FACULTAD DE ESTUDIOS ESTADÍSTICOS

GRADO EN ESTADISTICA APLICADA

Curso 2023/2024

Trabajo de Fin de Grado

TITULO: Valoración automática de viviendas

***Alumno*: Coral García Cardeñas**

***Tutor*: Lorenzo Escot Mangas**

Junio de 2024

UNIVERSIDAD COMPLUTENSE
MADRID

DEDICATORIA

A mi padre y mi hermano, por apoyarme siempre antes de irse y confiar en mí. Por estar aún sin estar presentes.

A mi madre, por ayudarme en todo y ser un ejemplo de que no hay que rendirse.

A mis amigos, por estar siempre a mi lado y hacerme tan feliz.

Resumen

La valoración de inmuebles es un proceso fundamental para determinar la estimación del valor de mercado de una propiedad. Profesionales especializados, como arquitectos y tasadores, evalúan diversos factores, incluyendo la ubicación, superficie y características físicas de las viviendas, de acuerdo con las normas vigentes. En los últimos años, el uso del Big Data y los modelos de valoración automatizada (AVM's) han revolucionado este campo. Los AVM's utilizan algoritmos y técnicas de inteligencia artificial para estimar los valores de las propiedades, entrenando modelos con grandes bases de datos que incorporan técnicas econométricas espaciales. Estas técnicas consideran las características geográficas y espaciales de los datos, mejorando significativamente la precisión de las estimaciones.

En este trabajo, se propone evaluar el impacto de las técnicas econométricas espaciales en la valoración de propiedades inmobiliarias. El objetivo es desarrollar modelos de valoración automatizada que incorporen componentes espaciales, como la ubicación geográfica, empleando el software R. Para ello, se recopilarán y limpiarán datos espaciales, asegurando su calidad para identificar patrones y relaciones significativas. Una vez recopilados y limpios, se realizará un análisis exploratorio para extraer información valiosa sobre cómo la ubicación y otras características influyen en el valor de las propiedades. Posteriormente, se desarrollarán modelos predictivos que incorporen la ubicación y otras características, evaluando su precisión y efectividad.

Los resultados se interpretarán y presentarán mediante técnicas de visualización, lo que permitirá una compresión más clara. Este enfoque no sólo aporta una perspectiva innovadora al campo de la valoración automática, sino que también tiene el potencial de mejorar significativamente la precisión y eficacia de las valoraciones, beneficiando a su vez a tasadores, compradores o vendedores del mercado inmobiliario.

Abstract

Real estate valuation is a fundamental process to discover the estimated market value of a property. Specialized professionals such as architects and appraisers, evaluate various factors, including the location, surface area and physical characteristics of homes, considering the current standards. In recent years, the use of Big Data and automated valuation models (AVMs) have changed this field. AVMs use algorithms and artificial intelligence techniques to calculate property values by training models with large databases that incorporate spatial econometric techniques. These techniques consider the geographical and spatial characteristics of the data, highly improving the precision of the estimates.

In this paper, we propose to evaluate the impact of spatial econometric techniques in real estate valuation. The objective is to develop automated valuation models that incorporate spatial components, such as geographic location, using R software. For this purpose, spatial data will be collected and cleaned, guaranteeing its quality, to identify significant patterns and relationships. Once collected and cleaned, an exploratory analysis will be performed to extract valuable information on how location and other characteristics influence property values. Subsequently, predictive models incorporating location and other characteristics will be developed and evaluated for accuracy and effectiveness.

The results will be interpreted and presented using visualization techniques, allowing for a clearer understanding. This approach not only brings an innovative perspective to the field of automatic valuation, but also has the potential to significantly improve the accuracy and efficiency of valuations, guaranteeing benefits to appraisers, buyers or sellers in the real estate market.

Índice

Tablas

Figuras

1. Introducción

La valoración de inmuebles se define como el cálculo del valor de tasación de la propiedad en el mercado inmobiliario para determinar el precio que tiene un inmueble en un momento específico. Se utiliza en diversas transacciones y decisiones financieras relacionadas con la propiedad, como la compra, venta, refinanciación, seguro y tasación fiscal. Los profesionales encargados de llevar a cabo las tasaciones, como arquitectos o arquitectos técnicos, evalúan aspectos como la ubicación, la superficie, características físicas, antigüedad y estado de conservación, entre otros, siguiendo estándares y procedimientos establecidos por leyes y regulaciones.

Es importante tener en cuenta que el valor de tasación de un inmueble también puede variar según diversos factores del entorno socioeconómico incluyendo las condiciones del mercado, la oferta y la demanda o las condiciones económicas locales y regionales. Por lo tanto, las tasaciones deben realizarse de manera objetiva y profesional para garantizar que reflejen con precisión el valor real del activo inmobiliario en el momento de la evaluación.

1.1. Big data y modelos AVM

El concepto Big Data hace referencia a grandes volúmenes de datos, generalmente no estructurados, empleados para la compilación, gestión y análisis utilizando métodos tradicionales de procesamiento de datos. Esta revolución en la información ha transformado la manera en que entendemos y utilizamos los datos en prácticamente todos los sectores, incluido el mercado inmobiliario.

Los Modelos de Valoración Automatizada (AVM) son herramientas que emplean algoritmos matemáticos y estadísticos, así como de inteligencia artificial para estimar de manera automatizada el valor de un inmueble en una ubicación geográfica específica y en un momento determinado. Dentro del grupo de métodos de valoración avanzados se encuentran el modelo de valoración hedónica, métodos basados en análisis de series temporales, modelos espaciales y de kriging, redes neuronales artificiales y árboles de decisión, entre otros.

La integración del Big Data con los modelos AVM permite una valoración más precisa y oportuna de las propiedades, al considerar una amplia gama de factores y variables, tales como datos socioeconómicos, las tendencias del mercado, características de la propiedad y preferencias de los compradores. Además, este enfoque presenta diversas ventajas, como reducción de costes, ahorro de tiempo y papel, eficiencia y objetividad. No obstante, es importante tener en cuenta algunos inconvenientes, como la falta de validez de los AVM como herramienta de tasación y posibles sesgos en los datos.

Además, es esencial considerar la estadística espacial y la econometría espacial en el análisis de datos inmobiliarios. La estadística espacial se enfoca en el análisis de datos con referencias geográficas, permitiendo comprender cómo se distribuyen las variables en el espacio, identificar patrones y realizar predicciones. Por otro lado, la econometría espacial busca cuantificar las relaciones entre variables económicas considerando su ubicación geográfica. Estos métodos permiten modelar y comprender fenómenos económicos que no pueden ser completamente capturados por modelos tradicionales debido a la estructura espacial de los datos.

En resumen, la incorporación del Big Data impulsa la evolución de los modelos de valoración automatizada, ofreciendo una mayor precisión y eficiencia en el proceso de valoración de

inmuebles. Sin embargo, es crucial abordar los desafíos asociados para garantizar su efectividad y fiabilidad en el mercado inmobiliario.

1.2. Objetivos

Este trabajo tiene como objetivo principal examinar la relevancia y el impacto de la inclusión de métodos econométricos en el proceso de valoración de bienes inmuebles, centrándose en el empleo de modelos AVM y utilizando la herramienta de programación R como medio para su desarrollo y análisis. Se adoptará un enfoque práctico que integre el análisis espacial, lo que permitirá comprender de manera más completa cómo la ubicación geográfica influye en el valor de las propiedades.

Para llevar a cabo este estudio, en primer lugar, se procederá a la recopilación y limpieza de los datos espaciales, tales como la ubicación geográfica de las propiedades u otras características que puedan influir. Posteriormente se realizará un análisis exploratorio de estos datos para identificar patrones espaciales y relaciones significativas entre las variables. Seguidamente, se desarrollarán modelos de valoración automática que incorporen componentes espaciales utilizando técnicas econométricas avanzadas. Estos modelos permitirán predecir el precio de las propiedades en función de su ubicación geográfica y otras características. Se emplearán métodos de validación para evaluar la precisión y la efectividad de los modelos desarrollados.

Una vez que se hayan obtenido los resultados, se procederá a su interpretación y análisis en profundidad. Se buscará comprender cómo las variables espaciales influyen en el precio de las propiedades y se identificarán posibles patrones o tendencias en los datos. Además, se utilizarán técnicas de visualización para comunicar de manera clara y efectiva los hallazgos del estudio.

1.3. Estado del arte en el uso del R

El sistema de R es un entorno de software libre para computación estadística y para gráficos que ha experimentado un crecimiento significativo en el análisis de datos espaciales en los últimos años. Con una gran variedad de paquetes como ‘spatial’, ‘spdep’, ‘sf’, ‘raster’, ‘spatialreg’, entre otros, R ofrece una amplia gama de herramientas para manejar y analizar datos espaciales.

Uno de los hitos más importantes en este avance fue la creación del paquete ‘sp’, que introdujo clases y métodos estándar para datos espaciales en R. Estas clases han facilitado la organización y manipulación de datos, facilitando la compatibilidad entre diferentes paquetes de análisis espacial.

Por estas razones, se ha elegido utilizar R como la herramienta principal para llevar a cabo este trabajo, aprovechando la variedad de funciones y la comunidad activa de desarrolladores y usuarios que respaldan su desarrollo continuo.

2. Metodología

1.1. Análisis exploratorio

El análisis exploratorio de datos espaciales (AEDE) es una fase crucial en el proceso de su comprensión y su distribución geográfica, y que debe ser la etapa inicial de un estudio econométrico que involucre datos georreferenciados. Cuando disponemos de estos datos, debemos utilizar herramientas y técnicas tanto gráficas como numéricas que permitan:

- Describir y visualizar distribuciones espaciales.
- Identificar localizaciones atípicas.
- Descubrir patrones de asociación espacial, clústeres o puntos calientes/fríos.
- Sugerir regímenes espaciales u otras formas de heterogeneidad espacial.

1.1.1. Efectos espaciales

Los efectos espaciales se refieren a cómo la ubicación geográfica influye en las variables de interés. Es esencial comprender si existen patrones espaciales, es decir, si ciertas áreas tienen valores similares o diferentes en comparación con sus vecinas. Hay dos tipos principales de efectos espaciales: la dependencia espacial y la heterogeneidad espacial.

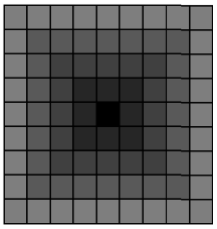
Autocorrelación espacial

La dependencia espacial, también conocida como autocorrelación espacial (AE), es la ausencia de independencia entre observaciones en un punto del espacio, donde la proximidad geográfica influye en los valores de las variables. Esto nos ayuda a entender si hay patrones espaciales en los datos y cómo se relacionan entre sí las diferentes áreas. Como expresa la primera ley de la geografía de Tobler (1979), "*Todas las cosas están relacionadas entre sí, pero las cosas más próximas en el espacio tienen una relación mayor que las distantes*".

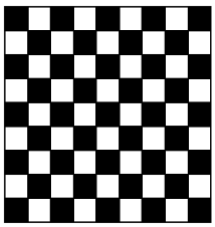
Esto se podría explicar por tres condiciones, como errores de medición para observaciones en unidades espaciales contiguas, problemas de agregación espacial y la influencia del espacio en los procesos de interacción, como la difusión y las jerarquías espaciales.

La autocorrelación espacial puede ser positiva o negativa:

- AE positiva: las observaciones similares tienden a agruparse juntas en el espacio. Esto es, las áreas cercanas entre sí tienen valores similares.
- AE negativa: las observaciones similares tienden a dispersarse en el espacio. Las áreas cercanas son diferentes entre sí.
- AE nula: la variable se distribuye de forma aleatoria (no existen patrones en la distribución espacial).



AE positiva



AE negativa

Heterogeneidad espacial

La heterogeneidad espacial (HE) está relacionada con la falta de uniformidad sobre el espacio, donde diferentes áreas o unidades espaciales tienen efectos diferentes en los fenómenos estudiados. Esto puede reflejarse de dos formas:

- Inestabilidad estructural: los parámetros de un modelo, como la regresión, varían en función de su ubicación. En otras palabras, no son homogéneos para el conjunto de datos.

$$Y_i = X_i\beta_i + \varepsilon_i \quad i = 1, \ldots, n$$

- Heterocedasticidad: errores de medición debido a la falta de variables significativas. Varianzas diferentes en el término de error, que puede representarse como:

$$Var(\varepsilon_i) = \sigma^2{}_i$$

donde $\sigma^2{}_i$ indica que la varianza de la perturbación aleatoria es diferente para cada observación muestral i.

La mayoría de estos problemas, pueden ser resueltos por técnicas de econometría estándar.

1.1.2. Estimación de la autocorrelación espacial

Matriz de pesos espaciales

Para cuantificar las relaciones espaciales, se utiliza lo que se conoce como matriz de pesos espaciales. Esta matriz es crucial para estos modelos, ya que define cómo se relacionan espacialmente las observaciones entre sí. Existen diferentes formas de construir esta matriz, siendo las más comunes la contigüidad (para datos de polígonos o regiones) y la distancia (para puntos).

En nuestro caso, al estar tratando con puntos geoespaciales, utilizaremos la matriz de pesos espaciales basada en la distancia. Estas distancias permiten definir qué tan cerca o lejos están las observaciones entre sí, y de esta manera determinar cómo influyen unas sobre otras en términos espaciales.

La estructura de la matriz de pesos espaciales viene representada de la siguiente manera:

$$w = \begin{pmatrix} 0 & w_{12} \ldots w_{1n} \\ w_{21} & 0 \cdots w_{2n} \\ & \ldots \\ w_{n1} & w_{n2} \cdots 0 \end{pmatrix}$$

Para llegar a ella, se utilizan diferentes transformaciones, pero nosotros nos centraremos en la relación basada en la inversa de la distancia, de la siguiente manera:

$$w_{ij} = \begin{cases} 1/d_{ij}^{\gamma} = d_{ij}^{-\gamma}, & \forall ij \in \{1, \ldots, N\} \\ 0 & \forall i = j \end{cases} \qquad \text{con } \gamma > 0$$

El parámetro γ nos permite penalizar en menor o mayor medida la proximidad espacial en función de la distancia. En nuestro caso, utilizaremos $\gamma = 2$.

Esta transformación respeta la ley de Tobler mencionada anteriormente, ya que los pesos son mayores cuando las observaciones están más cercanas o menores cuando están más alejadas.

Antes de estimar la autocorrelación espacial, se recomienda estandarizar la matriz obtenida. Aunque no es algo obligatorio, tiene muchos efectos positivos. Consiste en ajustar la matriz para que la suma de los elementos de cada fila sea igual a uno.

$$w_{i,j}^{s} = \frac{w_{ij}}{\Sigma_j w_{ij}} \quad tal\ que \quad \sum_j w_{i,j}^{s} = 1$$

Esto facilita la comparación de la influencia relativa de las observaciones además de facilitar la implementación de métodos estadísticos que las utilicen, haciendo que funcionen de manera más robusta y eficiente.

Retardo espacial

El retardo espacial de una variable Y en una ubicación i es la media ponderada de los valores de Y en ubicaciones vecinas. Dicho de otra forma, el retardo espacial se define como la influencia que tienen las observaciones sobre otras cercanas. La matriz de pesos mencionada anteriormente se utiliza para construir las variables retardadas espacialmente:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \longrightarrow w \cdot y = \begin{pmatrix} \Sigma_j w_{1j}^{s} y_j \\ \Sigma_j w_{2j}^{s} y_j \\ \dots \\ \Sigma_j W_{nj}^{s} y_j \end{pmatrix}$$

Estadísticos globales de autocorrelación espacial

La autocorrelación espacial global mide el grado general de similitud espacial en toda el área de estudio. Es una medida que indica si hay tendencia general de agrupamiento o dispersión.

Una vez hallada la matriz de pesos espaciales, podemos estimar la autocorrelación espacial, para ello se utilizan diferentes medidas de asociación o dependencia espacial global, contrastando la hipótesis nula de “ausencia de autocorrelación espacial”.

- **I de Moran:**

$$I = \frac{n}{s_0} \frac{\Sigma_{ij} w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\Sigma_i (y_i - \bar{y})^2} \quad \text{siendo} \quad S_0 = \sum_i \sum_j w_{ij}^{s}$$

  - -1: correlación espacial perfecta positiva
  - 0: ausencia de correlación espacial
  - 1: correlación espacial perfecta negativa

- **C de Geary:** es un índice de comparaciones por pares entre las diferentes zonas.

$$c = \frac{N-1}{2s_0}\frac{\Sigma_{ij} w_{ij}(y_i - y_j)}{\Sigma_{i=1}(y_i - \bar{y})^2} \; i \neq j$$

Valores menores que 1 indican autocorrelación espacial positiva, mientras que valores superiores a 1 muestran autocorrelación espacial negativa.

- **G(d) de Getis y Ord:** medida de concentración (o escasez de concentración) espacial de una variable Y.

$$\mathrm{G(d)} = \frac{\sum_{i=1} \Sigma_{j=1} w_{ij}(d) y_i y_j}{\Sigma_{i=\Lambda} \sum_{j=1} y_i y_j} \; i \neq j$$

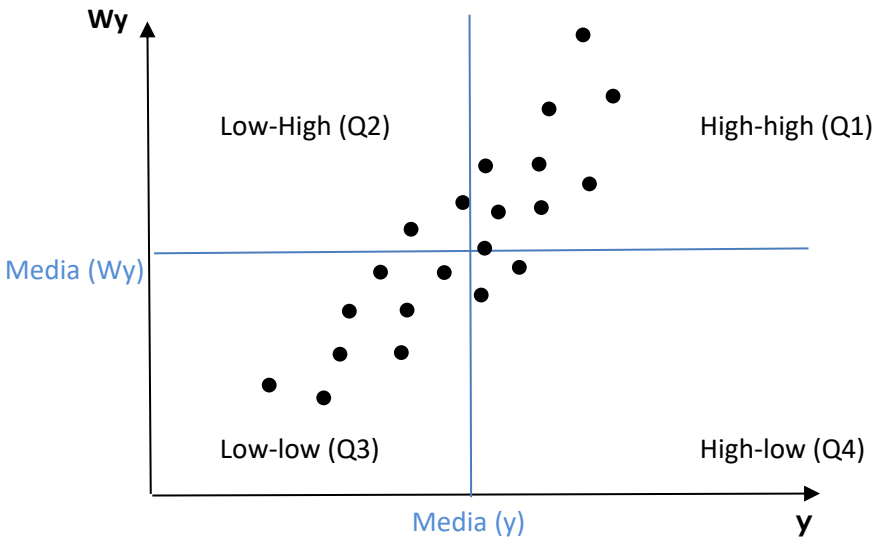
Estadísticos de asociación espacial local

La autocorrelación espacial local mide la similitud espacial en áreas más específicas. Permite identificar puntos calientes (clústeres de valores altos), puntos fríos (clústeres de valores bajos) y áreas con puntos atípicos.

El **Índice I de Moran Local** (**LISA**) es una de las medidas más comunes para calcular la autocorrelación espacial local. Se calcula como:

$$I_i = \frac{(x_i - \bar{x})}{m_2} \sum_{j=1}^{n} w_{ij}(x_j - \bar{x}) \text{ para } i \neq j \text{ donde } m_2 = \frac{\Sigma_{i=1}^{n}(x_j - \bar{x})^2}{n}$$

- Clústeres mediante Scatterplot de Moran:



  - Cuadrante I (Q1): observaciones con valores altos rodeados de observaciones con valores altos.
  - Cuadrante II (Q2): observaciones con valores bajos rodeados de observaciones con valores altos.
  - Cuadrante III (Q3): observaciones con valores bajos rodeados de observaciones con valores bajos.
  - Cuadrante IIII (Q4): observaciones con valores altos rodeados de observaciones con valores bajos.

1.2. Modelos de regresión

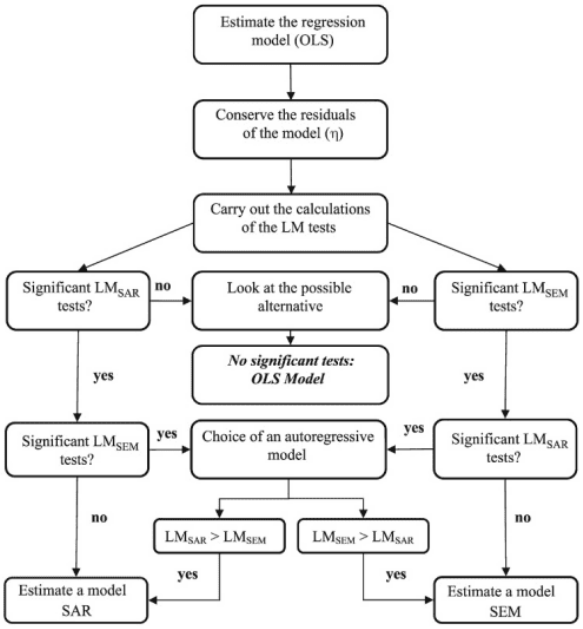
Los modelos de regresión son herramientas estadísticas que se utilizan para describir la relación entre una variable dependiente (o variable respuesta) y una o más variables independientes (o predictoras). Existen varios modelos de regresión, cada uno adecuado para diferentes tipos de datos y relaciones.

El modelo de regresión de Mínimos Cuadrados Ordinarios (MCO) es uno de los métodos más comunes en la econometría y la estadística para estimar los parámetros de un modelo de regresión lineal. El propósito del MCO es minimizar la suma de los cuadrados de los errores, que son las diferencias entre los valores observados y los valores predichos por el modelo.

Este modelo se puede expresar como:

$$Y_i = X_i\beta_i + \varepsilon_i \quad i = 1, \ldots, n$$

Después de haber detectado la autocorrelación espacial, verificamos si también se encuentra en el modelo de regresión, para ello, se pueden utilizar los multiplicadores de Lagrange (LM tests). Estos test permiten evaluar si es necesario incluir términos adicionales que capturen la dependencia espacial, con un término autorregresivo (SAR) o un término de error espacial (SEM), además permite contrastar la existencia conjunta de ambos tipos de dependencia espacial.



Fuente: "Econometría espacial con microdatos" Jean Dubé, Diego Legros.

1.2.1. Modelos de regresión espacial

- **Modelo espacial autorregresivo (SAR):**

Este modelo es útil en casos en los que el valor que adopta una variable en una ubicación depende de los valores de otras variables explicativas en la misma ubicación y del valor de esa misma variable en otras ubicaciones vecinas, incumpliéndose así el principio de independencia entre las observaciones muestrales. Para implementarlo, se introduce un término de rezago

espacial de la variable dependiente, que se incorpora al modelo como una variable explicativa adicional.

$$y = \rho w y + X\beta + u$$

$$u \sim N(0, \sigma^2 I)$$

- **Modelo de error espacial (SEM):**

Es el modelo más utilizado cuando el modelo básico de regresión lineal no logra explicar el efecto espacial, trasladando entonces ese efecto hacia los términos de error.

$$y = X\beta + u$$

$$u = \lambda w u + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I)$$

- **Modelo espacial autorregresivo combinado (SAC) (Kelijian-Prucha model):**

Este modelo incluye tanto un término autorregresivo espacial en la variable dependiente como un término de error espacial.

$$y = \rho w_1 y + X\beta + u$$

$$u = \lambda w_2 u + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I)$$

$$w_1 = w_2 = w$$

- **Modelo de error espacial de Durbin (SDEM):**

Incluye términos adicionales de las variables independientes multiplicadas por la matriz de pesos espaciales.

$$y = X\beta + w_1 X\gamma + u$$

$$u = \lambda w_2 u + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2 I)$$

- **Modelo autorregresivo espacial condicional (CAR):**

Este modelo se utiliza para analizar datos espaciales donde existe dependencia entre las observaciones en distintas ubicaciones geográficas. La idea principal es que la distribución de cada componente está condicionada a la de sus vecinos cercanos.

$$\left(Y_i | Y_{(i)}\right) \sim N\left(x_i'\beta + \sum_{j=1}^{n} c_{ij}\left(Y_j - x_j'\beta\right), \sigma_i^2\right)$$

- $y_{(i)} = \{y_j : j \neq i\}$
- $\beta = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^P$ parámetros de regresión
- $\sigma_i^2 > 0 \quad c_{ij} \geq 0$ parámetros de covarianza

M = diag$(\sigma_1^2, \ldots, \sigma_n^2)$

C = (c_{ij})

- $M^{-1}C$ es simétrica
- $M^{-1}(I_n - C)$ es definida positiva

$$Y \sim N_n(X\beta, (I_n - C)^{-1}M)$$

Suposiciones del modelo:

i. $M = \sigma^2 I_n \ \ \sigma^2 > 0$
ii. $C = \phi W$, ϕ un 'parámetro espacial' desconocido y $W = (w_{ij})$ matriz de pesos conocida $w_{ij} > 0$ si i y j son vecinos.

De esta forma logramos capturar de manera explícita la dependencia espacial, facilitando un análisis más preciso y realista.

3. Recopilación y preprocesamiento de datos

3.1. Obtención de los datos

Para la realización de este trabajo hemos utilizado "idealista18", un paquete de R que contiene diferentes conjuntos de datos georreferenciados originalmente publicados en idealista y que además han sido enriquecidos con datos oficiales del catastro español. Este paquete proporciona tres conjuntos de datos para las ciudades de Madrid, Valencia y Barcelona, junto con listados de polígonos de barrios y puntos de interés para cada ciudad. En nuestro análisis, nos enfocaremos en el conjunto de datos correspondiente al año 2018 de la ciudad de Madrid.

- **Madrid_Sale**: Este conjunto de datos, es un objeto de tipo "sf". Consta de 94.815 observaciones y 42 variables diferentes que describen características de cada vivienda.
- **Madrid_Polygons**: Se trata de un objeto "sf" que contiene polígonos que representan los barrios de Madrid, un identificador y el nombre del barrio. En total, hay 135 barrios incluidos.
- **Madrid_Pois**: Este conjunto es un objeto "lista" que proporciona coordenadas de puntos de interés en Madrid. Incluye tres elementos: las coordenadas del centro de la ciudad, de las estaciones de metro y de la calle principal.

Estos conjuntos de datos nos proporcionan una amplia información geográfica y características de viviendas que van a ser fundamentales para nuestro análisis.

Para la obtención de los datos, importaremos esta biblioteca y leeremos adecuadamente cada conjunto de datos.

Además, para la representación de algunos mapas, hemos descargado cartografías de los 21 distritos de Madrid, descargadas del portal de datos abiertos del Ayuntamiento de Madrid.

3.2. Análisis descriptivo

El análisis descriptivo es un proceso fundamental en cualquier proyecto de análisis de datos. Permite comprender como se distribuyen las distintas variables y ayuda a identificar posibles anomalías en los datos que necesiten ser abordados.

Antes de realizar el análisis descriptivo, es crucial entender las variables que conforman el conjunto de datos de **Madrid_Sale.** A continuación, describimos brevemente el significado de cada variable:

	Variable	Descripción	Tipo
1	ASSETID	Identificador único para cada vivienda	Categórica nominal
2	PERIOD	Mes en el que se extrajeron los datos (*)	Categórica ordinal
3	**PRICE**	Precio en idealista en euros (variable dependiente)	Numérica continua
4	UNITPRICE	Precio unitario (€/m^2)	Numérica continua
5	CONSTRUCTEDAREA	Área construida (m^2)	Numérica continua
6	ROOMNUMBER	Número de habitaciones	Numérica discreta
7	BATHNUMBER	Número de baños	Numérica discreta
8	HASTERRACE	Indicador de si la propiedad tiene terraza 1 – Tiene terraza 0 – No tiene terraza	Categórica dicotómica
9	HASLIFT	Indicador de si la propiedad tiene ascensor 1 – Tiene ascensor 0 – No tiene ascensor	Categórica dicotómica
10	HASAIRCONDITIONING	Indicador de si la propiedad tiene aire acondicionado 1 – Tiene aire acondicionado 0 – No tiene aire acondicionado	Categórica dicotómica
11	AMENITYID	Indica los servicios incluidos. 1 – Sin muebles, sin servicios de cocina 2 – Servicios de cocina, sin muebles 3 – Servicios de cocina, con muebles	Categórica ordinal
12	HASPARKINGSPACE	Indicador de si la propiedad tiene aparcamiento 1 – Tiene aparcamiento 0 – No tiene aparcamiento	Categórica dicotómica
13	ISPARKINGSPACEINCLUDEDINPRICE	Indicador de si el aparcamiento está incluido en el precio 1 – Está incluido 0 – No está incluido	Categórica dicotómica
14	PARKINGSPACEPRICE	Precio del aparcamiento (€)	Numérica continua
15	HASNORTHORIENTATION	Indicador de si la propiedad tiene orientación norte (**)	Categórica dicotómica
16	HASSOUTHORIENTATION	Indicador de si la propiedad tiene orientación sur (**)	Categórica dicotómica
17	HASEASTORIENTATION	Indicador de si la propiedad tiene orientación este (**)	Categórica dicotómica
18	HASWESTORIENTATION	Indicador de si la propiedad tiene orientación oeste (**)	Categórica dicotómica
19	HASBOXROOM	Indicador de si la propiedad tiene trastero 1 – Tiene trastero 0 – No tiene trastero	Categórica dicotómica
20	HASWARDROBE	Indicador de si la propiedad tiene armario empotrado 1 – Tiene armario empotrado 0 – No tiene armario empotrado	Categórica dicotómica
21	HASSWIMMINGPOOL	Indicador de si la propiedad tiene piscina 1 – Tiene piscina 0 – No tiene piscina	Categórica dicotómica
22	HASDOORMAN	Indicador de si la propiedad tiene portero 1 – Tiene portero 0 – No tiene portero	Categórica dicotómica
23	HASGARDEN	Indicador de si la propiedad tiene jardín 1 – Tiene jardín	Categórica dicotómica

		0 – No tiene jardín	
24	ISDUPLEX	Indicador de si la propiedad es un dúplex 1 – Es dúplex 0 – No es dúplex	Categórica dicotómica
25	ISSTUDIO	Indicador de si la propiedad es un estudio 1 – Es un estudio 0 – No es un estudio	Categórica dicotómica
26	ISINTOPFLOOR	Indicador de si la propiedad está en la última planta 1 – Está en la última planta 0 – No está en la última planta	Categórica dicotómica
27	CONSTRUCTIONYEAR	Año de construcción de la propiedad según el anunciante	Numérica discreta
28	FLOORCLEAN	Indicador de la planta en la que se encuentra la vivienda	Numérica discreta
29	FLATLOCATIONID	Indica el tipo de vistas que tiene el piso 1 - Externas 2 - Internas	Categórica dicotómica
30	CADCONSTRUCTIONYEAR	Año de construcción según el catastro	Numérica discreta
31	CADMAXBUILDINGFLOOR	Máxima planta del edificio según el catastro	Numérica discreta
32	CADDWELLINGCOUNT	Número de viviendas en el edificio según el catastro	Numérica discreta
33	CADASTRALQUALITYID	Identificador de la calidad catastral (0 mejor – 10 peor)	Numérica ordinal
34	BUILTTYPEID_1	Identificador del estado del piso (nueva construcción)	Categórica dicotómica
35	BUILTTYPEID_2	Identificador del estado del piso (segunda mano para restaurar)	Categórica dicotómica
36	BUILTTYPEID_3	Identificador del estado del piso (segunda mano en buen estado)	Categórica dicotómica
37	DISTANCE_TO_CITY_CENTER	Distancia de la propiedad al centro de la ciudad (km)	Numérica continua
38	DISTANCE_TO_METRO	Distancia de la propiedad al metro (km)	Numérica continua
39	DISTANCE_TO_CASTELLANA	Distancia de la propiedad al Paseo de la Castellana (km)	Numérica continua
40	LONGITUDE	Coordenada de longitud	
41	LATITUDE	Coordenada de latitud	

Tabla 1. Descripción de las variables

(*) Una vivienda puede encontrarse en más de un periodo cuando una propiedad puesta en venta en un trimestre se vendió en un trimestre posterior
(**) Las características de orientación no son ortogonales, una casa orientada al norte también puede estar orientada al este.

Con esta comprensión de las variables, procederemos con un análisis más detallado. Para ello, vamos a emplear estadísticos descriptivos y técnicas gráficas según el tipo de variable.

Sin embargo, antes de iniciar este proceso, es importante destacar que hemos tenido que dar el formato adecuado a las variables para el correcto procesamiento de los datos, ya que, a pesar de que muchas de las variables son categóricas, R las ha clasificado como numéricas. Esto nos permitirá realizar cálculos y visualizaciones precisas para examinar la distribución de las variables y comprender mejor el comportamiento de las viviendas en el mercado inmobiliario de Madrid.

Para empezar, como vemos en la descripción de la tabla, puede haber varias viviendas con el mismo identificador debido a que una vivienda puede encontrarse en más de un período cuando una propiedad puesta en venta en un trimestre se vendió en un trimestre posterior. Como no queremos valores duplicados de las viviendas y además no sabemos el valor real de cada una de ellas, procedemos a eliminar aquellas que estén duplicadas, en total son 19011, por lo que nos quedaríamos con un total de 75804 observaciones en nuestro dataset.

Variables numéricas

A continuación, vamos a calcular los principales estadísticos descriptivos para las variables numéricas de nuestro dataset.

Variable	Mínimo	Q1	Mediana	Media	Q3	Máximo	NA's
PRICE	21000	157000	257000	389544	457000	8133000	0
UNITPRICE	805.3	2206.7	3448.3	3641.6	4734	9997.6	0
CONSTRUCTEDAREA	21	62	82	100.3	115	985	0
PARKINGSPACEPRICE	1	1	1	666.1	1	925001	0
CONSTRUCTIONYEAR	1	1955	1968	1964	1986	2291	44574
FLOORCLEAN	-1	1	2	2.745	4	11	3114
CADCONSTRUCTIONYEAR	1623	1955	1967	1965	1983	2018	0
CADMAXBUILDINGFLOOR	0	5	6	6.361	8	26	0
CADDWELLINGCOUNT	1	12	21	38.72	39	1499	0
DISTANCE_TO_CITY_CENTER	0.0154	2.4037	4.1143	4.4822	6.2195	415.7526	0
DISTANCE_TO_METRO	0.0014	0.2128	0.3300	0.4754	0.5192	339.4774	0
DISTANCE_TO_CASTELLANA	0.0014	1.0405	1.9621	2.6813	3.8542	412.8037	0

Tabla 2. Estadísticos descriptivos variables continuas

De esta tabla podemos sacar las siguientes conclusiones:

- **Valores extremos**: encontramos valores extremos o atípicos en algunas variables, como el precio máximo del espacio de estacionamiento (925001), el año de construcción (2291) o el número máximo de pisos según el catastro.

- **Inconsistencias en los datos**: la variable "parkingspaceprice" parece tener valores incorrectos, ya que la mayoría de los registros tienen valor uno. Esto sugiere que podría ser mejor eliminar esta variable.

- **Datos inverosímiles**: la variable año de construcción muestra un mínimo de uno. Esto no puede ser posible ya que no puede haber una propiedad construida en el año uno. Además, tiene un gran número de valores faltantes, por lo podríamos considerar eliminarla y utilizar la variable año de construcción según el catastro en su lugar, ya que esta sí está completa.

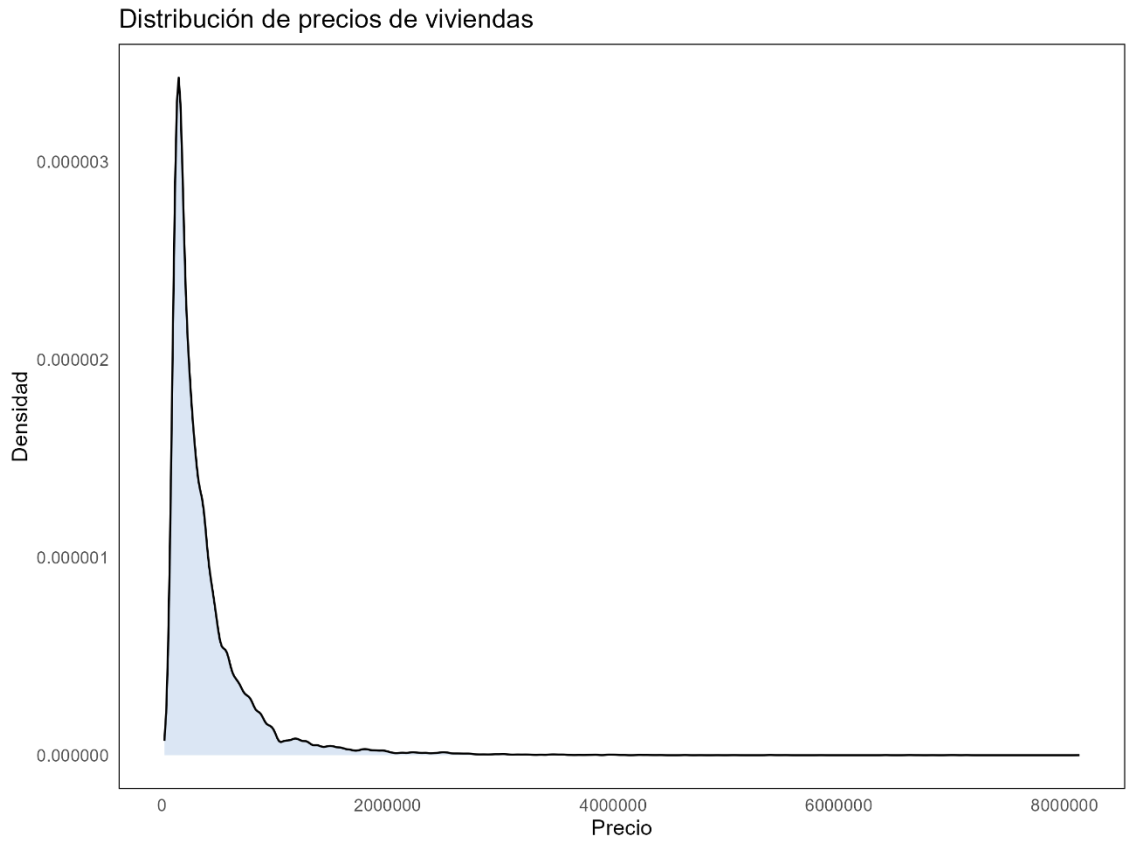


Figura 1. Distribución de los precios de viviendas

El comportamiento de los precios de las viviendas es crucial para entender el mercado inmobiliario y para el análisis que estamos llevando a cabo.

Como podemos observar en el gráfico, podemos destacar varios puntos importantes:

- La distribución de los precios no parece seguir una distribución normal, es asimétrica y sesgada hacia la derecha. Esto significa que hay más viviendas con precios más bajos y menos viviendas con precios más altos.

- El 75% de las viviendas tienen precios comprendidos entre 21.000 y 457.000 euros. Este rango cubre una amplia variedad de precios, desde viviendas más económicas hasta propiedades de precio medio.

- Es posible que haya valores atípicos o outliers en el extremo derecho del gráfico, representando viviendas con precios significativamente más altos que la mayoría. Estos valores atípicos pueden indicar viviendas de lujo u otros casos excepcionales que podrían distorsionar la distribución de los precios.

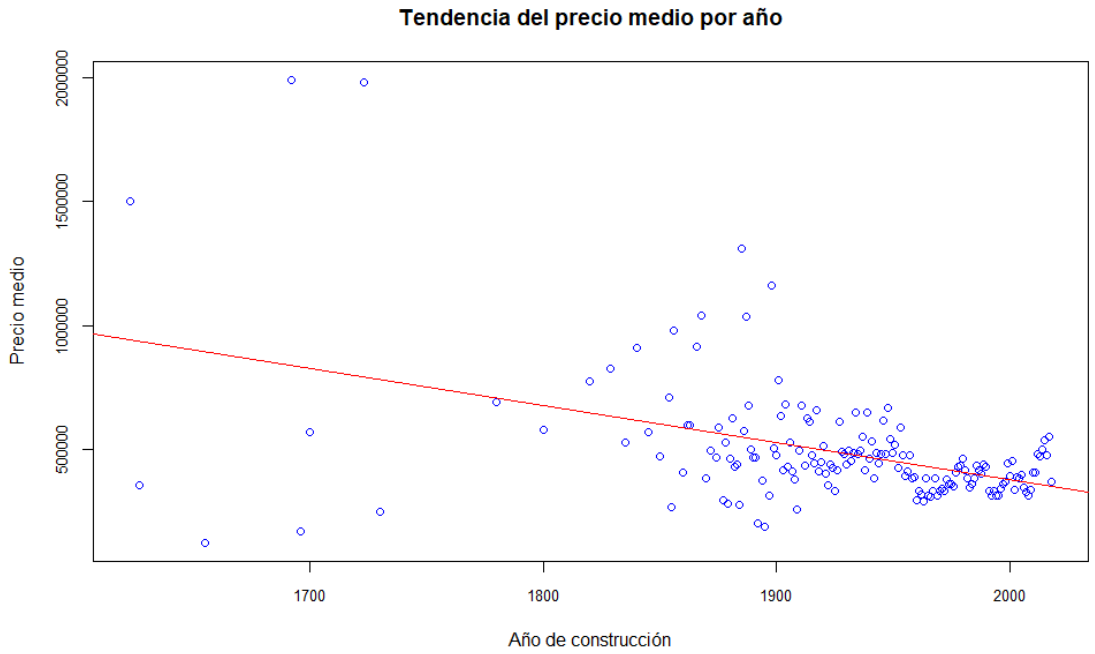


Figura 2. Gráfico de dispersión del precio medio por año

En este gráfico de dispersión podemos ver cómo se comporta el precio en relación con la antigüedad de las viviendas. Esto resulta interesante ya que de esta manera nos permite comprender si los precios han disminuido o no con el tiempo.

En este caso, vemos claramente una disminución del precio a lo largo del tiempo. Sin embargo, al estar representando el precio medio, los resultados podrían estar sesgados debido a una muestra muy pequeña de viviendas antiguas.

Variables categóricas

Para las variables categóricas, utilizaremos gráficos de barras para visualizar la distribución de las diferentes categorías.

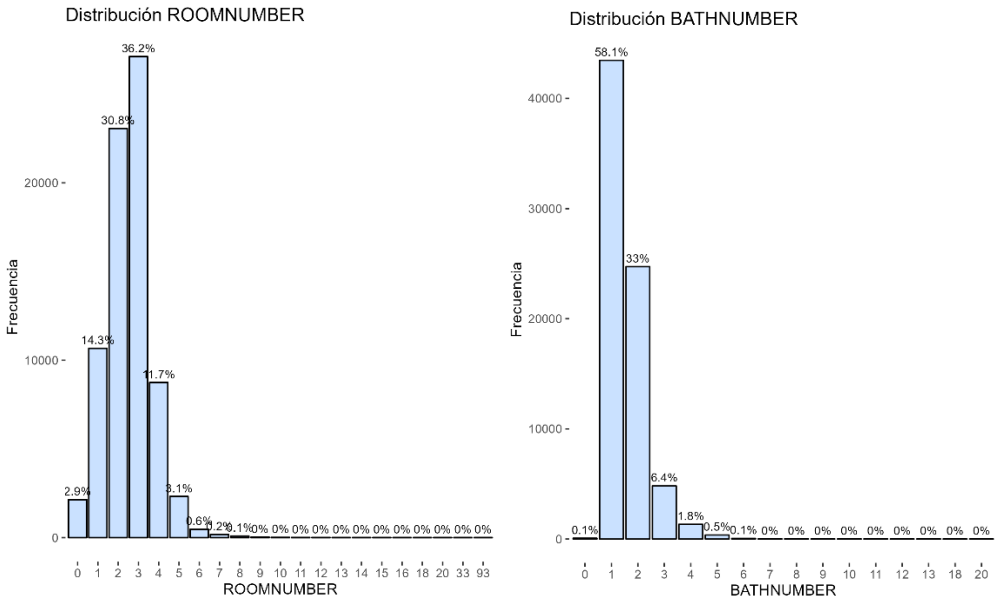


Figura 3. Frecuencias del número de habitaciones y baños

En estos gráficos observamos:

- **Número de habitaciones**: observamos que hay viviendas con 0 habitaciones, pero esto es posible ya que un 94.5 % de estas observaciones son estudios. Los estudios son pequeñas viviendas que combinan espacio de estar, dormitorio y a menudo cocina en una sola área. Esta variable podría estar altamente correlacionada con "isstudio" debido a este alto porcentaje. Para nuestro análisis, podríamos considerar eliminar aquellas viviendas que tengan más de 8 habitaciones, ya que esto es muy poco común.

- **Número de baños**: observamos algunas viviendas con 0 baños, por lo que podríamos eliminarlas, ya que es muy poco habitual que una vivienda no tenga baño, y al igual que con el número de habitaciones, podríamos eliminar aquellas con más de 8 baños.

Estos hallazgos sugieren que podría ser beneficioso filtrar las viviendas con números extremos de habitaciones y baños para mejorar la calidad de los datos y garantizar que reflejen con precisión las características típicas de las viviendas.

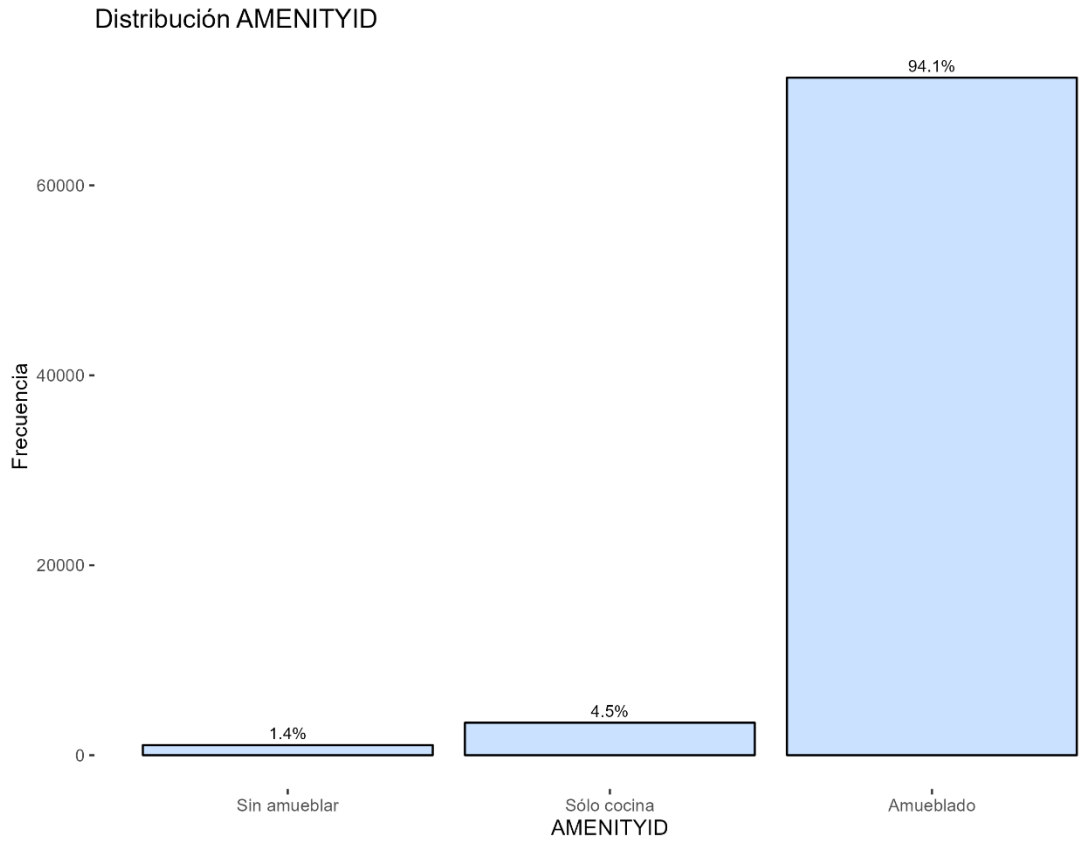


Figura 4. Frecuencias del nivel de amueblado

En el gráfico de distribución del estado de amueblamiento de las viviendas, observamos una clara predominancia de las viviendas amuebladas, que representan el 94.1% del total. Esto indica que la gran mayoría de las viviendas están completamente amuebladas.

Por lo general, las viviendas completamente amuebladas podrían tener un precio más alto debido a la comodidad y la reducción de gastos iniciales, lo que podría influir significativamente en nuestro análisis. Sin embargo, la baja proporción de viviendas sin amueblar (1.4%) y de aquellas con sólo la cocina amueblada (4.5%) plantea una limitación importante. La escasa representación de estas categorías podría dificultar la extracción de conclusiones robustas sobre su impacto en los precios de las viviendas.

Esta disparidad en la representación puede afectar la capacidad del modelo para generalizar sus predicciones a las tres categorías de amueblamiento. Para abordar esto, podríamos prescindir de esta variable.

En resumen, aunque el estado de amueblamiento debe ser considerado importante para el modelo de predicción, la baja representación de viviendas sin amueblar y con sólo la cocina puede limitar la precisión del análisis en estas categorías.

Variable	n	Variable	n
ROOMNUMBER		**BATHNUMBER**	
0	2149	1	43480
1	10681	2	24799
2	23073	3	5033
3	27224	4	1659
4	9066	5	570
5	2578	6	117
Other	1033	Other	146
HASTERRACE		**HASLIFT**	
0	49259	0	23550
1	26545	1	52254
HASPARKINGSPACE		**ISPARKINGSPACEINCLUDEDINPRICE**	
0	59307	0	59307
1	16497	1	16497
HASBOXROOM		**HASDOORMAN**	
0	56853	0	56983
1	18951	1	18821
ISSTUDIO		**AMENITYID**	
0	73752	1	1055
1	2052	2	3409
HASAIRCONDITIONING		**3**	71340
0	42546	**HASNORTHORIENTATION**	
1	33258	0	67875
HASSOUTHORIENTATION		1	7929
0	58273	**HASWESTORIENTATION**	
1	17531	0	64985
HASEASTORIENTATION		1	10819
0	60935	**HASSWIMMINGPOOL**	
1	14869	0	65029
HASWARDROBE		1	10775
0	33072	**ISDUPLEX**	
1	42732	0	73887
HASGARDEN		1	1917
0	62410	**FLATLOCATIONID**	
1	13394	1	60912
ISINTOPFLOOR		2	9730
0	74171	Na’s	5162
1	1633	**CADASTRALQUALITYID**	
BUILTTYPEID_1		**3**	9807
0	73388	4	19594
1	2416	5	16678
BUILTTYPEID_2		6	16596
0	61646	7	8509
1	14158	Other	4619
BUILTTYPEID_3		Na’s	1
0	16574		
1	59230		

Tabla 3. Distribución del número de categorías de las variables

Este análisis descriptivo aporta una base sólida para entender la distribución y las características de las viviendas de Madrid. Con esta información, podemos proceder a la depuración de los datos.

3.3. Depuración de los datos

Una vez concluido el análisis descriptivo de las variables que componen nuestro dataset, procedemos con la depuración de datos.

Vamos a empezar eliminando la variable año de construcción ya que tiene un porcentaje muy alto de valores missing y podemos utilizar en su lugar el año de construcción según el catastro, que está más completo. Eliminamos la variable precio del espacio de estacionamiento y "AMENITYID" ya que no nos servirán en nuestro análisis del precio de la vivienda. También eliminamos las variables ASSETID y PERIOD ya que no aportan información útil a la hora de predecir el precio de la vivienda. Esto nos ayudará a evitar sesgos y reducir la complejidad del modelo.

Para nuestro análisis, nos quedaremos con aquellas viviendas que su precio sea menor a 2.000.000€ para eliminar aquellas que puedan considerarse viviendas de lujo, lo que no asegurará que los resultados sean representativos del mercado general.

Una vez hecho esto, pasamos al tratamiento de valores missing. Para ello, hemos visto que prácticamente ninguna de nuestras variables tenía, únicamente "FLOORCLEAN" con 3114 observaciones missing, "FLATLOCATIONID" con 5162, y cadastralqualityid que tiene una. Para tratar estos valores, decidimos eliminar estas observaciones debido a que la cantidad de datos missing es relativamente pequeña y no afectará significativamente nuestro análisis, dada la gran cantidad de observaciones que tenemos.

A continuación, pasamos a eliminar aquellas observaciones con 0 baños y las que tengan un número de baños y habitaciones mayor de 8, ya que son características inusuales. Además, agruparemos aquellas las que tengan 5 o más habitaciones y 5 o más baños para que así sean más representativas y podamos asegurar que cada grupo tenga un tamaño de muestra suficiente para proporcionar estimaciones robustas.

También agruparemos la variable "cadastralqualityid" en dos grupos, 0_4 y 5_9. De esta manera, conseguiremos simplificar los parámetros de los futuros modelos y reducir la dimensionalidad, lo que resulta útil para mejorar la interpretabilidad.

Una vez corregidos los datos y tratado los valores missing, podemos pasar a la búsqueda y tratamiento de datos atípicos excluyendo la variable objetivo. Lo haremos únicamente de las variables numéricas, ya que de las demás no es necesario al ser variables binarias o haberlas agrupado.

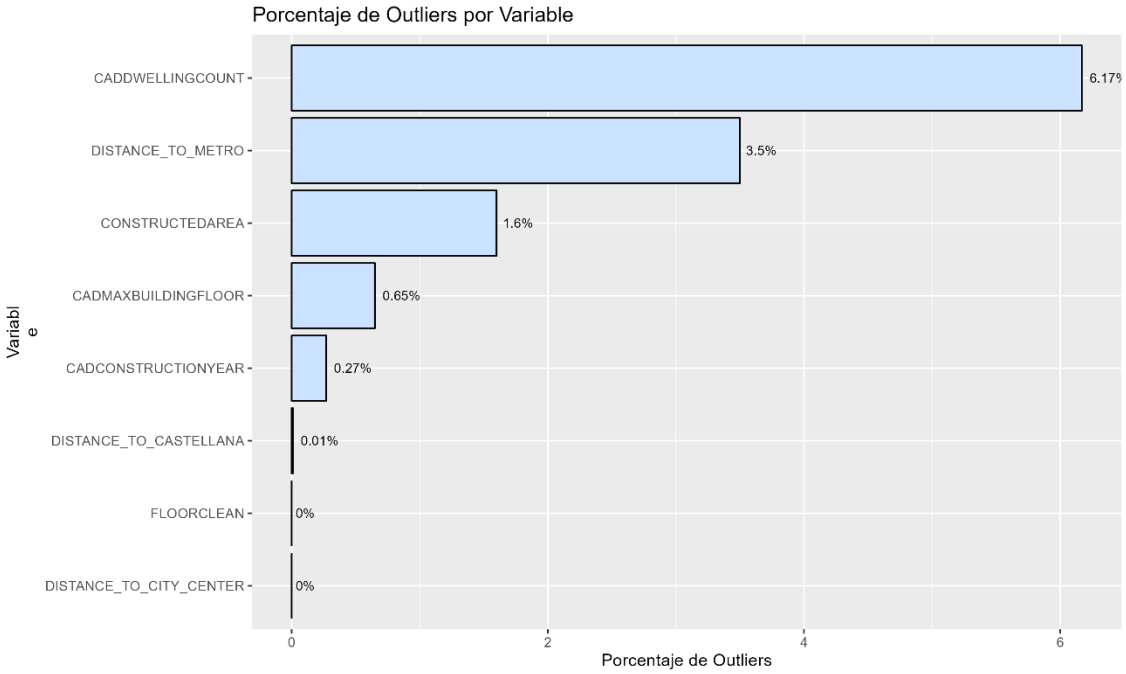


Figura 5. Porcentaje de outliers para las variables numéricas

Este gráfico muestra el porcentaje de valores atípicos de nuestro conjunto de datos, por lo que hemos eliminado todas aquellas observaciones que contenían alguno. Con esto, tendríamos nuestros datos limpios con un total de 60161 observaciones.

Antes de pasar al análisis exploratorio de datos espaciales, es importante evaluar la correlación de las variables numéricas, para ello calculamos la matriz de correlaciones:

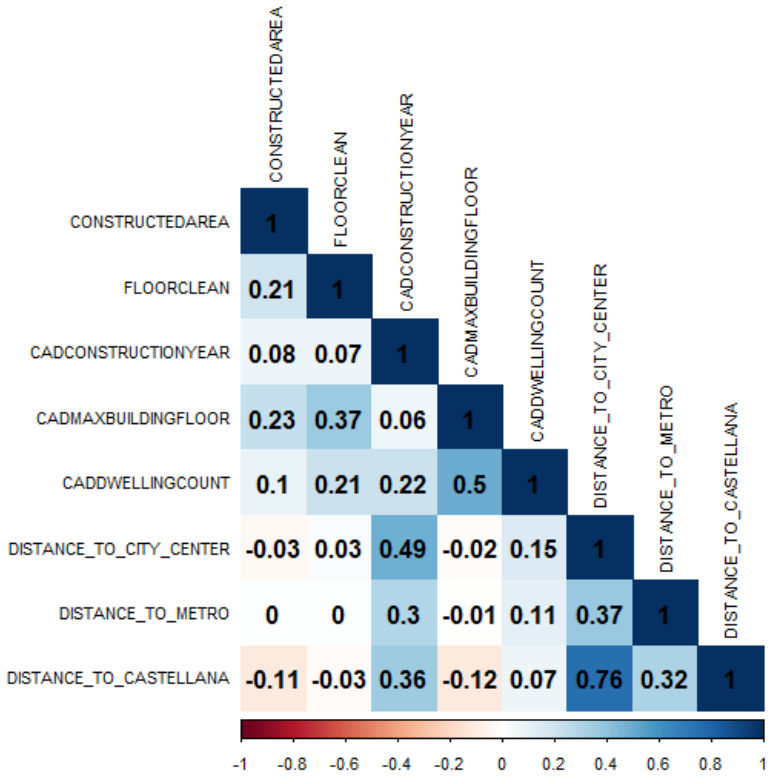


Figura 6. Matriz de correlaciones para las variables numéricas

Como podemos observar, las variables con una correlación más alta son distancia a la castellana y distancia al centro de la ciudad, con un valor de 0.76. Por lo general las variables no tienen una correlación muy fuerte por lo que no sería necesario eliminar ninguna.

Debido a que tenemos una alta cantidad de observaciones y que podría afectar a la eficiencia de los modelos, vamos a proceder a realizar un muestreo aleatorio simple sobre todos los datos para quedarnos con una muestra de 3.000 viviendas. El objetivo es reducir el tamaño del conjunto de datos manteniendo la representatividad del total.

Con esto podríamos dar por concluido el proceso de depuración de datos y pasar al análisis exploratorio de datos espaciales.

4. Análisis exploratorio de datos espaciales

El análisis exploratorio de datos espaciales (AEDE) es una fase fundamental en cualquier análisis espacial. Permite comprender las relaciones espaciales entre las variables, así como explorar patrones geográficos en nuestros datos. A través del AEDE, podemos identificar tendencias y anomalías que podrían influir en el análisis posterior y la toma de decisiones.

Antes de empezar con el análisis exploratorio con nuestra muestra de datos, vamos a representar el precio medio por unidad de metro cuadrado en los 21 distritos de Madrid para hacernos una idea inicial de cómo se distribuyen los precios.

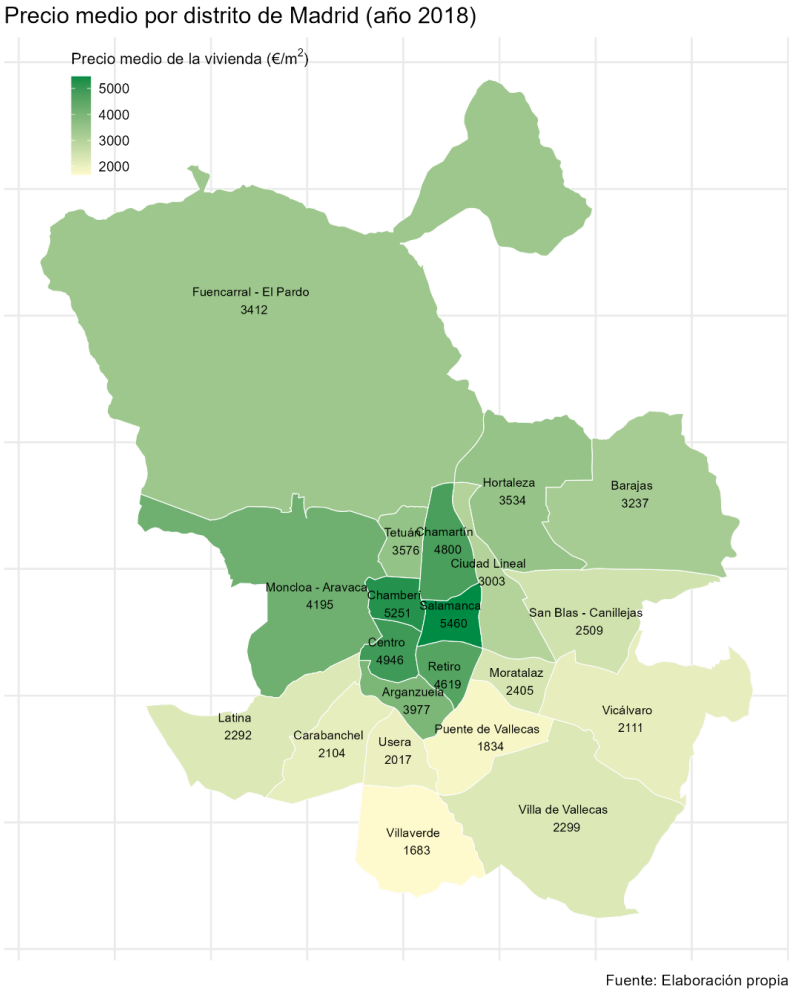


Figura 7. Mapa coroplético del precio medio de viviendas por distrito

En este mapa geográfico se ha representado el precio medio por metro cuadrado para los diferentes distritos de Madrid. Como podemos observar, cuanto más al sur, las viviendas tienden a tener precios más bajos, y cuanto más por el centro/norte los precios tienden a ser más altos. Por ejemplo, en el distrito de Salamanca el precio medio se encuentra en 5460€/m^2, y en Villaverde el precio medio se encuentra en 1683€/m^2. Esto puede deberse a factores como la accesibilidad a servicios o características socioeconómicas de las áreas. Además, es posible que existan clusters de precios más altos en ciertas zonas céntricas o al norte de Madrid, lo cual sugiere una mayor demanda o la presencia de barrios más prestigiosos.

Otra forma de visualizar la distribución de los precios, es con un mapa interactivo de la siguiente manera:

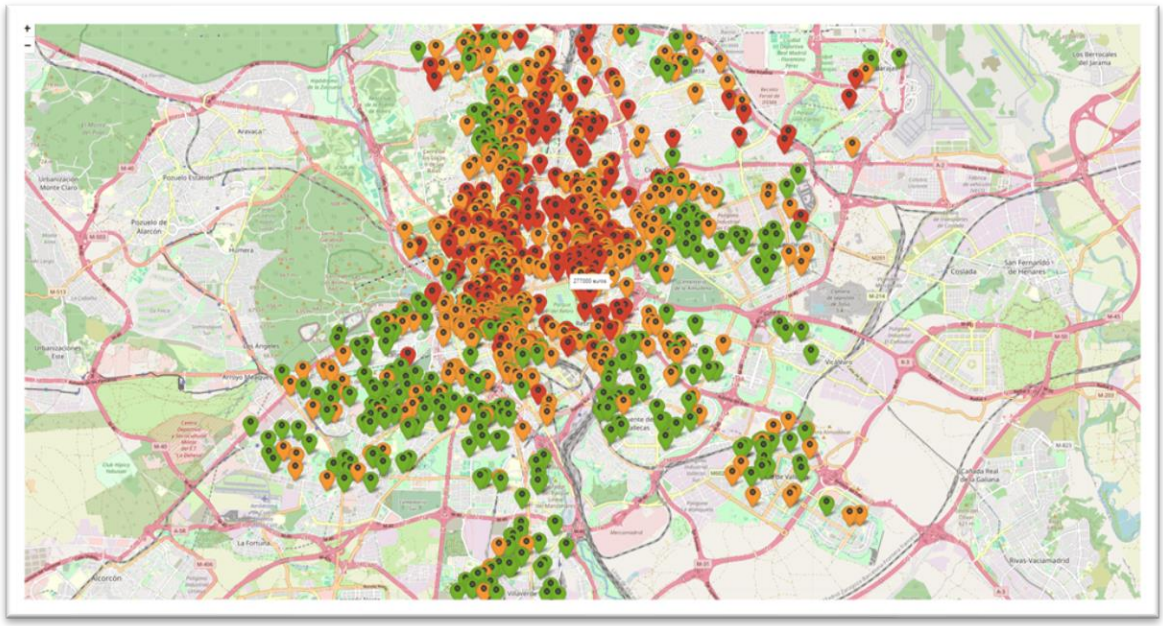


Figura 8. Captura mapa interactivo. Distribución de precios en Madrid

Como podemos ver, al igual que en el anterior mapa, vemos que los precios más altos (puntos rojos) tienden a estar más en el centro de la ciudad.

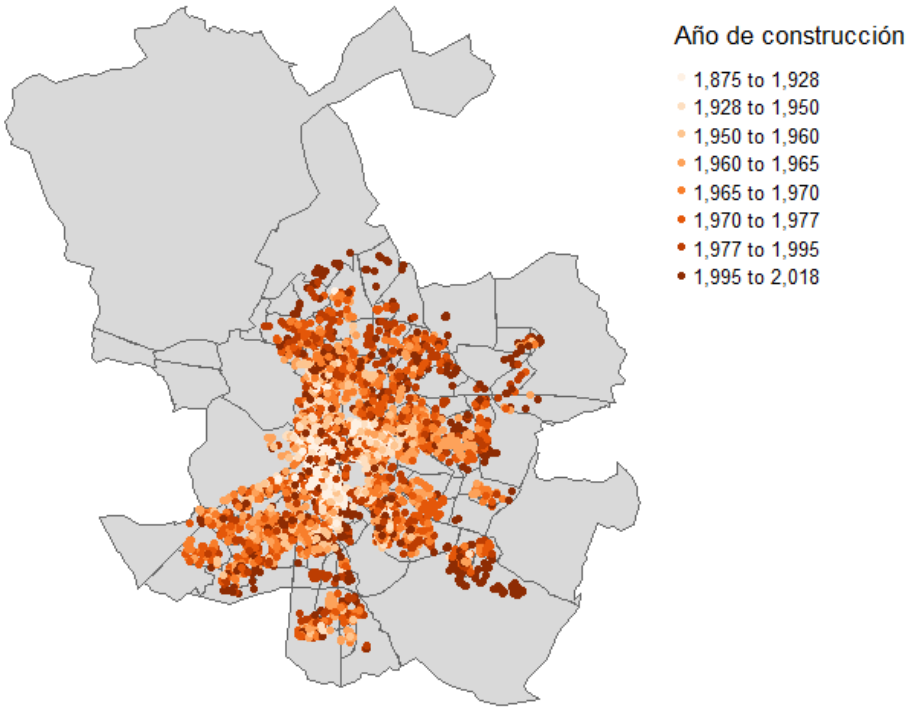


Figura 9. Distribución de viviendas en función del año de construcción

Este mapa muestra la distribución de las viviendas en función del año de construcción. Como podemos observar, las viviendas más antiguas tienden a estar ubicadas en el centro de la ciudad, lo que resulta lógico al irse expandiendo hacia la periferia con el tiempo.

Aunque los precios de las viviendas en el centro son generalmente más altos, la correlación entre el año de construcción y los precios es de -0.0678. Esto indica que, a pesar de las tendencias generales, no existe una relación fuerte entre el año de construcción y los precios de las viviendas. Factores como la ubicación, la superficie y el estado de la vivienda también influyen significativamente en los precios.

Los outliers espaciales son puntos de datos que se desvían significativamente de los patrones esperados en un contexto geográfico. Identificar estas localizaciones atípicas es importante para entender mejor las anomalías y posibles errores en los datos, así como para detectar fenómenos excepcionales.

Para identificar estas observaciones, primero debemos definir las relaciones de vecindad. Además, estas relaciones de vecindad nos ayudarán a capturar la dependencia espacial. Utilizamos la distancia euclidiana para determinar la proximidad entre puntos geográficos como es en este caso, al estar tratando con viviendas, especificamos un rango entre 0 km a 0.5 km. Esto significa que consideramos vecinos a aquellas viviendas que se encuentren a una distancia de hasta 0.5 km entre sí. Identificamos las viviendas que no tengan vecinos dentro del rango especificado y filtraremos eliminándolas de nuestra muestra para evitar que influyan en el análisis posterior.

A continuación, realizaremos una transformación calculando el cuadrado de la inversa a la distancia.

Esta transformación nos ayuda a capturar la idea de que la influencia entre puntos geográficos

disminuye con la distancia. Así, los vecinos más cercanos tienen mayor peso y los más distantes menor peso.

Matriz de pesos espaciales: Construimos una matriz donde cada elemento w_{ij} representa el peso espacial entre las viviendas i y j. Si la distancia d_{ij} entre i y j es menor o igual a 0.5 km, entonces $w_{ij} = 1/d_{ij}^2$, de lo contrario $w_{ij} = 0$.

Esta matriz nos permite modelar la dependencia espacial entre los datos, permitiéndonos capturar la relación entre los puntos adecuadamente para el análisis posterior.

Autocorrelación espacial

Uno de los efetos espaciales más importantes es la autocorrelación espacial. Como comentábamos en el gráfico de los precios de la vivienda, parecía que cuanto más en el centro los precios eran más elevados. El análisis de autocorrelación espacial global analiza las observaciones para determinar si una variable se encuentra distribuida de forma aleatoria en el espacio o si, por el contrario, existe un patrón espacial (una asociación significativa de valores similares o disimiles entre puntos vecinos).

A continuación, vamos a contrastar la hipótesis nula de "ausencia de autocorrelación espacial" con tres diferentes estadísticos globales:

1. Índice I de Moran:

Moran I statistic	Moran I statistic standard deviate	p-value
0.5484	41.448	< 2.2e-16

Tabla 4. Resultados prueba del índice I de Moran

   - **Índice I de Moran (**0.5484**):** Este valor positivo indica una autocorrelación espacial positiva, es decir, los precios de las viviendas en ubicaciones cercanas tienden a ser similares.
   - **Estadístico estándar (**41.448): Tiene un valor extremadamente alto, lo que indica una fuerte evidencia contra la hipótesis nula de ausencia de autocorrelación espacial.
   - **P-valor (< 2.2e-16):** el p-valor aproximadamente 0 indica que la autocorrelación espacial observada es significativa con un nivel de confianza muy alto.

2. Prueba C de Geary:

Geary C statistic	Geary C statistic standard deviate	p-value
0.42518	34.069	< 2.2e-16

Tabla 5. Resultados prueba C de Geary

   - **Geary C statistic (0.42518):** Un valor menor que 1 indica autocorrelación espacial positiva, lo que sugiere que los precios de viviendas en ubicaciones cercanas son similares.
   - **Estadístico estándar (34.069):** Un valor muy alto, lo que proporciona una

fuerte evidencia contra la hipótesis nula.

- **P-valor (< 2.2e-16):** Al igual que con el índice de Moran, un p-valor muy bajo indica autocorrelación significativa.

3. Prueba G(d) De Getis y Ord:

Global G statistic	Global G statistic standard deviate	p-value
0.000446	31.963	< 2.2e-16

Tabla 6. Resultados prueba G(d) de Getis y Ord

- **G(d) statistic (0.000446):** Este valor cercano a cero sugiere que las ubicaciones con altos valores de precios están agrupadas cerca de otras ubicaciones con altos valores, lo que indica una autocorrelación espacial positiva.
- **Estadístico estándar (31.963):** Un valor muy alto, lo que proporciona evidencia contundente contra la hipótesis nula de ausencia de agrupamiento espacial.
- **p-valor (< 2.2e-16):** P-valor aproximadamente 0 indica autocorrelación significativa, confirmando que los patrones espaciales observados no son aleatorios.

En resumen, los tres estadísticos globales utilizados para contrastar la hipótesis nula de "ausencia de autocorrelación espacial" en los precios de las viviendas indican de manera consistente la presencia de autocorrelación espacial positiva.

- Índice I de Moran: confirma la agrupación de precios similares en ubicaciones cercanas.
- Prueba C de Geary: reafirma la similitud de valores de precios en áreas vecinas.
- Prueba G(d) de Getis y Ord: indica la agrupación de valores altos de precios en proximidad espacial.

Dado que todos los estadísticos tienen p-valores extremadamente bajos, podemos rechazar con un alto nivel de confianza la hipótesis nula y concluir que existe un patrón significativo de autocorrelación espacial en los precios. Esto sugiere que los factores globales y la proximidad espacial son determinantes importantes en la distribución de los precios de las viviendas en el área estudiada, en este caso de Madrid.

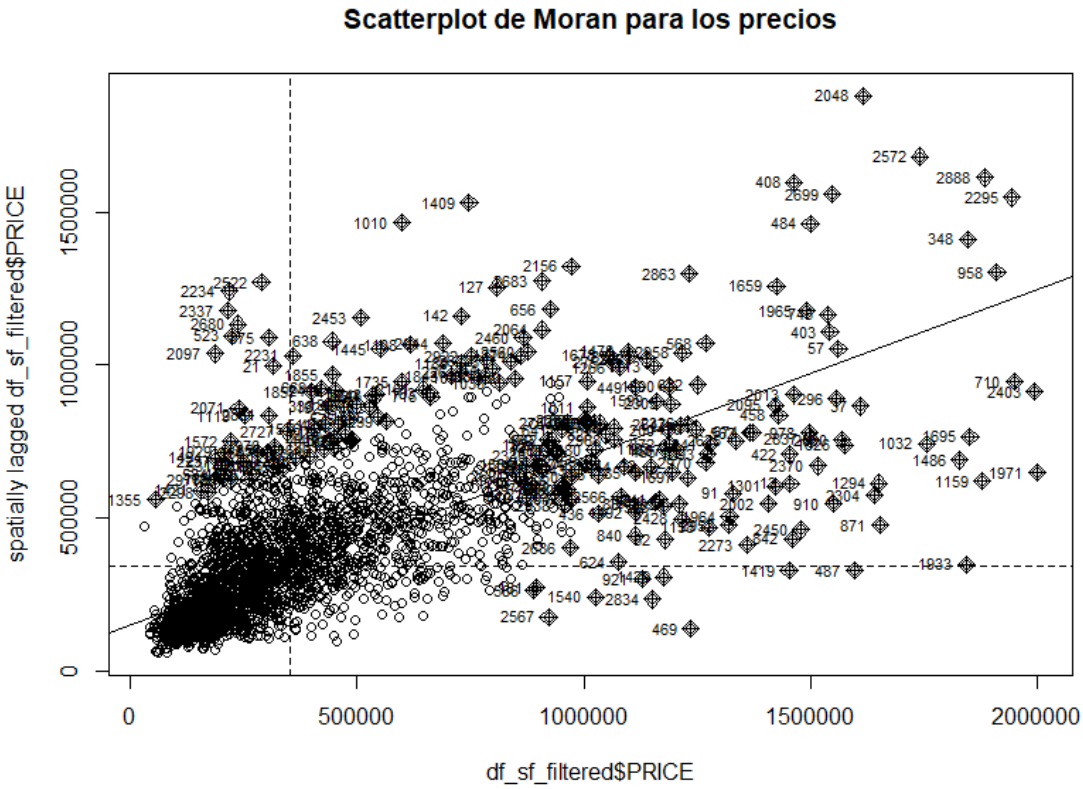


Figura 10. Diagrama de dispersión de Moran para los precios

El scatterplot de Moran es una herramienta útil para identificar y visualizar la dependencia espacial, proporcionando una comprensión más profunda de cómo los precios de las viviendas se agrupan espacialmente. Este gráfico muestra la relación entre los valores observados y los valores lag (promedio ponderado de los valores vecinos).

Retardo espacial

El retardo espacial de los precios indica la correlación entre los precios de las viviendas en un área determinada y los precios de las viviendas en las áreas adyacentes. Calculamos una nueva variable llamada “PRICEW” que contendrá estos valores retardados.

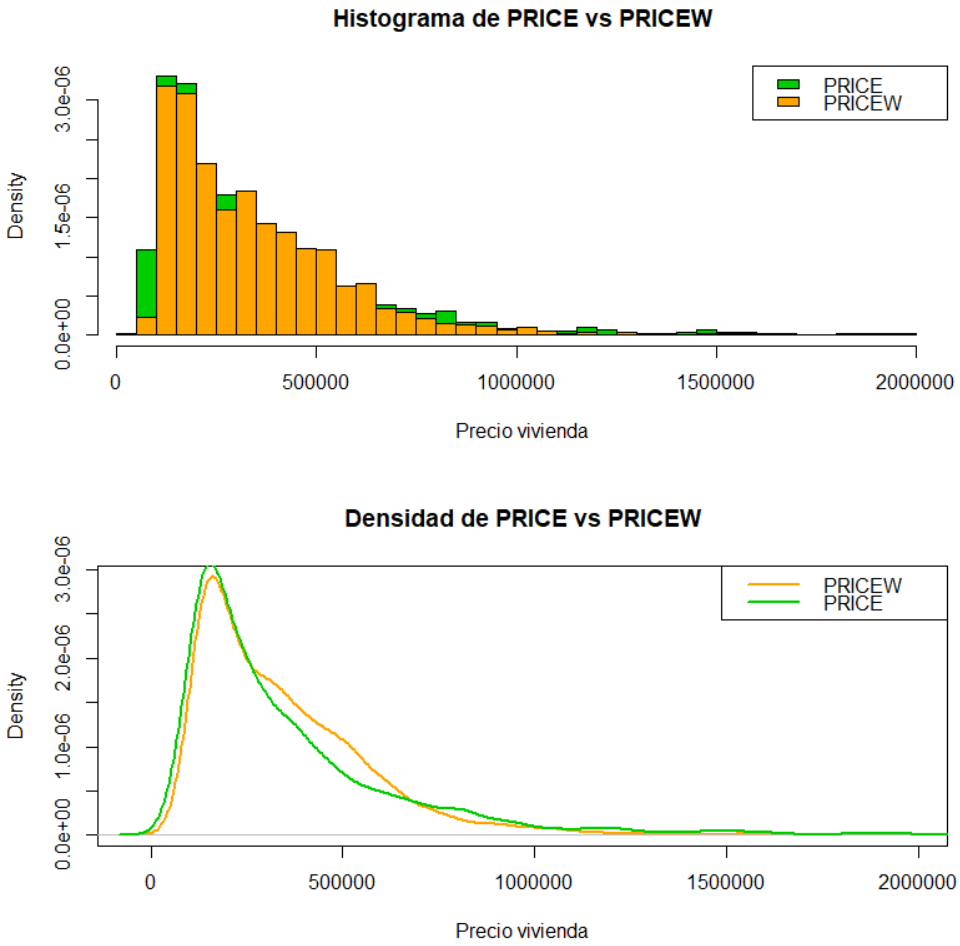


Figura 11. Gráficos de histograma y densidad para los precios y los precios retardados

Podemos observar en el histograma y la distribución de densidad una clara similitud entre las dos variables. Esto sugiere que los precios de las viviendas en un área están estrechamente relacionados con los precios en las áreas adyacentes. Esto es consistente con la existencia de autocorrelación espacial positiva que vimos anteriormente, donde las áreas cercanas tienen precios similares.

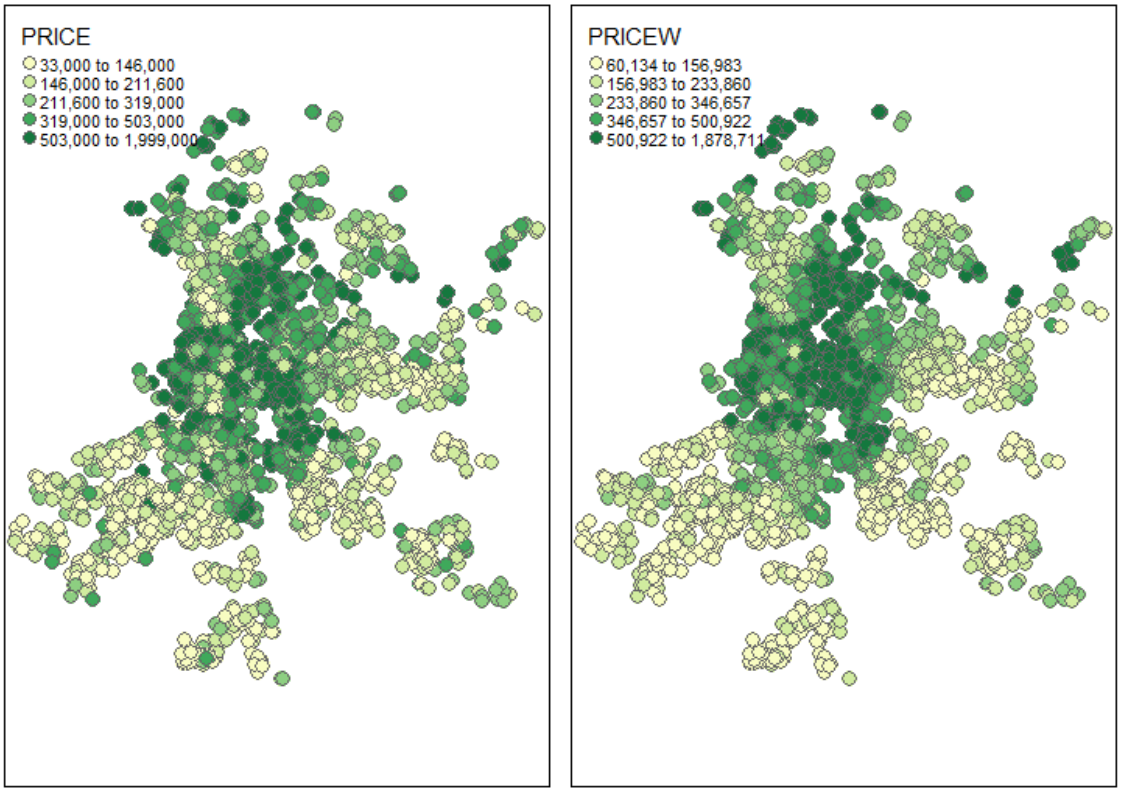


Figura 12. Representación de los precios y los precios retardados

Otra forma de representar los retardos es con puntos geográficos. El retardo espacial de los precios nos permite obtener la media móvil, que considera tanto la influencia local como su influencia en los vecindarios cercanos. Esto suaviza la distribución de los precios, proporcionando una mejor comprensión de las tendencias y patrones espaciales en los precios de la vivienda, tal y como vemos en esta representación. Este enfoque es especialmente útil para identificar influencias locales y vecinales, mejorando la precisión de las estimaciones de precios.

Los indicadores locales de asociación espacial o indicadores LISA (Local Indicator of Spatial Association), es un indicador cuyo objetivo principal es proyectar la asociación espacial local. De esta forma podemos ver representado en un mapa los clústeres formados y los valores significativamente altos de las variables “high-high” y “low-low”. Las variables “alto-bajo” y “bajo-alto” representar los clústeres compuestos por outliers o atípicos espaciales.

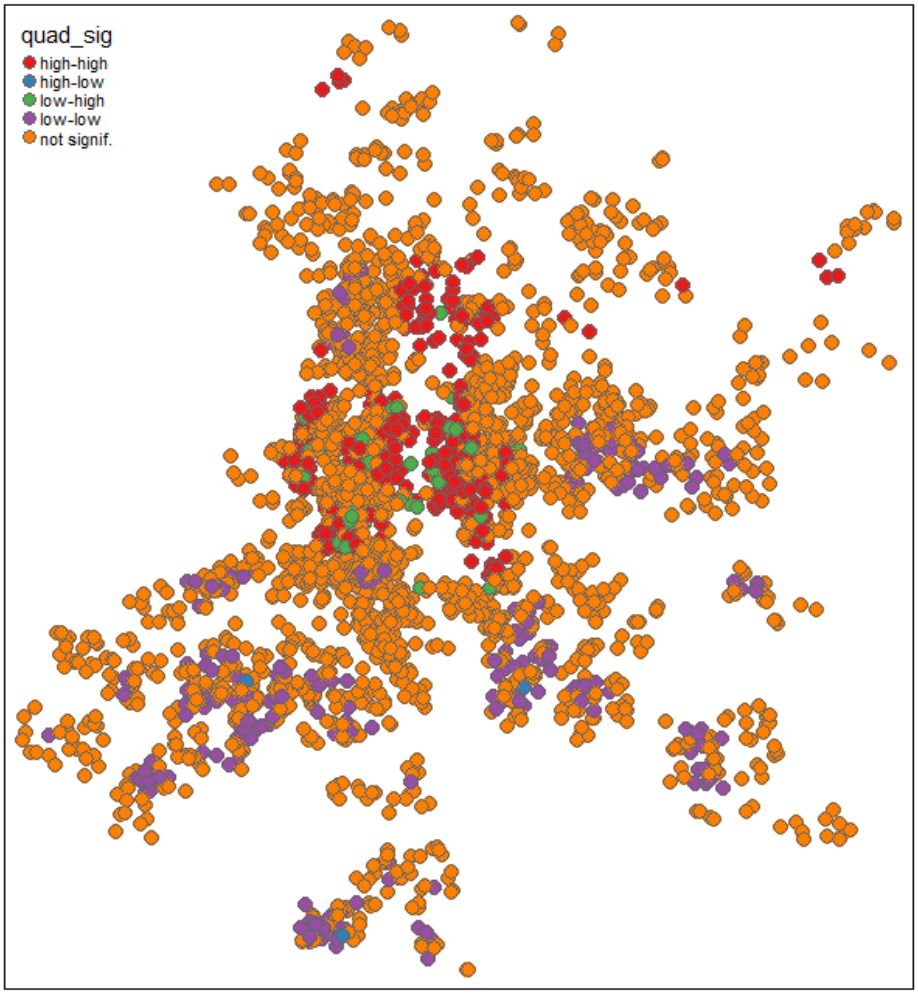


Figura 13. Mapa clúster LISA de los precios para Madrid

En este gráfico observamos que tenemos puntos rojos “high-high” por el centro y algunos por el norte, lo que significa que hay una dependencia espacial positiva significativa, con precios altos cerca de precios altos. Asimismo, en el sur y este, los puntos morados indican precios bajos cerca de precios bajos. Por el contrario, vemos muy pocos puntos verdes “high-low” y “low-high”, lo que indica que hay pocos outliers espaciales. La presencia de muchos puntos no significativos sugiere que, en gran parte del área estudiada, no hay una asociación espacial clara.

En resumen, el AEDE nos ha proporcionado una comprensión detallada de la distribución espacial de los precios de las viviendas en Madrid, identificando patrones significativos y dependencias espaciales que serán cruciales para el análisis posterior.

5. Modelado de datos espaciales

5.1. Modelo de regresión sin efectos espaciales

Para empezar, crearemos un modelo MCO (Mínimos Cuadrados Ordinarios) con todas las variables disponibles. Este modelo servirá como punto de partida para estimar los coeficientes de los parámetros y determinar la significancia de cada variable.

Posteriormente aplicaremos el método "stepwise" para eliminar todas aquellas variables que no sean significativas, obteniendo así un modelo más sencillo y eficiente. Tras el ajuste, obtendremos un modelo de regresión sin efectos espaciales simplificado, que tiene las siguientes características:

Coeficientes	Estimación	Std. Error	Pr(>|z|)
Intercept	*1425530.38*	*199471.21*	*1.12e-12 ****
CONSTRUCTEDAREA	*4284.13*	*98.32*	*< 2e-16 ****
ROOMNUMBER1	*24037.33*	*14880.98*	*0.106351*
ROOMNUMBER2	*7641.04*	*14555.70*	*0.599656*
ROOMNUMBER3	*-22126.60*	*15112.01*	*0.143253*
ROOMNUMBER4	*-85998.59*	*17270.41*	*6.74e-07 ****
ROOMNUMBER5 or more	*-103510.30*	*22754.47*	*5.61e-06 ****
BATHNUMBER2	*24265.92*	*7035.76*	*0.000571 ****
BATHNUMBER3	*177088.68*	*13895.75*	*< 2e-16 ****
BATHNUMBER4	*306900.10*	*27150.38*	*< 2e-16 ****
BATHNUMBER5 or more	*361609.31*	*75016.44*	*1.50e-06 ****
HASTERRACE1	*-21514.30*	*5324.84*	*5.47e-05 ****
HASLIFT1	*46286.49*	*6219.88*	*1.29e-13 ****
HASAIRCONDITIONING1	*17413.26*	*5178.41*	*0.000782 ****
HASSOUTHORIENTATION1	*-13675.53*	*5659.96*	*0.015744 **
HASSWIMMINGPOOL1	*40557.60*	*11182.14*	*0.000292 ****
HASDOORMAN1	*48552.76*	*6372.06*	*3.40e-14 ****
HASGARDEN1	*-20766.78*	*9122.70*	*0.022894 **
ISDUPLEX1	*-41625.53*	*17596.72*	*0.018068 **
FLOORCLEAN	*6270.94*	*1168.98*	*8.74e-08 ****
CADCONSTRUCTIONYEAR	*-702.92*	*102.89*	*1.01e-11 ****
CADMAXBUILDINGFLOOR	*-1876.46*	*1227.02*	*0.126300*
CADWELLINGCOUNT	*-346.98*	*123.36*	*0.004946 ***
CADASTRALQUALITYID5_9	*-56987.33*	*5715.77*	*< 2e-16 ****
BUILTTYPE_11	*47760.73*	*16924.46*	*0.004804 ***
BUILDTYPE_21	*-25647.51*	*6292.36*	*4.70e-05 ****
DISTANCE_TO_CITY_CENTER	*-6598.29*	*1667.89*	*7.80e-05 ****
DISTANCE_TO_METRO	*-65192.49*	*11335.87*	*9.78e-09 ****
DISTANCE_TO_CASTELLANA	*-16396.22*	*1945.11*	*< 2e-16 ****
Multiple R-squared		*0.8005*	
Adjusted R-squared:		*0.7986*	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1			

Tabla 7. Resultados modelo MCO

El modelo MCO nos proporciona información valiosa sobre cómo influye cada variable a la hora de predecir el precio de la vivienda. Algunas variables parecen ser más importantes que otras, como vemos por la significancia de cada una. Además, el R^2 tiene un valor de 0.8005, lo cual quiere decir que el modelo y las variables independientes explican un 80.05% de la variabilidad total de la variable dependiente, es decir, del precio.

Debido a que el análisis exploratorio de los precios ha revelado la existencia de dependencia espacial, es fundamental incorporar dicha dependencia en el modelo. Además, cualquier dependencia espacial que el modelo no logre explicar se reflejará en los residuos.

Si el modelo estimado no es capaz de capturar esa dependencia espacial, es decir, si los residuos muestran autocorrelación espacial, será necesario incluir explícitamente esa dependencia espacial en el modelo econométrico. El objetivo es ajustar el modelo hasta que los residuos se comporten como ruido blanco, es decir, distribuidos aleatoriamente con media cero y varianza constante.

$$Ruido\ blanco \sim N(0,\sigma^2)$$

Antes de pasar a los modelos con efectos espaciales, es interesante representar los residuos de este modelo de regresión para evaluar su distribución.

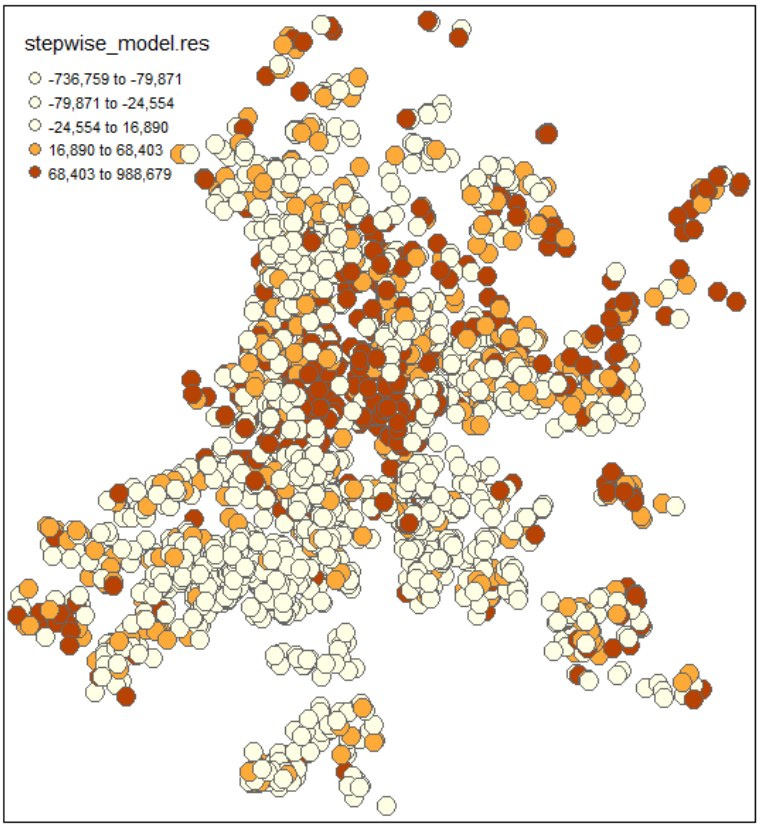


Tabla 8. Representación de los residuos del modelo Stepwise

Al analizar los residuos del modelo de regresión sin efectos espaciales, observamos que estos no se distribuyen de manera aleatoria. Esto indica que el modelo no está capturando adecuadamente la dependencia espacial existente en nuestros datos, por lo que es necesario incorporar la matriz de pesos espaciales en modelos de regresión para capturar esta dependencia espacial y así mejorar la precisión de las predicciones.

5.2. Desarrollo de modelos espaciales

Para abordar la dependencia espacial presente en los datos, necesitaremos considerar modelos espaciales. Hay una gran variedad de modelos espaciales que podemos utilizar, y la

utilización de algunas pruebas específicas pueden ayudarnos o darnos una idea del modelo espacial más adecuado para nuestros datos.

Los test de los multiplicadores de Lagrange son útiles para determinar si debemos emplear un modelo de regresión espacial autorregresivo (SAR) o un modelo de error espacial (SEM). Para ello, utilizando la matriz de distancias calculada anteriormente y el modelo MCO anterior, tenemos los siguientes resultados:

	Estadístico	Df	P-valor
RSerr	*426.23807*	*1*	*0.000000e+00*
RSlag	*590.06379*	*1*	*0.000000e+00*
adjRSerr	*59.54528*	*1*	*1.199041e-14*
adjRSlag	*223.37099*	*1*	*0.000000e+00*
SARMA	*649.60906*	*2*	*0.000000e+00*

Tabla 9. Resultados test Multiplicadores de Lagrange

A la vista de los resultados, tanto los tests de lag espacial como de error espacial son significativos, es decir, los dos tipos de dependencia espacial (sustantiva y residual) están presente en los datos, por lo que podemos considerar tanto el modelo SAR como el modelo SEM, o incluso un modelo mixto SARMA.

Tras probar con varios modelos diferentes, se muestran los resúmenes de los resultados obtenidos para todos ellos:

- **Modelo de error espacial (SEM)**

Coeficientes	Estimación	Std. Error	Pr(>|z|)
Intercept	*614740.668*	*188537.844*	*0.0011119*
CONSTRUCTEDAREA	*3854.546*	*86.795*	*< 2.2e-16*
ROOMNUMBER1	*24930.093*	*12753.059*	*0.0506028*
ROOMNUMBER2	*28181.638*	*12652.622*	*0.0259248*
ROOMNUMBER3	*10483.017*	*13134.266*	*0.4247878*
ROOMNUMBER4	*-39941.800*	*14974.323*	*0.0076451*
ROOMNUMBER5 or more	*-44457.412*	*19313.530*	*0.0213420*
BATHNUMBER2	*25028.286*	*6161.860*	*4.869e-05*
BATHNUMBER3	*152040.399*	*12170.794*	*< 2.2e-16*
BATHNUMBER4	*287268.967*	*24419.016*	*< 2.2e-16*
BATHNUMBER5 or more	*145737.992*	*63129.689*	*0.0209686*
HASTERRACE1	*-7894.025*	*4664.037*	*0.0905449*
HASLIFT1	*19827.576*	*5592.333*	*0.0003919*
HASAIRCONDITIONING1	*16236.351*	*4414.047*	*0.0002348*
HASSOUTHORIENTATION1	*-9271.410*	*4833.477*	*0.0550900*
HASSWIMMINGPOOL1	*46358.450*	*10428.055*	*8.767e-06*
HASDOORMAN1	*23917.416*	*5597.048*	*1.927e-05*
HASGARDEN1	*-13895.003*	*8021.646*	*0.0832400*
ISDUPLEX1	*-40011.127*	*15003.526*	*0.0076581*
FLOORCLEAN	*7282.767*	*1007.405*	*4.858e-13*
CADCONSTRUCTIONYEAR	*-266.860*	*97.103*	*0.0059921*
CADMAXBUILDINGFLOOR	*-1849.678*	*1192.003*	*0.1207246*
CADWELLINGCOUNT	*-176.922*	*119.681*	*0.1393335*
CADASTRALQUALITYID5_9	*-28650.049*	*5582.901*	*2.871e-07*
BUILTTYPE_11	*68920.378*	*15813.419*	*1.310e-05*
BUILDTYPE_21	*-33014.082*	*5395.853*	*9.451e-10*
DISTANCE_TO_CITY_CENTER	*-9220.080*	*2869.837*	*0.0013147*
DISTANCE_TO_METRO	*-66781.460*	*16959.028*	*8.223e-05*
DISTANCE_TO_CASTELLANA	*-25474.175*	*3498.595*	*3.306e-13*
Lambda		*0.53326*	

Number of observations	*2995*
Number of parameters estimated	*31*
Log likelihood	*-39189.03*
AIC	*78440*

Tabla 10. Resultados modelo de error espacial SEM

- **Modelo espacial autorregresivo (SAR)**

Coeficientes	Estimación	Std. Error	Pr(>|z|)
Intercept	*786348.937*	*180330.217*	*1.297e-05*
CONSTRUCTEDAREA	*3787.258*	*83.770*	*< 2.2e-16*
ROOMNUMBER1	*20281.276*	*NaN*	*NaN*
ROOMNUMBER2	*15473.127*	*NaN*	*NaN*
ROOMNUMBER3	*-3939.046*	*NaN*	*NaN*
ROOMNUMBER4	*-54321.632*	*NaN*	*NaN*
ROOMNUMBER5 or more	*-57801.217*	*NaN*	*NaN*
BATHNUMBER2	*19324.455*	*6202.087*	*0.0018345*
BATHNUMBER3	*150142.223*	*12342.266*	*< 2.2e-16*
BATHNUMBER4	*247883.648*	*24049.070*	*< 2.2e-16*
BATHNUMBER5 or more	*242896.173*	*63845.296*	*0.0001421*
HASTERRACE1	*-13786.060*	*4034.914*	*0.0006339*
HASLIFT1	*30476.924*	*5566.435*	*4.372e-08*
HASAIRCONDITIONING1	*15956.130*	*4440.845*	*0.0003268*
HASSOUTHORIENTATION1	*-12577.294*	*4877.238*	*0.0099152*
HASSWIMMINGPOOL1	*30785.629*	*9983.406*	*0.0020446*
HASDOORMAN1	*26219.047*	*5715.586*	*4.490e-06*
HASGARDEN1	*-21429.395*	*8069.919*	*0.0079199*
ISDUPLEX1	*-33026.765*	*15565.125*	*0.0338508*
FLOORCLEAN	*6355.398*	*1041.804*	*1.058e-09*
CADCONSTRUCTIONYEAR	*-428.150*	*92.777*	*3.934e-06*
CADMAXBUILDINGFLOOR	*-3842.983*	*1150.971*	*0.0008411*
CADWELLINGCOUNT	*-228.044*	*109.801*	*0.0378134*
CADASTRALQUALITYID5_9	*-24700.395*	*4835.107*	*3.246e-07*
BUILTTYPE_11	*58762.832*	*14976.00*	*8.716e-05*
BUILDTYPE_21	*-31403.747*	*5614.917*	*2.233e-08*
DISTANCE_TO_CITY_CENTER	*-7006.821*	*1426.015*	*8.943e-07*
DISTANCE_TO_METRO	*-47905.990*	*10043.069*	*1.842e-06*
DISTANCE_TO_CASTELLANA	*-7412.095*	*1729.193*	*1.816e-05*
Rho	*0.33363*		
Number of observations	*2995*		
Number of parameters estimated	*31*		
Log likelihood	*-39143.32*		
AIC	*78349*		

Tabla 11. Resultados modelo espacial autorregresivo (SAR)

- **Modelo de error espacial de Durbin (SDEM)**

Coeficientes	Estimación		Std. Error		Pr(>|z|)	
Intercept	*2075196.449*	***lag***	*407009.651*	***lag***	*3.421e-07*	***lag***
CONSTRUCTEDAREA	*3903.484*	*1157.363*	*88.721*	*175.713*	*< 2.2e-16*	*4.498e-11*
ROOMNUMBER1	*23057.998*	*-10497.828*	*13325.805*	*25952.895*	*0.0835719*	*0.6858484*
ROOMNUMBER2	*25017.366*	*-43010.260*	*13057.955*	*25060.041*	*0.0553814*	*0.0861093*
ROOMNUMBER3	*5804.345*	*-64817.955*	*13592.291*	*26049.450*	*0.6693559*	*0.0128368*
ROOMNUMBER4	*-47663.033*	*-102710.121*	*15569.810*	*30579.300*	*0.0022042*	*0.0007828*
ROOMNUMBER5 or more	*-57040.965*	*-140943.054*	*20525.244*	*41381.736*	*0.0054516*	*0.0006594*
BATHNUMBER2	*19164.813*	*-20676.545*	*6260.095*	*12788.689*	*0.0022029*	*0.1059250*
BATHNUMBER3	*146293.280*	*55357.756*	*12343.570*	*25499.207*	*< 2.2e-16*	*0.0299342*
BATHNUMBER4	*276890.034*	*31485.358*	*24325.442*	*47244.520*	*< 2.2e-16*	*0.5051337*

BATHNUMBER5 or more	*309543.931*	*781664.192*	*67440.499*	*136412.939*	*4.435e-06*	*1.004e-08*
HASTERRACE1	*-10645.211*	*-30214.946*	*4746.983*	*9736.999*	*0.0249277*	*0.0019150*
HASLIFT1	*26233.183*	*52601.314*	*5561.655*	*11285.74*	*2.396e-06*	*3.149e-06*
HASAIRCONDITIONING1	*16337.122*	*5939.253*	*4622.196*	*9400.705*	*0.0004086*	*0.5275253*
HASSOUTHORIENTATION1	*-12445.858*	*-14636.432*	*5058.110*	*10199.370*	*0.0138715*	*0.1512777*
HASSWIMMINGPOOL1	*37737.801*	*-13161.539*	*10275.356*	*18734.432*	*0.0002400*	*0.4823474*
HASDOORMAN1	*27645.950*	*48973.623*	*5719.191*	*11540.758*	*1.339e-06*	*2.200e-05*
HASGARDEN1	*-16814.765*	*-9638.859*	*8207.724*	*15443.617*	*0.0404962*	*0.5325407*
ISDUPLEX1	*-39844.705*	*553.888*	*15730.101*	*33598.382*	*0.0113084*	*0.9868470*
FLOORCLEAN	*7111.368*	*-1057.893*	*1047.917*	*2105.78*	*1.151e-11*	*0.6154048*
CADCONSTRUCTIONYEAR	*-252.179*	*-807.764*	*94.193*	*179.998*	*0.0074228*	*7.202e-06*
CADMAXBUILDINGFLOOR	*-3611.281*	*-356.426*	*1159.604*	*2067.450*	*0.0018442*	*0.8631239*
CADWELLINGCOUNT	*-180.954*	*-585.129*	*115.642*	*207.241*	*0.1176361*	*0.0047514*
CADASTRALQUALITYID5_9	*-18914.948*	*-54821.506*	*5447.953*	*10540.995*	*0.0005167*	*1.984e-07*
BUILTTYPE_11	*59870.809*	*-19357.913*	*15490.543*	*27703.283*	*0.0001111*	*0.4847028*
BUILDTYPE_21	*-29603.851*	*22194.465*	*5615.644*	*11608.329*	*1.352e-07*	*0.0558835*
DISTANCE_TO_CITY_CENTER	*36943.500*	*-40734.054*	*44944.992*	*45150.767*	*0.4110931*	*0.3669620*
DISTANCE_TO_METRO	*-34533.762*	*-11766.950*	*28390.646*	*34608.103*	*0.2238409*	*0.7338523*
DISTANCE_TO_CASTELLANA	*-120110.370*	*110447.037*	*46123.673*	*46391.602*	*0.0092118*	*0.0172772*
Lambda	*0.41119*					
Number of observations	*2995*					
Number of parameters estimated	*59*					
Log likelihood	*-39024.72*					
AIC	*78167*					

Tabla 12. Resultados modelo de error espacial de Durbin (SDEM)

- **Modelo espacial autorregresivo combinado (SACSAR)**

Coeficientes	Estimación	Std. Error	Pr(>|z|)
Intercept	*652292.295*	*174033.279*	*0.0001782*
CONSTRUCTEDAREA	*3888.115*	*97.013*	*< 2.2e-16*
ROOMNUMBER1	*23986.494*	*14653.770*	*0.1016551*
ROOMNUMBER2	*22933.453*	*14235.195*	*0.1071712*
ROOMNUMBER3	*3992.969*	*12171.710*	*0.7428714*
ROOMNUMBER4	*-47133.650*	*16816.968*	*0.0050670*
ROOMNUMBER5 or more	*-48676.134*	*21409.846*	*0.0229937*
BATHNUMBER2	*21744.592*	*6226.396*	*0.0004788*
BATHNUMBER3	*151917.501*	*12427.767*	*< 2.2e-16*
BATHNUMBER4	*273727.251*	*25005.744*	*< 2.2e-16*
BATHNUMBER5 or more	*181304.570*	*70742.810*	*0.0103811*
HASTERRACE1	*-11178.221*	*5354.798*	*0.0368416*
HASLIFT1	*25916.911*	*5639.563*	*4.316e-06*
HASAIRCONDITIONING1	*16414.494*	*5181.416*	*0.0015352*
HASSOUTHORIENTATION1	*-11110.975*	*5008.749*	*0.0265335*
HASSWIMMINGPOOL1	*37704.082*	*10009.645*	*0.0001654*
HASDOORMAN1	*24210.318*	*5638.952*	*1.760e-05*
HASGARDEN1	*-19069.297*	*7918.886*	*0.0160367*
ISDUPLEX1	*-33962.404*	*18055.669*	*0.0599742*
FLOORCLEAN	*6882.378*	*1022.378*	*1.677e-11*
CADCONSTRUCTIONYEAR	*-349.944*	*90.463*	*0.0001096*
CADMAXBUILDINGFLOOR	*-3221.337*	*1154.896*	*0.0052824*
CADWELLINGCOUNT	*-200.280*	*107.237*	*0.0618125*
CADASTRALQUALITYID5_9	*-25363.875*	*5488.631*	*3.816e-06*
BUILTTYPE_11	*66927.020*	*15646.282*	*1.890e-05*
BUILDTYPE_21	*-33636.487*	*6276.871*	*8.378e-08*
DISTANCE_TO_CITY_CENTER	*-8217.292*	*1854.125*	*9.341e-06*

DISTANCE_TO_METRO	*-54848.377*	*12056.190*	*5.380e-06*
DISTANCE_TO_CASTELLANA	*-10772.442*	*2333.059*	*3.888e-06*
Rho		*0.26174*	
Lambda		*0.22474*	
Number of observations		*2995*	
Number of parameters estimated		*32*	
Log likelihood		*-39119.75*	
AIC		*78304*	

Tabla 13. Resultados modelo espacial autorregresivo combinado (SACSAR)

- **Modelo de autorregresión espacial condicional (CAR)**

Coeficientes	Estimación	Std. Error	Pr(>|z|)
Intercept	*728582.520*	*194278.050*	*0.0001767*
CONSTRUCTEDAREA	*3762.526*	*90.472*	*< 2.2e-16*
ROOMNUMBER1	*-6626.547*	*13397.744*	*0.6208813*
ROOMNUMBER2	*7226.274*	*13265.568*	*0.5859329*
ROOMNUMBER3	*-7236.935*	*13775.082*	*0.5993300*
ROOMNUMBER4	*-73822.087*	*15693.291*	*2.550e-06*
ROOMNUMBER5 or more	*-41772.375*	*20290.757*	*0.0395240*
BATHNUMBER2	*18044.899*	*6437.054*	*0.0050585*
BATHNUMBER3	*140316.028*	*12677.580*	*< 2.2e-16*
BATHNUMBER4	*339260.415*	*25564.661*	*< 2.2e-16*
BATHNUMBER5 or more	*242727.305*	*66439.016*	*0.0002588*
HASTERRACE1	*-11209.371*	*4872.328*	*0.0214132*
HASLIFT1	*16218.887*	*5812.394*	*0.0052643*
HASAIRCONDITIONING1	*14527.662*	*4631.722*	*0.0017094*
HASSOUTHORIENTATION1	*-10195.235*	*5074.353*	*0.0445192*
HASSWIMMINGPOOL1	*35356.416*	*10840.863*	*0.0011086*
HASDOORMAN1	*24347.254*	*5841.088*	*3.069e-05*
HASGARDEN1	*-27859.223*	*8409.158*	*0.0009231*
ISDUPLEX1	*-59578.795*	*15695.654*	*0.0001471*
FLOORCLEAN	*5541.616*	*1055.867*	*1.534e-07*
CADCONSTRUCTIONYEAR	*-281.309*	*100.082*	*0.0049422*
CADMAXBUILDINGFLOOR	*-2892.179*	*1226.580*	*0.0183776*
CADWELLINGCOUNT	*102.776*	*123.147*	*0.4039557*
CADASTRALQUALITYID5_9	*-25469.305*	*5702.838*	*7.967e-06*
BUILTTYPE_11	*116794.166*	*16505.187*	*1.481e-12*
BUILDTYPE_21	*-32249.173*	*5656.054*	*1.186e-08*
DISTANCE_TO_CITY_CENTER	*-9408.551*	*2626.319*	*0.0003404*
DISTANCE_TO_METRO	*-132847.984*	*15855.885*	*< 2.2e-16*
DISTANCE_TO_CASTELLANA	*-26387.020*	*3178.353*	*< 2.2e-16*
Lambda		*0.72695*	
Number of observations		*2995*	
Number of parameters estimated		*31*	
Log likelihood		*-39239.73*	
AIC		*78541*	

Tabla 14. Resultados modelo de autorregresión espacial condicional (CAR)

Estos cinco modelos diferentes examinan la relación entre diversas características de las viviendas y su precio, teniendo en cuenta también, a diferencia del modelo MCO, la dependencia espacial.

- Cada coeficiente estimado indica el cambio esperado en el precio de una vivienda asociado con un aumento unitario en la variable independiente, manteniendo todas las demás variables constantes. Esto nos permite comprender cómo cada característica específica afecta al precio de la vivienda.

- Los valores en la columna “Pr(>|z|)” indican la significancia estadística de cada coeficiente. Aunque anteriormente ya hicimos una pre-selección de variables con el método “stepwise”, al entrenar nuevos modelos podemos ver que puede seguir existiendo variables no significativas. Valores bajos, normalmente menores que 0.05, indican una mayor significatividad.

- Lambda (λ) y rho (ρ) son los parámetros que indican la fuerza y dirección de la autocorrelación espacial en los errores del modelo y en la variable predictora respectivamente. Estos parámetros son importantes para comprender cómo la proximidad espacial entre las viviendas puede influir en los precios.

- El logaritmo de verosimilitud y el criterio de información de Akaike (AIC) proporcionan información sobre la bondad de ajuste de los modelos. Un AIC más bajo indica un mejor ajuste del modelo, lo que significa que el modelo explica mejor la variabilidad en los datos.

- Todos los modelos se han ajustado con 2995 observaciones.

Estas características nos ayudarán a evaluar y comparar los diferentes modelos, así como a comprender cómo las diferentes características de las viviendas influyen en sus precios, teniendo en cuenta la dependencia espacial.

Nuestro objetivo en este proyecto, es encontrar el mejor modelo que consiga recoger toda la dependencia espacial, o, dicho de otra forma, que los residuos del modelo sean ruido blanco.

Para ello, aplicaremos el test de Moran para examinar los residuos de cada modelo, lo que nos ayudará a determinar si el modelo captura adecuadamente la dependencia espacial.

Moran I test under randomisation					
Model	**Moran I statistic**	**Standard deviate**	**Expectation**	**Variance**	**p-value**
MCO	*0.2818375308*	*21.316*	*-0.0003340013*	*0.0001752254*	*2.2e-16*
SEM	*-0.0392513553*	*-2.9412*	*-0.0003340013*	*0.0001750764*	*0.9984*
SAR	*0.0685681585*	*5.2076*	*-0.0003340013*	*0.0001750605*	*9.564e-08*
SDEM	*-0.0117456426*	*-0.86254*	*-0.0003340013*	*0.0001750404*	*0.8058*
SACSAR	*-0.0126221951*	*-0.92876*	*-0.0003340013*	*0.0001750518*	*0.8235*
CAR	*-0.1800603623*	*-13.584*	*-0.0003340013*	*0.0001750460*	*1*

Tabla 15. Resultados test de Moran para los residuos mediante aleatorización

Además, para reforzarlo, realizaremos también una simulación por Monte-Carlo para evaluar la validez de os resultados obtenidos mediante el test de Moran I bajo randomización.

Monte-Carlo simulation of Moran I		
Model	**statistic**	**p-value**
SEM	*-0.039251*	*0.995*
SAR	*0.068568*	*0.000999*
SDEM	*-0.011746*	*0.8052*
SACSAR	*-0.012622*	*0.8192*
CAR	*-0.18006*	*0.999*
alternative hypothesis: greater		

Tabla 16. Simulación por Monte-Carlo para los residuos

Con estos resultados, podemos descartar el modelo SAR, ya que debido a su bajo p-valor (0.000999), podemos concluir que existe evidencia estadística suficiente para poder rechazar la hipótesis nula de que no existe autocorrelación espacial en los residuos. En cambio, los demás modelos, logran capturar toda o casi toda la dependencia espacial.

El mejor modelo para nuestro caso, sería el modelo CAR. Para comprobar su efectividad, representamos gráficamente los residuos para analizar su distribución y de esta manera asegurarnos de que no haya patrones espaciales significativos.

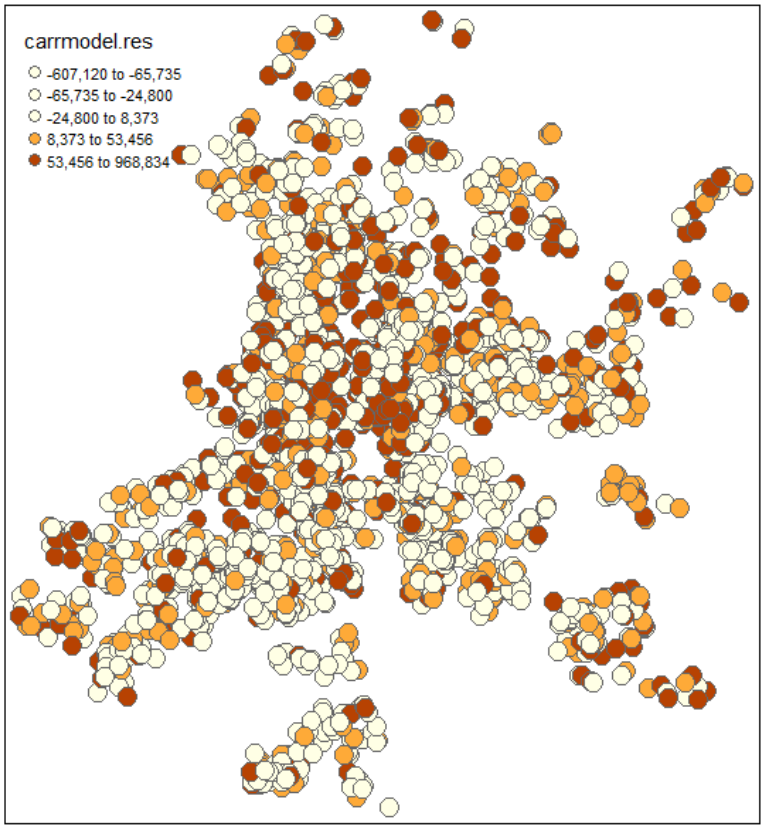


Figura 14. Representación de los residuos del modelo CAR

Como vemos en la representación, este modelo logra capturar toda la dependencia espacial al estar los puntos distribuidos aleatoriamente, asegurando que los residuos del modelo sean ruido blanco.

Interpretación del modelo

Una vez elegido el modelo ganador, podemos interpretar algunos de sus coeficientes de la siguiente manera:

- El precio base de la vivienda es de 728.582’52 €.
- Por cada metro cuadrado adicional, el precio de la vivienda aumenta en 3.762’526 €.
- Tener solamente una habitación disminuye el precio en 6.626.547 € comparado con no tener habitaciones.
- Tener piscina aumenta el precio en 35.356,416 € con respecto a no tener.
- Por cada km que se aleje la vivienda del centro de la ciudad, el precio disminuye en 9.408’551 €.

De esta forma, podemos comprender mejor los resultados del modelo y cómo cada una de las características afectan a los precios.

6. Conclusiones

El principal objetivo de este proyecto fue evaluar el impacto de las técnicas econométricas espaciales en la valoración de viviendas mediante el desarrollo de AVM's que incluyan componentes espaciales. Inicialmente, realizamos la recopilación, visualización y limpieza de los datos, una de las partes más importantes de cualquier análisis ya que nos permite comprender cómo se comportan las variables e identificar relaciones significativas entre ellas. Una vez conseguimos tener los datos limpios y preparados, realizamos un análisis exploratorio para identificar patrones geográficos. De esta manera, logramos extraer información valiosa sobre cómo la ubicación y otras características influyen en el valor de las viviendas.

Posteriormente, se desarrollaron y compararon varios modelos de econometría espacial, destacándose el modelo autorregresivo condicional (CAR) por su capacidad para capturar toda la dependencia espacial, asegurando que los residuos fuesen ruido blanco, lo que fue validado mediante los test de Moran y simulaciones por Monte-Carlo. La inclusión de componentes espaciales ha permitido capturar la influencia de las propiedades vecinas, resultando unas estimaciones más precisas y minimizando los errores y sesgos de los modelos. De esta manera, profesionales de este sector pueden ayudarse de estos modelos para la toma de decisiones en la compra y venta de las propiedades. Además, el uso del Big Data proporciona métodos más eficientes y menos costosos.

Para mejorar los resultados obtenidos, podemos considerar algunas aplicaciones. La aplicación de transformaciones en las variables podría mejorar la precisión de los modelos. Incorporar datos temporales y el uso de modelos espacio-temporales nos permitiría analizar las tendencias a lo largo del tiempo y predecir cómo los valores de las viviendas pueden cambiar en el futuro, incluyendo también factores socioeconómicos. Además, otra idea podría ser desarrollar herramientas que permitan a los usuarios ingresar datos o características de una propiedad para obtener una valoración al momento, esto podría incluir aplicaciones móviles o plataformas web.

En resumen, este proyecto ha demostrado que la incorporación de técnicas econométricas espaciales y el uso de modelos avanzados proporcionan beneficios significativos. Las mejoras propuestas no sólo pueden ayudar a mejorar la precisión, sino que también pueden ampliar su aplicabilidad en el mercado inmobiliario.

Bibliografía

Anselin, L. (1988). Spatial Econometrics: Methods and Models (1988a ed.). Kluwer Academic.

Anselin, L., & Rey, S. J. (s/f). Modern spatial econometrics in practice: AGuideto Geo Da, Geo DaSpace and PySAL. Sergerey.org. Recuperado el 3 de junio de 2024, de https://sergerey.org/giasp16/pdfs/anselin_rey_weights.pdf

Bivand, R. S., Pebesma, E. J., & Gomez-Rubio, V. (2008). Applied spatial data analysis with R (2008a ed.). Springer.

Ceballos, I. A. B. E. (s/f). Algunos conceptos de la econometría espacial y el análisis exploratorio de datos espaciales. Edu.co. Recuperado el 3 de junio de 2024, de https://repository.eafit.edu.co/server/api/core/bitstreams/e172a144-1e8f-4fca-91bc-80541200e57f/content

De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. Annals of the Institute of Statistical Mathematics, 64(1), 107–133. https://doi.org/10.1007/s10463-010-0298-1

Dube, J., & Legros, D. (2014). Spatial Econometrics using Microdata: Dubé/spatial econometrics using microdata. ISTE Ltd and John Wiley & Sons.

Elhorst, J. P. (2010). Applied spatial econometrics: Raising the bar. Spatial Economic Analysis, 5(1), 9–28. https://doi.org/10.1080/17421770903541772

Fischer, M. M., & Getis, A. (Eds.). (2009). Handbook of applied spatial analysis: Software tools, methods and applications (2010a ed.). Springer.

Haining, R. (2010). Spatial data analysis: Theory and practice. Cambridge University Press. https://doi.org/10.1017/cbo9780511754944

Llavero, J. (2024, marzo 8). 10 factores que influyen en el precio de una vivienda. Blog iad. https://blog.iadespana.es/consejos/vender/precio-de-una-vivienda/

Muñoz, A., & Margarita, S. (2014). Econometría espacial: Método y aplicaciones. Universidad Autónoma de Bucaramanga UNAB.

Portal de datos abiertos del Ayuntamiento de Madrid. (s/f). Madrid.es. Recuperado el 3 de junio de 2024, de https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?vgnextoid=a4f36d34fa86c410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD

Recio, D. (s/f). GeoAnalysis-idealista18-WL: Data cleansing and geospatial analysis of the idealista18 dataset using Wolfram Language. Focuses on neighborhood-level visualizations to uncover real estate market trends. Ideal for data enthusiasts and real estate analysts.

Serrano, R. M., & Valcarce, E. V. (2000). TECNICAS ECONOMETRICAS PARA EL TRATAMIENTO DE DATOS ESPACIALES : LA ECONOMETRIA ESPACIAL. Edicions Universitat Barcelona.

Toloza-Delgado, J., Melo-Martínez, O., & Azcarate-Romero, J. (2021). Determinantes del precio de la vivienda nueva en Bogotá para el año 2019: una aproximación a través de un modelo semiparamétrico de regresión espacial. Ingeniería y Ciencia, 17(34), 23–52. https://doi.org/10.17230/ingciencia.17.34.2

Torres, G. A. A. (s/f). Dependencia Espacial: Detección, Validación y Modelación. Edu.co. Recuperado el 3 de junio de 2024, de https://repository.eafit.edu.co/server/api/core/bitstreams/2f47925c-e233-4832-

bb7f-d3a91f8c815d/content

Torres, J. A. J. (2019, noviembre 28). Detección y reemplazo de outliers con R. Adictos al trabajo. https://www.adictosaltrabajo.com/2019/11/28/deteccion-y-reemplazo-de-outliers-con-r/