions of geographically weighted regression in R. Geo-spatial Information Science, 25(4), 536–549. https://doi.org/10.1080/10095020.2022.2064244
60. Lucas, M. P., Lamb, R., & Siciliano, S. D. (2022). Optimizing automated kriging to improve spatial interpolation of climate variables. *Journal of Hydrometeorology, 23*(4), 739–756. https://doi.org/10.1175/JHM-D-21-0171.1
61. Mahoney, M. W., Johnson, L. K., Silge, J., Frick, H., Kuhn, M., & Beier, C. M. (2023). Assessing the performance of spatial cross-validation approaches for models of spatially structured data. arXiv preprint arXiv:2303.07334. https://arxiv.org/abs/2303.07334
62. McCord, M. J., McCord, J., Davis, P. T., Haran, M., & Bidanset, P. (2019). House Price Estimation Using an Eigenvector Spatial Filtering Approach. International Journal of Housing Markets and Analysis, 13(5), 845–867. https://doi.org/10.1108/IJHMA-09-2019-0097
63. McMillen, D. P. (2010). Issues in spatial data analysis. Journal of Regional Science, 50(1), 119–141. Enlace: Wiley Online Library
64. Murakami, D., & Griffith, D. A. (2019). Eigenvector Spatial Filtering for Large Data Sets: Fixed and Random Effects Approaches. Geographical Analysis, 51(1), 23–49. https://doi.org/10.1111/gean.12156
65. Murakami, D., Sugasawa, S., Seya, H., & Griffith, D. A. (2024). Sub-model Aggregation for Scalable Eigenvector Spatial Filtering: Application to Spatially Varying Coefficient Modeling. Geographical Analysis. https://doi.org/10.1111/gean.12393
66. Nikou, M., & Tziachris, P. (2022). Prediction and Uncertainty Capabilities of Quantile Regression Forests in Estimating Spatial Distribution of Soil Organic Matter. ISPRS International Journal of Geo-Information, 11(2), 130. https://doi.org/10.3390/ijgi11020130

67. Oliver, M. A., & Webster, R. (2015). Basic Steps in Geostatistics: The Variogram and Kriging. SpringerBriefs in Agriculture. https://doi.org/10.1007/978-3-319-15865-5
68. Open Geospatial Consortium. (2011). Simple Feature Access — Part 1: Common Architecture (OGC/ISO 19125-1). https://www.ogc.org/publications/standard/sfa/
69. Peng, Z., & Inoue, R. (2022). Identifying Multiple Scales of Spatial Heterogeneity in Housing Prices Based on Eigenvector Spatial Filtering Approaches. ISPRS International Journal of Geo-Information, 11(5), 283. https://doi.org/10.3390/ijgi11050283
70. Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning.* MIT Press.
71. Rennert, L. (2022). A meta-analysis of the impact of rail stations on property values: Applying a transit planning lens. Transport Policy, 163, 165–180. https://www.sciencedirect.com/science/article/pii/S0965856422001677 ScienceDirect
(Versión institucional abierta: London School of Economics ePrints). https://eprints.lse.ac.uk/115511/ eprints.lse.ac.uk
72. Rey-Blanco, D., García-Palomares, J. C., & Gutiérrez, J. (2024). Improving hedonic housing price models by integrating accessibility measures. Expert Systems with Applications, 245, 122904. https://www.sciencedirect.com/science/article/pii/S0957417423015610
73. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography, 40(8), 913–929. https://doi.org/10.1111/ecog.02881
74. Rosen, S. (1974). Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy, 82*(1), 34 https://www.jstor.org/stable/1830899
75. Rossiter, D. G. (2018). Co-kriging with the gstat package of the R environment for statistical computing (Technical Note, July 14, 2018). Cornell University. https://www.css.cornell.edu/faculty/dgr2/_static/files/R_PDF/CoKrigeR.pdf
76. Royal Institution of Chartered Surveyors. (2020). RICS Valuation – Global Standards Red Book Global Standards. RICS. Edición efectiva desde el 31/01/2020.
77. Sampson, P. D., Szpiro, A. A., Sheppard, L., Lindström, J., & Kaufman, J. D. (2013). A regionalized national universal kriging model using Partial Least Squares regression for estimating annual PM2.5 concentrations in epidemiology. Environmental Health Perspectives, 121(8), 873–879.
78. Schabenberger, O., & Gotway, C. A. (2005). Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC. https://doi.org/10.1201/9781315275086
79. SciPy. (2024a). scipy.stats.shapiro — Shapiro–Wilk test. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html Documentación de Numpy y Scipy

80. SciPy. (2024b). scipy.stats.normaltest — D'Agostino–Pearson test. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.normaltest.html Documentación de Numpy y Scipy
81. SciPy. (2024c). scipy.stats.anderson — Anderson–Darling test. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.anderson.html Documentación de Numpy y Scipy
82. Sharma, H., Dua, N., & Tammana, R. (2024). An Optimal House Price Prediction Algorithm: XGBoost. arXiv preprint arXiv:2402.04082. https://arxiv.org/abs/2402.04082
83. Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* Springer. https://doi.org/10.1007/978-1-4612-1494-6
84. Tyrväinen, L., & Miettinen, A. (2000). Property prices and urban forest amenities. *Journal of Environmental Economics and Management*, *39*(2), 205-223. https://doi.org/10.1006/jeem.1999.1097
85. Valavi, R., Elith, J., Lahoz-Monfort, J. J., & Guillera-Arroita, G. (2019). blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. Methods in Ecology and Evolution, 10(2), 225–232. https://doi.org/10.1111/2041-210X.13107
86. Villada-Medina, J., Hernández-Muñoz, J., & Arango-Londoño, C. (2022). *A house price modeling based on clustering and kriging: The Medellín case*. Revista RRS, 12(4), 1–17. https://rrs.scholasticahq.com/article/66198-a-house-price-modeling-based-on-clustering-and-kriging-the-medellin-case
87. Virtanen, P., Gommers, R., Oliphant, T. E., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nature Methods, 17(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2
88. Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction with Applications* (3rd ed.). Springer. https://doi.org/10.1007/978-3-662-05294-5
89. Waller, L. A., & Gotway, C. A. (2004). Applied Spatial Statistics for Public Health Data. Wiley. Enlace: Google Books+1
90. Webster, R., & Oliver, M. A. (2007). *Geostatistics for Environmental Scientists* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9780470517277
91. Yang, S., Wong, S. W. K., & Kou, S. C. (2021). Inference of dynamic systems from noisy and sparse data via manifold-constrained Gaussian processes. Proceedings of the National Academy of Sciences, 118(15), e2020397118. https://doi.org/10.1073/pnas.2020397118

10. ANEXOS


1.
VALOR DE METRO CUADRADO DE ÁREA PRIVADA
Avenida Carrera 72
Avenida Carrera 45
Canal Córdoba
Canal Torca
Canal El Cedro
Source: Esri, TomTom, Garmin, FAO, NOAA, USGS, © OpenStreetMap contributors, and the GIS User Community
N
UNIVERSIDAD DISTRITAL FRANCISCO JOSE DE CALDAS
Facultad de Ingeniería
Proyecto Curricular de Ingeniería Catastral y Geodesia
UBICACIÓN
COLOMBIA
BOGOTÁ D.C.
LEYENDA
Valor m2
$ 4.070.000 - $ 4.830.000
$ 4.830.000 - $ 5.210.000
$ 5.210.000 - $ 5.520.000
$ 5.520.000 - $ 5.850.000
$ 5.850.000 - $ 6.310.000
$ 6.310.000 - $ 6.970.000
$ 6.970.000 - $ 8.360.000
ESCALA
1:30.000
0
1.750.000
3.500.000
Meters
SISTEMA DE REFERENCIA ESPACIAL
Proyección: Transverca de Mercator
Falso Este: 5000000,0
Falso Norte: 2000000,0
Meridiano Central: -73,0
Factor de Escala: 0,9992
Latitud de Origen: 4,0
Unidad Lineal: Meter(1,0)
MAPA 1.
VALORES POR MANZANA DE METRO CUADRADO DE ÁREA PRIVADA