

Capítulo 4

Modelo de utilidad de localización

“Localización, localización, localización.”

— Harold Samuels, promotor inmobiliario

4.1 Introducción

La vivienda es un bien inusual en tres aspectos: heterogeneidad, durabilidad e inmovilidad (Kiel y Zabel, 2008), este último factor apunta a la localización como un criterio fundamental en la toma de decisión al adquirirla y, además, determina en buena medida su valor. Esta intuición fue respaldada desde hace décadas por numerosas investigaciones, como Friedman y Weinberg (1981), o Hanushek y Quigley (1979), que sugerían que muchos hogares deciden donde vivir en función a los ingresos familiares. El método de los precios hedónicos, mencionado en capítulos anteriores, se podría usar para medir la influencia de la ubicación, sin embargo, no se ha logrado un consenso general de cómo especificar las covariables de localización en los modelos. A menudo se construyen de manera arbitraria, lo que no logra controlar fenómenos como la heterogeneidad espacial¹ y la dependencia espacial² (Anselin y Rey, 2014); la autocorrelación espacial (Anselin y Griffith, 1988); el cambio de calidad de la vivienda; la multicolinealidad entre variables (Orford, 2017) y la heterocedasticidad (Fletcher *et al.*, 2000).

La aparición de la heterogeneidad y la autocorrelación espacial se relacionan con

¹La heterogeneidad espacial (Anselin y Griffith, 1988) consiste en la falta de uniformidad de los efectos (incluida la dependencia espacial) sobre el espacio geográfico.

²En términos de asociación (correlación), la dependencia espacial implica precios parecidos para lugares cercanos, en términos más amplios se refiere a la influencia del entorno en las variaciones del precio.

una especificación deficiente de los atributos de ubicación (Helbich *et al.*, 2014), y a pesar de los numerosos estudios realizados al efecto, es aún una cuestión por resolver, como indica Bourassa (2021). Por su parte Heyman (2018), en una revisión sistemática de la especificación de la localización en modelos hedónicos, señala que la mayoría de los casos la especifican de una forma poco elaborada, o a través de características de área agregadas arbitrariamente. En nuestra opinión, esto es debido a tres factores: disponibilidad de datos, partición espacial inadecuada y coste computacional de cálculo. En definitiva, estas variables suelen estar incompletas, desactualizadas o especificadas arbitrariamente.

El primer desafío a abordar es cómo especificar la ubicación en el espacio, ya que los diseños urbanos son heterogéneos y la influencia de los factores varía enormemente de una zona a otra (LeSage y Pace, 2009). La segunda cuestión es la disponibilidad de información: los datos pueden ser inadecuados o agregarse con un criterio incompatible con el dominio del problema. Por ejemplo, las secciones censales de población están diseñadas para describir la distribución y las características de la población, no para reflejar las características de sus submercados de viviendas³. El tercer problema es que los atributos de ubicación no incorporan la utilidad marginal de la misma, al estar generalmente basadas en distancias euclidianas. Los atributos de utilidad basados en tiempos de desplazamiento o índices gravitatorios rara vez se utilizan en la industria y la investigación, debido a su dificultad de parametrización y a su coste computacional.

La delimitación geográfica de los submercados (geográficos) de la vivienda es eficaz en la mejora de los modelos hedónicos. Sin embargo, aún es una cuestión sin resolver completamente (Bourassa *et al.*, 2021), aunque recientemente encontramos avances en estudios que hacen uso del aprendizaje automático como Wu, Wei y Li (Wu *et al.*, 2020) o Rey *et al.* (2023).

El presente capítulo describe una metodología sistemática para la creación de atributos de ubicación aplicables a modelos de precios hedónicos de la vivienda. Estas variables se construyen como índices de accesibilidad de tipo gravitatorio⁴ de forma automática. El método resuelve los problemas habituales en la creación de variables auxiliares de localización: que sea un proceso genérico, es decir, que sea independiente del aspecto de accesibilidad a incorporar; que la variables creadas sean coherentes y robustas en términos de utilidad; y finalmente, que su

³ Esta cuestión se refiere al denominado problema del área modificable (MAUP), que estudia la variabilidad de la correlación en función del tipo de regiones por el que se divide el espacio (Wong, 2004), es decir que los resultados pueden variar sensiblemente en función de uso un tipo de división zonal u otro.

⁴ Una medida gravitatoria es aquella cuya intensidad es inversamente proporcional a la distancia entre un punto de interés y los elementos de información a los que se refiere.

especificación muestre coherencia entre la utilidad y los precios. Cada índice de accesibilidad de ubicación sintetiza numéricamente las oportunidades que tiene una vivienda cercana y que afectan a su precio, como oferta de ocio, escuelas, lugares de trabajo, etcétera.

Para comprobar los resultados de la metodología, se evalúa empíricamente el funcionamiento de una serie de índices de accesibilidad para la ciudad de Madrid con el conjunto de datos de oferta, comparando su rendimiento sobre cinco algoritmos de modelado de precios hedónicos.

4.1.1 Medidas de accesibilidad

La accesibilidad ha sido un tema central en la planificación física desde la segunda mitad del siglo XX (Batty, 2009). Los primeros usos del término se remontan a 1920, aunque sin embargo, fue Hansen (1959) quien propuso inicialmente una metodología para el uso de la accesibilidad en la planificación urbana. En ella definía la accesibilidad como “...el nivel de interacciones con una serie de oportunidades como compras, actividad residencial y empleo...”. Estudios relacionados en otros campos como geografía poblacional (véase (Stewart, 1947)), definieron el potencial gravitatorio ponderando la suma de fuerzas para explorar las reglas de distribución y equilibrio poblacional.

Batty (2009) propuso que un índice de accesibilidad asocia un grado de oportunidad a un lugar con el coste de materializarlo. Además, los índices de accesibilidad suelen presentarse en una forma compuesta que resume lo fácil o difícil que es hacer realidad una serie de oportunidades para un lugar determinado. El coste, también llamado impedancia, se puede medir como tiempo o distancia. Batty identifica tres tipos de accesibilidad: el primero, define lo cerca que está un individuo de una “oportunidad” como una operación calculada de forma directamente proporcional con su dimensión⁵, e inversamente proporcional a su distancia; el segundo, se centra en la distancia de un lugar a otro, ya sea la distancia euclíadiana o la distancia del tiempo de viaje; y el tercero, basado en un enfoque mixto de primer y segundo tipo, por ejemplo medidas que utilizan la sintaxis espacial⁶ (Hillier y Hanson, 1989).

Todos los planteamientos son una evolución del modelo de Von Thünen (1826), que en su trabajo “el Estado aislado” formula que el precio que un agente está dispuesto a pagar por la tierra depende de dos factores: la productividad de los cultivos y los costes de transporte. De modo que, el agricultor paga por lugares

⁵El tamaño de una oportunidad se refiere al nivel de intensidad de la misma, por ejemplo número de comercios cercanos o metros cuadrados de oficina cercanos.

⁶La sintaxis espacial es un campo que estudia la configuración de elementos espaciales y sus relaciones topológicas. Para las zonas urbanas, estudia los elementos de su red viaria y su malla urbana.

que maximizan su beneficio, que es el equilibrio entre costes e ingresos. Por lo tanto, la distancia a los mercados induciría costos que reducirían el precio de la tierra (en términos de competencia perfecta).

La teoría de la localización⁷, derivada del planteamiento de Von Thünen, se empezó a aplicar a las áreas urbanas en los años sesenta y setenta del siglo XX (Alonso *et al.*, 1964), (Mills, 1972), (Wingo, 1961) y (Muth, 1969).

La literatura resalta la importancia de la ubicación con respecto al centro de la ciudad, y denomina como “valor de situación” a la tasación monetaria de todas las ventajas de ubicación encontradas alrededor de un lugar. Por lo tanto, un modelo basado en la distancia al centro de la ciudad (CBD), donde se encuentran la mayoría de los servicios, puede explicar el aumento en el valor de la vivienda. Esta prima de valor es producto de la mayor utilidad percibida por el propietario al requerir un menor tiempo de desplazamiento hasta dónde se encuentran estos servicios. Witte, Sumka y Erekson (Witte *et al.*, 1979) publican uno de los primeros estudios que utilizan la variable distancia al CBD desde el barrio, en una aplicación de la teoría de los mercados implícitos de Rosen (1974). D’Acci (2019) revisa la numerosa literatura al respecto, concluyendo que la calidad de la ubicación (es decir, las características del área a través de muchos dimensiones) se capitaliza por el valor del inmueble. Ilustra el análisis realizando un estudio de caso sobre la ciudad italiana de Turín.

En el caso del precio de la vivienda, la ubicación determina tener una serie de ventajas o desventajas, generando utilidades o desutilidades que afectan el precio de venta de la propiedad. Aunque este enfoque proviene originalmente de Court (1939), se hizo más popular en la década de 1960 (Griliches, 1961). Los primeros enfoques para introducir el lugar como parte de los modelos hedónicos. Kain y Quigley (1970) incluyeron las características estructurales de la unidad de vivienda, las características del vecindario y la distancia al CBD. Witte *et al.* (1979) es una de las primeras investigaciones que utiliza la distancia al centro desde el barrio como una aplicación de la teoría de los mercados implícitos de Rosen. Posteriormente, este modelo se enriqueció con otras características del barrio (Bowen *et al.*, 2001).

Existen otras formas de incorporar la localización, como los de efectos fijos de ubicación, basadas en variables *dummy*. Cada una de ellas representa uno de los posibles lugares (barrios, secciones censales, por ejemplo). Tienen la ventaja de ser fáciles de especificar, pero pueden dar lugar a un gran número de covariables. Además, como indica Heyman (2019) en su estudio para Oslo, este

⁷La teoría de la localización es una disciplina de la geografía y la economía que estudia la relación entre las actividades económica y su situación geográfica (Britannica, 2014).

tipo de características tienen un poder explicativo limitado en comparación con las variables de ubicación relativa.

La distancia euclíadiana al CBD es un método directo y simple para incorporar la accesibilidad, no obstante, el enfoque monocéntrico no es aconsejable en las configuraciones urbanas actuales, como describen Waddell (1993) o Heikkila (1989), que cuestionan la validez de modelos monocéntricos en favor del uso de modelos policéntricos. Para abordar la cuestión, Song (1994) crea una serie de medidas de accesibilidad, demografía y planificación urbana en un modelo de precios de la vivienda. En este sentido, Knaap y Song (2003) aplicaron una serie de medidas cuantitativas utilizando Sistemas de Información Geográfica (GIS) a modelos hedónicos, y lograron analizar la contribución de cada medida al precio. Definieron seis características que afectan a las viviendas unifamiliares: diseño de calles, sistemas de circulación, conectividad, tamaño de bloque y configuración de malla cuadrada. Sobre estos primeros modelos, se pueden encontrar trabajos que desarrollan estas medidas con datos abiertos, en grandes redes urbanas (Blanchard y Waddell, 2017; Liu *et al.*, 2022).

A pesar del gran número de contribuciones de los últimos años, algunos autores, como Handy (2020), sostienen que la adopción de la accesibilidad en la planificación urbana ha evolucionado poco en las últimas décadas. Sin embargo, se han publicado recientemente algunos estudios prometedores que reintroducen el concepto de accesibilidad dentro del marco de pensamiento del análisis urbano. En paralelo, la creciente abundancia de fuentes abiertas y las nuevas capacidades de tratamiento de la información han producido nuevas variables espaciales genéricas (Vecchio y Martens, 2021). Como, por ejemplo, los indicadores genéricos globales propuestos por Boeing (2022), o las características espaciales urbanas denominadas “Spatial Signatures” propuestas por Arribas-Bel y Fleischman (2022). Estas últimas, se pusieron a disposición como fuentes de datos abiertas (Samardzhiev *et al.*, 2022).

El uso de fuentes abiertas de internet para la creación de características de zona en modelos hedónicos es creciente. Por ejemplo, Xiao (2017) utiliza los datos de POIs de OSM⁸ en los modelos de la vivienda en Beijing y demuestra su capacidad para controlar la autocorrelación espacial. Por otra parte, Hu (2019) usa esta misma fuente para delimitar los usos del suelo en la ciudad de Guangzhou (China). Li (2019) utiliza portales inmobiliarios, puntos de interés en Baidu e imagen satélite, para la construcción de indicadores de accesibilidad y estructura urbana para la estimación de precios. En este sentido, Čeh (2018) aplica la información de puntos de interés a la construcción de características de

⁸Open Street Map, para más información véase el Capítulo 2.

esta naturaleza para la capital de Eslovenia. Más recientemente, Liu et al. (2022) publicaron un método para estimar indicadores de accesibilidad globales basados en datos abiertos, que a pesar de haberse diseñado para estimar el precio del suelo, como este caso, resuelve el problema de la ausencia de replicabilidad, comparabilidad y reproducibilidad de los métodos basados en fuentes abiertas (Brunsdon y Comber, 2021). Cellmer (2023) demuestra la relación entre el precio de la vivienda y la densidad de ciertos tipos de POI, incidiendo el potencial de esta información para la segmentación del espacio urbano.

Bowes e Ihlanfeldt (2001), Bartholomew y Ewing (2011), Agostini y Palmucci (2017), Lieske *et al.* (Lieske *et al.*, 2021), y Choi, Park y Uribe (2022) estudiaron el efecto del acceso del transporte público en los precios del suelo, incluidos los efectos directos e indirectos de las estaciones de transporte. Estos estudios descubrieron que las estaciones situadas lejos del centro de la ciudad tienen impactos positivos, creando islas de valores inmobiliarios más altos, de la misma forma que demostraron la influencia en los precios de las estaciones de tránsito.

La aproximación del método propuesto trata *ex-ante*, y desde un punto de vista de utilidad, el sesgo que introduce la dependencia espacial. Lo que contrasta con las aproximaciones de análisis espacial cuyo objetivo principal es explicar la influencia de la localización en un fenómeno de estudio, por tanto tratar *ex-post* su existencia y efectos. Como explican Montero y *et al.* (2015): “...en la geoestadística el aspecto más importante en el análisis geoestadístico es cuantificar la correlación espacial entre las observaciones (...) y usar esta información para lograr los objetivos anteriores⁹...”. Este planteamiento tiene la ventaja de reproducir el comportamiento de los precios del suelo, no en base a la localización sino en función de las causas que lo producen, introduciendo una herramienta de análisis econométrico de gran valor.

Nuestra contribución ayuda a superar la imprecisa especificación estándar de la localización en la modelización hedónica, como las variables ficticias de localización, mediante la introducción de indicadores interpretables, altamente granulares y fáciles de calcular, capaces de producir regresiones mejor ajustadas (Diewert y Shimizu, 2021).

4.2 Metodología

El modelo parte de cuatro fuentes de datos: idealista, catastro, información censal de INE y cartografías de Open Street Map, todas ellas descritas en el Capítulo 2. Principalmente, en la construcción de las variables, se usarán un conjunto de

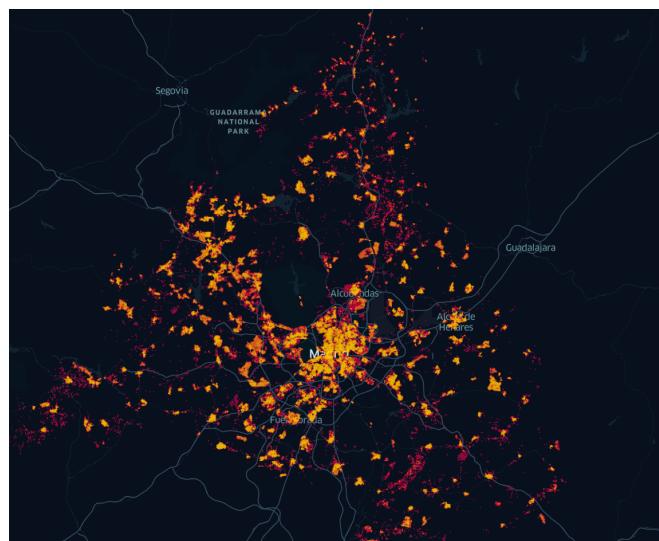
⁹Referido a reproducir un proceso espacial observado para un conjunto de observaciones.

anuncios de Idealista del año 2018, no solamente registros de alquiler sino que también de compraventa¹⁰.

Las medidas de accesibilidad se construyen agregando superficies catastrales, número de inmuebles, y puntos de interés de OSM (descritos en los subepígrafes 2.4.4 y 2.4.5), que permiten medir los diferentes tipos de usos inmobiliarios alrededor de cada vivienda. La información vial de OSM se ha utilizado para construir la topología de red de transportes a pie y coche, que es necesaria para la definición de las isócronas sobre las que se calculan las medidas.

El dato catastral permite, además, identificar todas las ubicaciones “semilla”, que son las posibles ubicaciones de una vivienda, y por tanto, los únicos lugares donde un propietario podría materializar las oportunidades. La ubicación exacta se calcula como el centroide¹¹ de las fincas de tipo residencial en la Comunidad de Madrid.

Figura 4.1. Localizaciones semilla utilizadas, el color indica la frecuencia



Fuente: elaboración propia.

La metodología propuesta crea un conjunto de índices de accesibilidad que capturan la contribución de la ubicación a los precios de la vivienda, efecto ligado al principio de utilidad de la localización (Rey-Blanco *et al.*, 2023b). De una forma muy resumida, el proceso desarrollado para construir los índices de accesibilidad se basa en seleccionar aquellas variables de accesibilidad que potencialmente reduzcan los errores espaciales del modelo. Existen diferentes estudios que relacionan la incorporación de atributos de accesibilidad con la reducción de la autocorrelación espacial en los residuos del modelo, por ejemplo Morali y Yilmaz

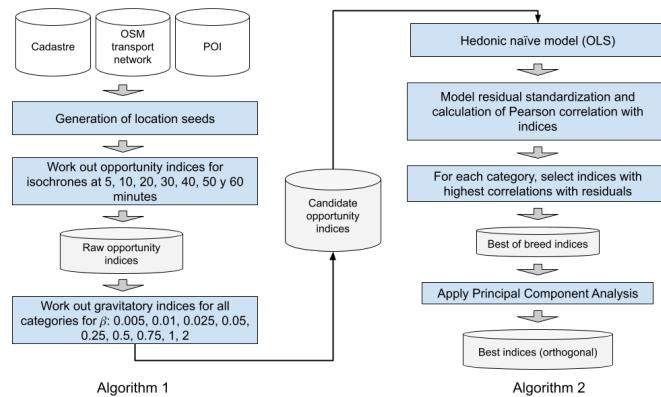
¹⁰Se opta por incluir datos de compraventa para incrementar el nivel de soporte del método.

¹¹El centroide representa el centro geométrico de una forma poligonal.

(2020).

El proceso general se realiza en tres pasos desarrollados en los dos algoritmos presentados en la Figura 4.2. El primero crea la familia de los índices candidatos básicos, mientras que el segundo se encarga de seleccionar el más adecuado, mediante un proceso heurístico. Posteriormente, se realiza un análisis de componentes principales para lograr un conjunto de variables de accesibilidad ortogonales.

Figura 4.2. Proceso general de construcción de índices de accesibilidad



Fuente: elaboración propia.

Para demostrar su validez se comprueba que las variables creadas aportan información de ubicación útil en el proceso de formación del precio de la vivienda, usando distintos enfoques de modelado hedónico: tradicionales y no tradicionales, basados en aprendizaje automático.

En la especificación de categorías de índices de accesibilidad, se ha seguido la clasificación propuesta por (Heyman *et al.*, 2018), que también utiliza el término “medidas de oportunidad” para referirse a este tipo de índices. Un índice de accesibilidad se define por una serie de oportunidades que ofrece una determinada ubicación, calculada al aplicar una función de impedancia gravitatoria, basada en el coste de desplazamiento¹², desde las oportunidades hasta la ubicación. Se ha decidido trabajar con dos medios de transporte para estimar las medidas: a pie y en coche.

Como se ha tratado en capítulos anteriores, no existe un consenso en los parámetros o forma funcional que debe seguir un modelo hedónico, y particularmente aún en la actualidad no existe una forma canónica para la especificación de los atributos de localización (Bourassa *et al.*, 2021). Dado que el objetivo principal de esta investigación es determinar la superioridad de los

¹²En nuestro caso los costes de transporte se expresan en tiempos de desplazamiento, aunque existen alternativas como la distancia en metros o el coste de combustible.

índices de accesibilidad propuestos sobre alternativas metodológicas más simples. Esta evaluación se realizará en términos de precisión y capacidad de representar la influencia de la localización. Con el fin de representar funcionalmente la relación entre estas variables y el precio de la vivienda, y para mantener su inteligibilidad se manejará una especificación simplificada, mediante una forma funcional estándar con variables ficticias de tiempo.

Tabla 4.1. Categorías de variables

Característica	Motivación
Estructural	Características estructurales que capturan la contribución de las características físicas de la propiedad, como superficie, número de habitación o estado de conservación.
Mercado	Incorpora la principal dinámica de oferta/demanda del mercado donde se ubica el inmueble
Localización	Explica la contribución de la ubicación en el precio del suelo, incluye características del barrio, índices de accesibilidad y otras características geográficas
Dummy de tiempo	Captura el ajuste de tiempo a lo largo del tiempo, efectos de tendencia y estacionalidad

Fuente: elaboración propia

Esta formulación expresa el modelo como una función lineal de los atributos de propiedad, dentro de las cuatro categorías de características descritas en la Tabla 4.1. El modelo de precio hedónico de la vivienda asume que el precio de una propiedad n en el período t , p_n^t , es una función de un número fijo de características o rasgos $q = 1, \dots, Q$, alcanzable por coche o andando, $m = \{\text{coche, a pie}\}$, que se miden por una serie de cantidades $Z_{nq}^{tm} = \{\hat{A}_{nk}^{tm*}, S_{nj}^t, M_{nl}^t, D_{nk}^t\}$ observados en $t = 1, \dots, T$ períodos, más un término de error aleatorio ε_n^t . Es decir,

$$p_n^t = \beta_0 + \sum_k \beta_k \cdot \hat{A}_{nk}^{tm*} + \sum_j \beta_j \cdot S_{nj}^t + \sum_l \beta_l \cdot M_{nl}^t + \sum_t \delta \cdot D_n^t + \varepsilon_n^t, m = \{\text{coche, a pie}\}, [4.1]$$

donde: \hat{A}_{nk}^{tm*} representa los índices de *accesibilidad* (*A*) de ubicación en términos de $k = 1, \dots, K$ *oportunidades* de ubicación a las que se puede llegar en automóvil y caminando; S_{nj}^t denota los atributos $j = 1, \dots, J$ *estructural* (*S*) de la propiedad; M_{nl}^t captura las características de oferta y demanda $l = 1, \dots, L$ del *submercado* (*M*) al que pertenece el inmueble; y D_n^t representa las variables *dummy* de tiempo (*D*).

Se anticipa que los \hat{A}_{nk}^{tm*} índices de accesibilidad usados en el modelo [4.1] son el resultado de un proceso de optimización que, partiendo de un conjunto de índices de accesibilidad básicos, A_{nk}^{tm} (es decir, sin la notación $\hat{\cdot}$), determinan su mejor definición maximizando su correlación con los residuos de un modelo MCO *naïve*¹³ que omite la accesibilidad pero incluye el resto de atributos ($S_{nj}^t, M_{nl}^t, D_{nk}^t$), y posteriormente transforma estos índices correlacionados entre sí, A_{nk}^{tm*} (denominado “óptimo” y denotado con el superíndice ‘*’), en ortogonales a través del análisis de componentes principales (PCA) (Pearson, 1901), obteniendo finalmente \hat{A}_{nk}^{tm*} .

Finalmente, como variable dependiente (p_n^t) se utiliza el precio por metro cuadrado, ya que ayuda a reducir la heterocedasticidad del modelo. Bajo los supuestos de error clásicos, en particular una media cero y una varianza constante, el modelo se estima a partir de los datos agrupados correspondientes a todos los períodos de tiempo, representados por variables ficticias dicotómicas de tiempo.

4.2.1 Especificación de índices de oportunidad

Las variables sobre las que se crean los índices de accesibilidad resumen numéricamente las oportunidades que ofrece la ubicación. En áreas densamente pobladas es conveniente calcular la accesibilidad tanto en automóvil como a pie. Los índices para ambos medios de transporte se basan en las mismas fuentes de información y, como se muestra en la siguiente subepígrafe, es el algoritmo de optimización quien decide qué importancia tiene cada modo. Con este procedimiento se evita la introducción de la subjetividad que supondría la elección de atributos diferentes por cada índice.

Cada oportunidad, que contribuye a un índice de accesibilidad, se calcula como un índice gravitatorio de los valores de cada variable dentro de una serie de isócronas, las cuales, se adaptan de acuerdo a los dos modos de transporte (Wee y Vickerman, 2021). Para el modo de transporte automóvil, se considera el vector de distancias $d_{i(m=coche)}$, accesible desde cualquier lugar en los siguientes tiempos de viaje: $i(coche) = \{1, \dots, I\} = \{5, 10, 20, 30, 40, 50 \text{ y } 60 \text{ minutos}\}$. Para los índices peatonales $d_{i(m=a \text{ pie})}$, se consideran los tiempos¹⁴. Para el modo de transporte de coche se considera el vector de distancias $d_i(m = coche)$, alcanzables desde cada ubicación en los siguientes tiempos: $i(coche) = 1, \dots, I = 5, 10, 20, 30, 40, 50 \text{ y } 60 \text{ minutos}$

¹³Un modelo sencillo sin especificación de variables de zona.

¹⁴La definición de isocronas que van de 5 a 30 minutos a pie, o de 10 a 60 minutos en coche, son de uso común en la literatura sobre accesibilidad. Ewing y Cervero (2010) y Handy y Niemeier (1997) sugieren utilizar umbrales de 10 minutos a pie para medir la accesibilidad a puntos de interés en configuraciones urbanas habituales. Frank et al. (2010) recomiendan tomar distancias de 5 minutos a pie para parques y paradas de transporte público, y límites mayores para otros destinos.

e $i(a pie) = \{1, \dots, I\} = \{5, 10, 20 \text{ y } 30 \text{ minutos}\}$. Las distancias de conducción se calculan asumiendo el límite legal máximo de velocidad de la vía, mientras que las distancias a pie suponen una velocidad de 5 km/h.

Al definir los índices de accesibilidad básicos A_{nk}^{tm} , se asume que la localización de la vivienda aporta una determinada utilidad a los propietarios, consecuencia de su cercanía a una serie de oportunidades. Éstas se representan a través de una serie de variables observadas en la ubicación n en el momento t .

La utilidad que produce una oportunidad es decreciente en función de los costes de transporte, expresado como una función de penalización exponencial inversamente proporcional a las distancias al tiempo de desplazamiento. En particular, para cada variable k , el índice básico de accesibilidad de ubicación, A_{nk}^{tm} , agrega sus valores para todas las isócronas I (por ejemplo, número de paradas de autobús a 5, 10, 20 y 30 minutos a pie), cada uno ponderado por su tiempo de viaje relativo, en un solo escalar. Esto corresponde a la especificación, basada en Levison y Krizek (2005), de la expresión analítica:

$$A_{nk}^{tm} = \sum_{i(m)=1}^I O_k(X, Y, d_{i(m)}) \cdot e^{-\beta_m d_{i(m)}}, m = \text{coche, a pie}, \quad [4.2]$$

donde $O_k(X, Y, d_{i(m)})$ representa un tipo de oportunidad a una distancia d desde un lugar ubicado en las coordenadas X, Y , dentro de los límites geográficos marcados por una serie de isócronas. La medida de distancia $d_{i(m)}$ corresponde a los rangos de tiempos de viaje antes mencionados (en minutos). El parámetro β_m representa la caída exponencial del índice para la impedancia aplicada (en este caso, el tiempo de distancia calculado).

El índice de accesibilidad [4.2] es específico para cada modo de transporte, ya sea caminando o conduciendo. En el epígrafe 4.3 se argumenta la selección de los valores óptimos de A_{nk}^{tm*} sobre un rango de valores de β_m evaluados por el algoritmo de optimización.

Los índices de accesibilidad gravitatorios A_{nk}^{tm} se construyen agregando el número de elementos para cada $k = 1, \dots, K$ oportunidad observada en una ubicación determinada n en el tiempo t . La Tabla 4.2 muestra las $K=26$ variables utilizadas en este caso para representar estas oportunidades, agrupadas en cinco categorías. Estas categorías incluyen *Transporte público, Transporte privado, Actividad económica más Servicios básicos, Social y Recreativo*.

Tabla 4.2. Catálogo de índices de oportunidad base

Categoría	Subcategoría	variable	Medida	Fuente
Transporte público	bus	TRANSPORT.BUS	frecuencia	OSM
	metro	TRANSPORT.METRO	frecuencia	OSM
	tren	TRANSPORT.TRAIN	frecuencia	OSM
Transporte privado	aeropuerto	TRANSPORT.AIRPORT	frecuencia	OSM
	autopista	ROUTING.HIGHWAY	longitud	OSM
	rutas	ROUTING.COMPLEXITY	densidad	OSM
Instalaciones urbanas	suelo	CAD.URBANLAND	superficie	catastro
	hotel	HOTEL	frecuencia	OSM
	hotel	VACATIONAL	frecuencia	airbnb
	comida	FOOD	frecuencia	OSM
	público	TOURISM	frecuencia	OSM
	público	MONEY	frecuencia	OSM
	educación	CAD.PUBLIC	superficie	catastro
	educación	CAD.SCHOOL	superficie	catastro
	educación	EDUCATION	frecuencia	OSM
	turismo	TOURISM	frecuencia	OSM
Social	salud	CAD.HEALTH	superficie	catastro
	comercio	SHOP	frecuencia	OSM
	comercio	CAD.COMMERCE	superficie	catastro
	agricultura	CAD.AGRICULTURE	superficie	catastro
	venues	CAD.VENUES	superficie	catastro
	religioso	CAD.RELIGION	superficie	catastro
	residencial	CAD.RESIDENTIAL	superficie	catastro
Recreativo	parque	PARK	frecuencia	OSM
	deportivo	CAD.SPORT	superficie	catastro
	deportivo	SPORT	frecuencia	OSM

Fuente: elaboración propia

La función de utilidad, $O_k(X, Y, d_{i(m)})$, se basa en la clasificación propuesta por Heyman, Law y Berghauser Pont (2018). Este índice representa una cantidad de oportunidades en una isócrona en el tiempo-distancia $d_{i(m)}$. Cada variable se agrega en función de su naturaleza, por ejemplo, para POIs se usa el conteo de elementos; en el caso de superficies construidas, se usa la suma de áreas, y en el caso de métricas de la malla urbana, se calculan mediante la suma de longitudes

de segmentos y densidades de calles.

Las 5 categorías dan lugar a tres grupos de índices (para una mayor información véase el Anexo I): :

- *Índices de transporte público y privado*: estos índices agregan los elementos de cada tipo (paradas de autobús, estaciones de metro, etc.) en una determinada impedancia tiempo-distancia. Las oportunidades de transporte privado utilizan la longitud en metros de las carreteras y la densidad de la red vial en metros cuadrados. Estas últimas medidas dan como resultado una mayor utilidad, al proporcionar una mejor conectividad en los suburbios de una ciudad, o una desutilidad, ya que implica mayores niveles de contaminación y ruido.
- *Actividad económica e Índice de servicios básicos*: se refieren al acceso a actividades económicas, servicios básicos, equipamiento residencial, empleo y ocio. Es habitual la concentración de muchas de estas variables en lugares específicos de la geografía. Por ejemplo, los alquileres de hoteles, alimentos, turismo y vacaciones generalmente se encuentran en grandes cantidades en áreas turísticas específicas. Se combinan medidas calculadas sobre puntos o superficies, estas últimas basadas en la suma del área total de ciertos usos del suelo (industrial, oficina, etc.). Estas medidas se usan habitualmente en el cálculo de índices de accesibilidad caminando (Frank *et al.*, 2010). El fundamento detrás de esta elección es que la cantidad de metros cuadrados de propiedades residenciales actuaría como un indicador indirecto de la población, mientras que la cantidad de metros cuadrados de oficinas (FAR), espacio industrial y comercial, son indicadores indirectos de las concentraciones brutas de empleo (Giuliano y Small, 1991).
- *Índices sociales y recreativos*: recogen variables socio-económicas relevantes y servicios recreativos, respectivamente.

4.2.2 Modelo de espacio discreto y granularidad flexible

La “maldición de la dimensionalidad”¹⁵ es habitual en los problemas de análisis espacial. Esta cuestión surge a la hora de especificar el modelo hedónico [4.1], por la gran cantidad de combinaciones de elementos que interactúan dentro de un área designada; en nuestro caso, las numerosas características de accesibilidad $O_k(X, Y, d_{i(m)})$ correspondientes al número de paradas de transporte público (autobús, tranvía, tren), actividades económicas y servicios públicos (comercio, educación, salud), por ejemplo.

¹⁵Es decir las consecuencias negativas de trabajar en contexto donde existe un gran número de variables (Hastie *et al.*, 2017).

La reducción de dimensiones se logra convirtiendo el espacio de coordenadas continuo (generalmente representado como un vector de números reales: latitud y longitud) en un espacio discreto, como hace por ejemplo Uber (2018) con su sistema H3. Nuestro enfoque se basa en un sistema de cuadrícula discreta global (DGGS¹⁶), construido sobre un espacio teselado de formas poligonales (Bondaruk, 2019). Como resultado, este DGGS crea un identificador numérico, denominado índice, para cada celda del mosaico que representa de forma única cada posición en el espacio. El tamaño de celda no debe ser ni demasiado fina ni demasiado gruesa, para mantener un equilibrio entre coste de cálculo y precisión. Por razones técnicas, se decide utilizar la teselación *H3*¹⁷ desarrollada por la empresa Uber (2018), su ventaja reside en su facilidad de uso y su disponibilidad en numerosas bases de datos y lenguajes de programación.

La principal ventaja de reducir las dimensiones de análisis es la importante disminución de los tiempos de cálculo de los algoritmos, que permiten mantenerlos dentro de unos límites razonables. Adicionalmente, se limita el número de áreas consideradas, restringiendo más aún el universo de trabajo a las semillas de ubicación de tipo residencial.

Nuestra hipótesis para decidir tomar como válido este un espacio simplificado, se basa en el fundamento de que los incrementos de utilidad marginales son irrelevantes dentro de áreas pequeñas (es decir, fuera de las celdas hexagonales). Por lo tanto, para una resolución 10 en *H3*, el error promedio en la distancia se traduciría en 50 metros que es en promedio menos de un minuto a pie y con una resolución 10 sería 150 metros mucho menos de un minuto en coche. Como las distancias de tiempo de viaje de menos de un minuto no supondrían una diferencia en utilidad percibida, se decide asumir una ligera pérdida de precisión en detrimento de las ganancias de rendimiento del proceso.

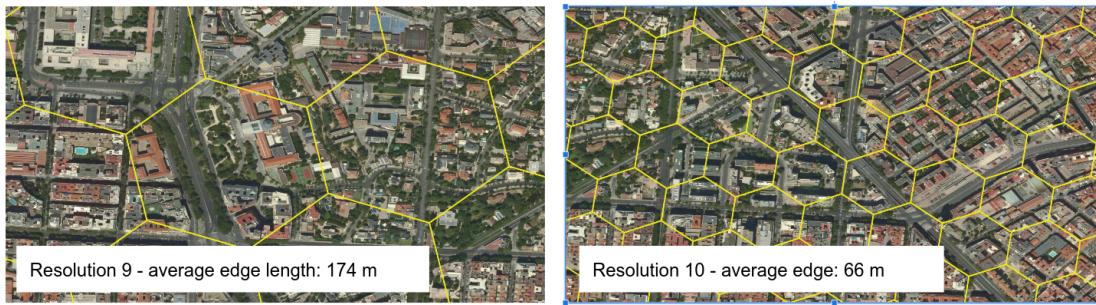
En el espacio de coordenadas teselado, los atributos de ubicación se precisan sobre una rejilla hexagonal multinivel para el conjunto de zonas semilla, correspondientes a todos los centroides de parcelas residenciales. En la Figura 4.3 se muestra el grado de granularidad de las medidas de accesibilidad, para la resolución del tiempo de viaje en automóvil 9 y el 10 para el modo de transporte a pie. La resolución 10, está formada por hexágonos con una longitud aproximada de arista de 66 metros¹⁸, para la resolución 9 la arista mide 174 metros.

¹⁶Discrete Global Grid System.

¹⁷El sistema de coordenadas H3 de Uber divide todo el planeta como un mosaico compuesto por hexágonos, y pentágonos, que se pueden agrupar a diferentes niveles de profundidad.

¹⁸En un hexágono regular, la longitud de arista mide lo mismo que el radio de la circunferencia sobre la que se inscriben los puntos del polígono.

Figura 4.3. Detalle de características de las zonas semilla utilizadas, para resoluciones H3 9 y 10



Fuente: elaboración propia.

4.2.3 Cálculo de índices gravitatorios

Para calcular los índices de accesibilidad ortogonales, A_{nk}^{tm*} , definitivos se selecciona el mejor índice de accesibilidad para cada una de las $K = 26$ variables presentadas en la Tabla 4.2. Esta selección parte del conjunto de índices sin procesar, A_{nk}^{tm} , calculados en función de las $O_{i(k)}(X, Y, d_{i(m)})$ características, para una serie de configuraciones de la función de desgaste exponencial β_m . Por tanto, un índice $Oportunidad_k(X, Y, d_{i(m)})$ es esencialmente una medida de una variable contenida dentro de una isócrona a tiempo-distancia $d_{i(m)}$, cuya medición se realiza acumulativamente a uno de los siguientes niveles: como conteo de *elementos*, como suma de *áreas*, suma de *longitudes* o como *densidad*, especificado como:

$$Oportunidad_k(X, Y, d_{i(m)}) = \sum_{o=1}^O M(X, Y, o_x, o_y, d_{i(m)}) \cdot C_k(o_x, o_y), \quad \forall o \in O, \quad [4.3]$$

$$M(X, Y, o_x, o_y, d_{i(m)}) = \begin{cases} 1, & \text{if distance } (X, Y) \rightarrow (o_x, o_y) \leq d_{i(m)} \\ 0, & \text{otherwise,} \end{cases}$$

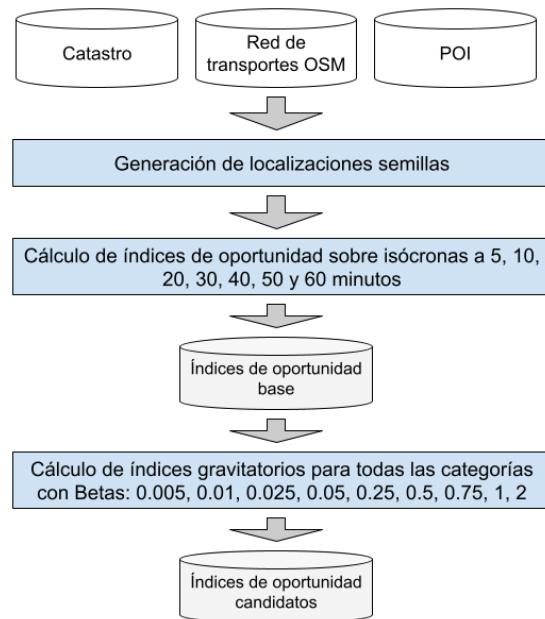
donde para cada elemento o en el universo completo de oportunidades O , con un par de coordenadas (o_x, o_y) , definimos una medida de contribución C_k , cuya definición depende de la función de agregación a aplicar en cada familia k (recuento, suma o densidad). $M(X, Y, o_x, o_y, d_{i(m)})$ es una función dicotómica utilizada para filtrar todas las oportunidades de contribución elegibles a una distancia $d_{i(m)}$.

El conjunto de $k = 1, \dots, K$ índices de accesibilidad óptimos A_{nk}^{tm*} , se obtiene eligiendo el β_m que maximiza la correlación con el errores de un modelo de regresión por mínimos cuadrados (denominado *naïve*), el cual omite las

variables de accesibilidad pero incluye el resto de atributos $S_{nj}^t, M_{nl}^t, D_{nk}^t$. Posteriormente, estos índices óptimos se transforman en un conjunto de índices de accesibilidad ortogonales, mediante el análisis de componentes principales. Esta transformación reduce la necesidad de tratamiento de covariables al eliminar la colinealidad entre los índices, mejorando así el rendimiento del modelo de regresión .

Todo el proceso tiene lugar en tres pasos y dos algoritmos como se observa en la Figura 4.4 y Figura 4.6. El primero, crea la familia de índices de accesibilidad brutos candidatos; mientras que el segundo, se encarga de seleccionar los de mejor desempeño mediante un enfoque heurístico y, posteriormente, realiza el análisis de componentes principales.

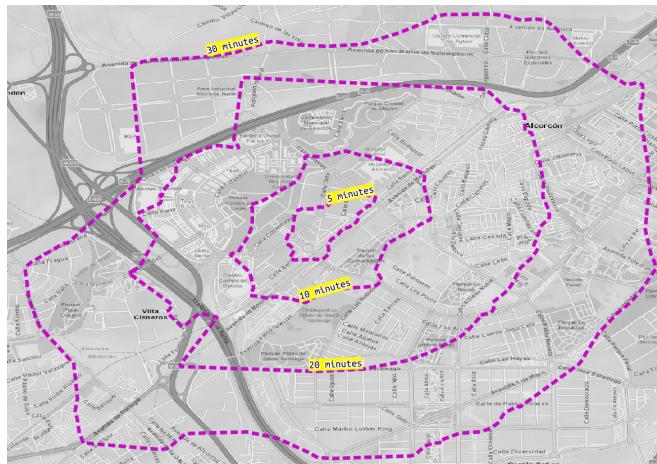
Figura 4.4. Parte I - Generación de índices de accesibilidad candidatos



Fuente: elaboración propia.

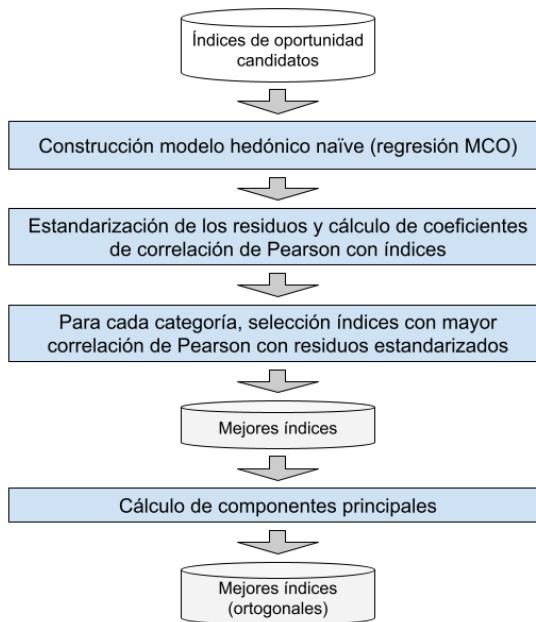
Sobre las ubicaciones de semilla se calculan las áreas isócronas asociadas para la serie de distancias de tiempo en los dos medios de transporte: $d_{i(m=coche)}$ y $d_{i(m=a pie)}$. La Figura 4.5 ilustra la forma de los anillos concéntricos definidos alrededor de una semilla específica.

Figura 4.5. Anillos de isócronas son las áreas accesibles a pie a 5, 10, 20, 30 minutos desde una ubicación semilla



Fuente: elaboración propia.

Figura 4.6. Parte II - Selección de los mejores índices de accesibilidad



Fuente: elaboración propia.

Posteriormente, para cada uno de los anillos, se calculan todos los índices de oportunidad $O_k(X, Y, d_{i(m)})$ y se consolidan en el índice de accesibilidad de ubicación, con distintos valores de β_m : 0.005, 0.01, 0.025, 0.05, 0.25, 0.5, 0.75, 1 y 2. Por lo tanto, se obtiene una familia de 9 índices de accesibilidad sin procesar para cada β_m , según la expresión:

$$A_{nk}^{tm} = \sum_{i(m)=1}^I O_{i(k)}(X, Y, d_{i(m)}) \cdot e^{-\beta_m \cdot d_{i(m)}}, m = \{\text{coche, caminar}\}, \beta_m = 1, \dots, 9 \quad [4.4]$$

Con todos los índices candidatos, 9 por variable de oportunidad, se inicia el segundo algoritmo (Figura 4.6) que selecciona aquellos que presentan un mejor desempeño potencial en el modelo de precios hedónicos.

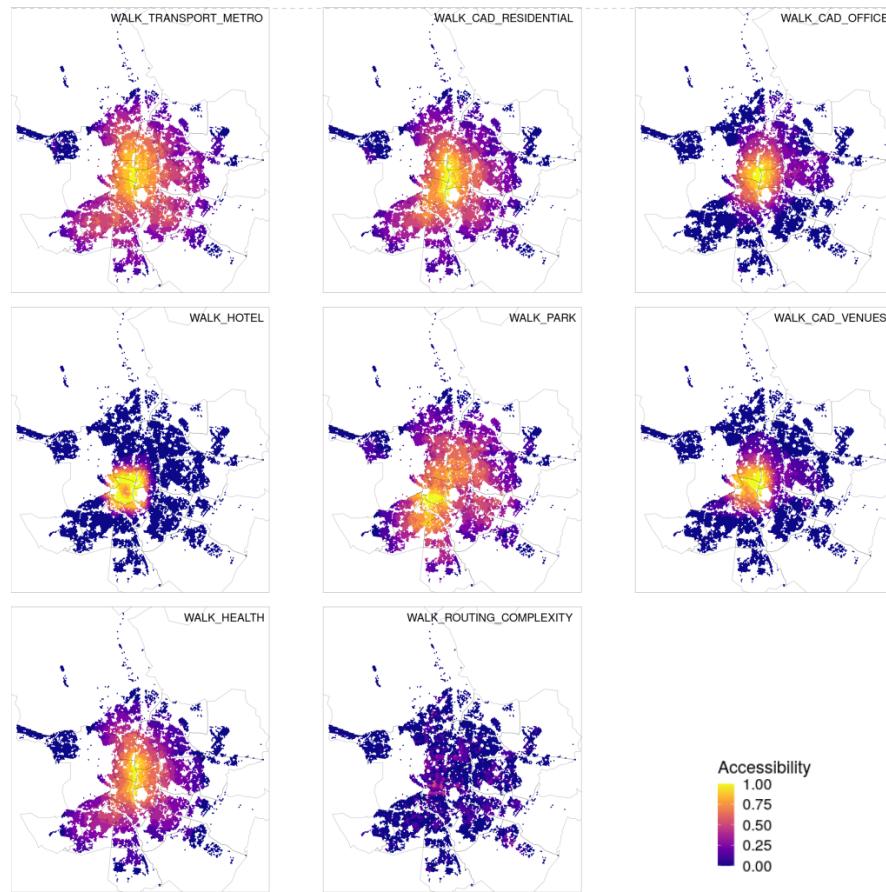
Para evitar la evaluación de todas las combinaciones posibles de variables y configuraciones, se sigue un enfoque heurístico univariante, el cual consiste en seleccionar la β_m que tiene la mayor correlación con los residuos de un modelo hedónico de precios *naïve* calculado sin información de ubicación. Fundamentado sobre la base de que en ausencia de variables de ubicación, presentadas en la Tabla 4.2, el residuo de un modelo debe mostrar un alto grado de correlación espacial (hipótesis corroborada en nuestra prueba empírica en el apartado 4.3.3). En un proceso no estacionario espacialmente, cualquier variable correlacionada con los residuos también estaría correlada con los atributos espaciales omitidos, por lo que sería un buen candidato como predictor del modelo.

El enfoque heurístico, además, selecciona la mejor configuración de cada índice accesibilidad individual (en este caso las β), con la expectativa de reducir el sesgo espacial del modelo *naïve*. Este procedimiento constituye una versión simplificada de un algoritmo de *boosting* (Friedman y Weinberg, 1981).

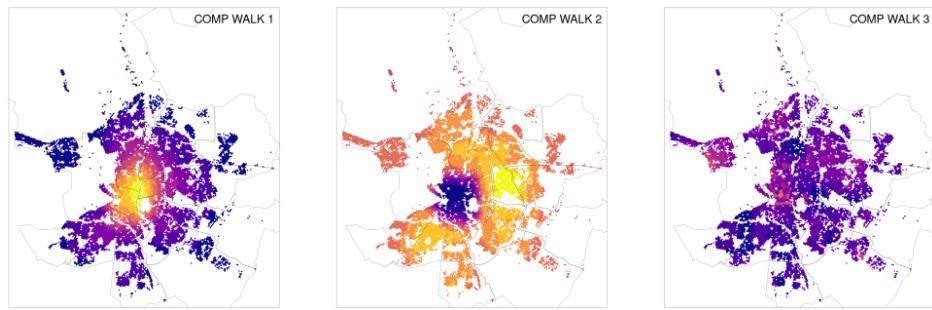
Dado que el conjunto de índices de accesibilidad óptimos A_{nk}^{tm*} pueden estar correlacionados entre sí, se aplica un modelo de análisis de componentes principales que elimina la colinealidad entre las variables (Abdi y Williams, 2010). La multicolinealidad no reduce el poder predictivo, pero es perjudicial en la inferencia estadística al reducir la interpretabilidad de los coeficientes, al igual que reduce la confiabilidad de la medida R^2 (Orford, 2017). En consecuencia, los componentes de accesibilidad obtenidos \hat{A}_{nk}^{tm*} son ortogonales entre sí y, por lo tanto, pueden emplearse para estimar el modelo hedónico de precios de vivienda sin preocuparse por la posible multicolinealidad de las variables.

En el Anexo 4d se presentan los β_m con mejores desempeños por cada índice, se observa que el modo coche exige un decaimiento exponencial más fuerte con un valor medio igual a $\beta_{coche} = 0,05$, mientras que el valor medio a pie es $\beta_{caminar} = 0,005$. Estos valores implican un factor de decaimiento para un viaje de 10 minutos equivalente a 39,35% y 4,89%, respectivamente (calculados como $1 - e^{-\beta_m \cdot 10 \text{ minutos}}$).

La Figura 4.7 muestra los índices de accesibilidad óptimos A_{nk}^{tm*} para las variables seleccionadas. Los colores más claros indican un mayor acceso a las oportunidades. Por ejemplo, los hoteles (*WALK HOTELS*) están muy concentrados en el centro de la ciudad, mientras que los servicios de salud están distribuidos de manera más uniforme en toda la ciudad (*WALK HEALTH*).

Figura 4.7. Índices de accesibilidad básicos

Fuente: elaboración propia.

Figura 4.8. Índices de accesibilidad ortogonales

Fuente: elaboración propia.

El PCA arroja 4 índices de accesibilidad ortogonales (\hat{A}_{nk}^{tm}) para el modo de transporte a pie, que representan el 88% de la variación de las medidas de accesibilidad en bruto. La Figura 4.8 muestra los tres primeros componentes (en términos de sus valores propios) para caminar. Estos componentes se nombran como *COMP_WALK*, tanto en las figuras como en los resultados de la regresión.

La interpretación de estos componentes se realiza a través del estudio de la

contribución de las distintas variables, de la Tabla 4.2, en las cargas del modelo PCA. Los grados de contribución, medidos como el valor absoluto de las cargas, se muestran en la Tabla 4.3, se puede asociar el grado de contribución de cada factor al conjunto de índices gravitatorios ortogonales. Además, para mejorar su legibilidad aplicamos una rotación Varimax a los valores originales (Kaiser, 1958). Esta transformación, basada en maximizar las varianzas de las cargas al cuadrado, mantiene la estructura original de los datos, aunque maximiza la distinción y diferenciación de correlaciones de variables y factores. Los resultados detallados del método de transformación se proporcionan en la siguiente Tabla 4.4 que presenta los valores propios, así como la extracción y las sumas rotadas de las cargas elevadas al cuadrado.

Como se puede observa en la Tabla 4.3, el primer componente principal *COMP_WAK_1* se refiere a áreas con un alto grado de servicios de ocio, áreas comerciales y bien conectadas con el transporte público y con una alta existencia de apartamentos de vacaciones. El segundo componente destaca el anillo exterior inmediato del centro de la ciudad, áreas urbanas residenciales prósperas desde un punto de vista urbano. En consecuencia, se aprecia una correlación negativa con los servicios turísticos, hoteles y restaurantes y oficinas, pero aún bien conectados y con una alta presencia de comercio y áreas residenciales. El tercer componente no es tan interpretable como los dos primeros, y destaca áreas específicas de la ciudad que presentan comportamientos diferentes a los generales dentro de sus distritos. Por ejemplo, se identifican valores altos en calles importantes como la calle principal de Madrid (Gran Vía) y áreas sujetas a fenómenos de gentrificación en el centro de la ciudad. Podemos ver esta combinación en la tabla de cargas, ya que este factor favorece características de áreas pequeñas en el centro de la ciudad: turísticas, hoteleras y comerciales, o de áreas más pequeñas con mucha oferta comercial pero menor conectividad con el transporte público que en el centro de la ciudad.

Tabla 4.3. Cargas de los componentes principales - Modo a pie

Componente	COMP	COMP	COMP	COMP
	WALK	WALK	WALK	WALK
	1	2	3	4
VACATIONAL	0.93	0.19		
TRANSPORT BUS	0.72	0.63		
TOURISM	0.92	0.23		
SHOP	0.85	0.44		
HOTEL	0.95	0.13		
FOOD	0.93	0.29		
CAD VENUES	0.91	0.35		
CAD RELIGION	0.77	0.59	0.12	
CAD PUBLIC	0.88	0.39	0.10	
CAD OFFICE	0.75	0.56	0.13	
CAD HOTEL	0.83	0.51		
CAD COMMERCE	0.70	0.68		
TRANSPORT METRO	0.54	0.74	0.15	
SPORT	0.18	0.86	0.10	
PARK	0.46	0.78		
HEALTH	0.64	0.70		
EDUCATION	0.52	0.80	0.16	
CAD SCHOMCO	0.57	0.77	0.15	
CAD RESIDENTIAL	0.58	0.76	0.13	
CAD INDUSTRY		0.85		
ROUTING	0.11	0.12	0.98	
COMPLEXITY				
CAD SPORT		0.17		0.97
TRANSPORT TRAIN	0.35	0.28		
ROUTING HIGHWAY	0.21	0.25		

Fuente: elaboración propia

Tabla 4.4. Tabla de sedimentación de los componentes principales - modo a pie

Componente	Autovalores iniciales			Suma de las cargas al cuadrado			Suma de las cargas al cuadrado rotadas		
	Total	% Var.	% Acuml.	Total	% Var.	% Acuml..	Total	% Var.	% Acuml..
COMP WALK 1	18.080	0.723	0.723	18.080	0.723	0.723	10.941	0.438	0.438
COMP WALK 2	2.145	0.086	0.809	2.145	0.086	0.809	8.239	0.330	0.767
COMP WALK 3	1.007	0.040	0.849	1.007	0.040	0.849	1.023	0.041	0.808
COMP WALK 4	0.888	0.036	0.885						

Fuente: elaboración propia

Tabla 4.5. Cargas de componentes principales - Modo coche

Componente	COMP	COMP	COMP	COMP
	CAR 1	CAR 2	CAR 3	CAR 4
TRANSPORT BUS	0.95	0.24		0.15
TRANSPORT AIRPORT	0.71	0.56		0.15
TOURISM	0.91	0.34		0.13
SHOP	0.93	0.31		0.14
HOTEL	0.96	0.19		0.11
HEALTH	0.95	0.25		0.14
CAD VENUES	0.94	0.30		0.15
CAD URBAN LAND	0.81	0.45		0.17
CAD SCHOMCO	0.90	0.37		0.15
CAD RELIGION	0.93	0.32		0.14
CAD PUBLIC	0.92	0.35		0.15
CAD OFFICE	0.90	0.37		0.14
CAD INDUSTRY	0.79	0.54		0.15
CAD HOTEL	0.94	0.30		0.14
CAD COMMERCE	0.92	0.33		0.15
CAD AGRICULTURE	0.68	0.54		0.13
TRANSPORT TRAIN	0.66	0.68		0.14
SPORT	0.49	0.85		0.10
PARK	0.50	0.85		0.11
EDUCATION		0.97		
CAD SPORT	0.29	0.74		
CAD RESIDENTIAL	0.49	0.85		0.11
ROUTING HIGHWAY			1	
ROUTING	-0.41	-0.20		-0.89
COMPLEXITY				

Fuente: elaboración propia

Tabla 4.6. Tabla de sedimentación de componentes principales - Modo coche

Componente	Autovalores iniciales			Suma de las cargas al cuadrado			Suma de las cargas al cuadrado rotadas		
	Total	% Var.	% Acuml.	Total	% Var.	% Acuml..	Total	% Var.	% Acuml..
COMP CAR 1	19.128	0.797	0.797	19.128	0.797	0.797	14.040	0.585	0.585
COMP CAR 2	2.273	0.095	0.892	2.273	0.095	0.892	6.412	0.267	0.852
COMP CAR 3	0.994	0.041	0.933						
COMP CAR 4	0.648	0.027	0.960						

Fuente: elaboración propia

4.2.4 Especificación de la validación de modelos

Para comprobar la robustez de los resultados aplicado a modelos de precios hedónicos de la vivienda, se consideran varias metodologías sobre tres variaciones del conjunto de datos:

- *Referencia*: utiliza los datos originales de los anuncios de Idealista presentados sin las variables de accesibilidad eliminadas. Es similar al modelo *naïve* mencionado en el proceso de construcción de índices, pero con un mayor número de atributos, en particular algunos referidos al ámbito local como medidas de dinámicas del mercado a nivel de distrito¹⁹ y datos sociodemográficos²⁰. Se asume este modelo como un escenario de referencia para comparar la mejora económica de los atributos de los índices de accesibilidad propuestos.
- *Dummy*: esta especificación modela la ubicación mediante la inclusión de una variable binaria ficticia para cada distrito (*dummy* de ubicación). Cada una de estas variables captura la contribución marginal de cada área geográfica en los precios.
- *Accesibilidad*: la especificación más completa e incluye los índices de accesibilidad ortogonal obtenidos a través del análisis de componentes principales.

Para estimar los precios de la vivienda se comparan varias de técnicas de modelado: por una parte, modelos econométricos tradicionales basados en regresión, y por otra, modelos de aprendizaje automático. Los primeros se usan predominantemente en el mundo académico, y los segundos, en la industria en valoraciones masivas de inmuebles (Valier, 2020).

El enfoque econométrico, que se usa el modelo MCO estándar, permite probar si los índices de ubicación ortogonal funcionan bien en términos de signos esperados²¹ y significación estadística. Por este motivo, se ha evitado el uso de enfoques menos interpretables de forma global como las regresiones locales. Dado que el objetivo principal de nuestro estudio es maximizar la precisión en la predicción de los precios de la vivienda, en el enfoque de aprendizaje automático se utilizan modelos de árboles de regresión ensamblados. Estos modelos son capaces de superar algunas de las limitaciones de los modelos de regresión, como la colinealidad o heterocedasticidad, aunque son más difíciles de ajustar y tienen más riesgo de sobreajustarse.

¹⁹Se incluye la proporción de viviendas en compra y alquiler y número medio de contactos por anuncio en el distrito.

²⁰Se incluye el nivel de educación y densidad de población del distrito.

²¹El signo se refiere a si la contribución al precio es positiva (mayor valor del índice implica mayores precios) o negativa (mayor valor, menores precios).

Se diseñan cuatro métodos, los dos primeros puramente econométricos y los dos últimos basados en técnicas de aprendizaje automático:

- Regresión de mínimos cuadrados ordinarios (MCO). Calcula los parámetros de una función lineal minimizando la suma de los residuos cuadrados.
- Modelos lineales generalizados regularizados Lasso y Elastic-Net (LERG). Esta aproximación también estima un modelo de regresión lineal usando un descenso de gradiente sobre el que se aplica un proceso de regularización²². El método realiza una regularización de penalizaciones L1 (*Lasso*) y L2 (*Ridge*), que es especialmente adecuado para casos como este, con un gran número de regresores.
- Árboles de partición recursiva o árboles RP. Propuesto originalmente por Breiman (1984) y basado en árboles binarios, es un modelo de árbol de regresión de tipo CART²³.
- *Random Forests*: este método²⁴ que estima la magnitud de la regresión como un consenso de varios modelos (Breiman, 2001).

Los cuatro métodos tienen como variable dependiente los precios de la vivienda en €/m² construidos. Para evitar sesgos de muestreo, se utiliza una estrategia de remuestreo de tipo validación cruzada con K=5 (Hastie y Tibshirani, 2017; LeCun *et al.*, 2015), el valor de K se decide sobre el trabajo de Arlot (2010) que argumenta empíricamente que el óptimo se encuentra entre 5 y 10. En este proceso, los datos se mezclan y se dividen en 5 grupos de igual tamaño sobre los que se repite el experimento 5 veces. Las métricas finales del modelo se calculan como el promedio de las medidas para las 5 iteraciones.

La configuración de hiperparámetros de cada algoritmo se determina mediante una búsqueda en cuadrícula (*grid search*). Este proceso prueba varias configuraciones y selecciona los parámetros con mejor rendimiento (para más detalles sobre la parametrización véase el Anexo 4c de este capítulo).

La Tabla 4.7 recoge los resultados el resultado de un modelo hedónico estándar basado en mínimos cuadrados. Se observa que los coeficientes de cada componente de accesibilidad exhibe el signo esperado y es estadísticamente significativo. Se toman los cuatro primeros componentes para el modo a pie junto con otro del modo coche, explicando los primeros el 88% de la varianza y el segundo el 4%. Se decide tomar el tercer componente del coche, al mantener la restricción de ortogonalidad con las covariables de accesibilidad del modo a pie (mostrando un coeficiente de correlación de Pearson promedio para los

²²Se ha utilizado el paquete de R *glmnet* (Simon *et al.*, 2011).

²³El árbol de tipo CART (Breiman *et al.*, 1984) es un algoritmo de árboles de decisión y regresión que se construye de forma recursiva a través de un árboles de decisión binarios, en este caso se ha utilizado el paquete de R *rpart* (Therneau *et al.*, 2015).

²⁴Se utiliza el paquete *ranger* de R (Wright y Ziegler, 2015).

componentes de caminar de 0,16 en términos absolutos).

Tabla 4.7. Coeficientes de regresión por mínimos cuadrados

	coeficiente	std.error	t valor	Pr(> t)	signif.
(Intercept)	3035.07	23.16	131.06	< 2e-16	***
CONSTRUCTEDAREA	-0.81	0.05	-14.90	< 2e-16	***
FLATLOCATION	159.70	4.41	36.24	< 2e-16	***
ROOMNUMBER	-160.94	2.21	-72.76	< 2e-16	***
ISSTUDIO	-57.70	16.67	-3.46	5.38e-04	***
ISPENTHOUSE	443.48	6.90	64.25	< 2e-16	***
HASLIFT	508.33	4.24	119.92	< 2e-16	***
MAXBUILDINGFLOOR	20.32	0.63	32.18	< 2e-16	***
HASANNEX	249.37	2.68	93.17	< 2e-16	***
COMP_WALK_1	324.17	1.87	173.26	< 2e-16	***
COMP_WALK_2	58.61	1.00	58.75	< 2e-16	***
COMP_WALK_3	111.91	1.90	58.81	< 2e-16	***
COMP_WALK_4	168.62	1.77	95.35	< 2e-16	***
COMP_CAR_3	-736.53	19.97	-36.89	< 2e-16	***
RENTSALE_RATIO	139.23	6.04	23.07	< 2e-16	***
ONMARKET_SALE	4191.73	115.98	36.14	< 2e-16	***
ONMARKET_RENT	3139.32	81.44	38.55	< 2e-16	***
DEMAND	-27.69	0.25	-111.29	< 2e-16	***
PERIOD_2018_01_31	-353.88	7.35	-48.14	< 2e-16	***
PERIOD_2018_02_28	-306.01	7.39	-41.39	< 2e-16	***
PERIOD_2018_03_31	-258.98	7.30	-35.49	< 2e-16	***
PERIOD_2018_04_30	-210.42	7.36	-28.59	< 2e-16	***
PERIOD_2018_05_31	-163.17	7.37	-22.15	< 2e-16	***
PERIOD_2018_06_30	-129.71	7.31	-17.74	< 2e-16	***
PERIOD_2018_07_31	-86.55	7.24	-11.96	< 2e-16	***
PERIOD_2018_08_31	-49.55	7.21	-6.87	6.28e-12	***
PERIOD_2018_09_30	-35.90	7.38	-4.87	1.14e-06	***
PERIOD_2018_10_31	-10.67	7.23	-1.48	1.40e-01	
PERIOD_2018_11_30	0.04	6.85	0.01	9.96e-01	
CADASTRALQUALITYID_1	509.35	25.55	19.93	< 2e-16	***
CADASTRALQUALITYID_2	272.94	19.16	14.25	< 2e-16	***
CADASTRALQUALITYID_3	92.66	17.56	5.28	1.32e-07	***
CADASTRALQUALITYID_4	-126.82	17.23	-7.36	1.84e-13	***
CADASTRALQUALITYID_5	-358.96	17.33	-20.71	< 2e-16	***
CADASTRALQUALITYID_6	-363.26	17.40	-20.88	< 2e-16	***
CADASTRALQUALITYID_7	-371.33	17.78	-20.89	< 2e-16	***
CADASTRALQUALITYID_8	-433.18	21.10	-20.53	< 2e-16	***

Fuente: elaboración propia

Signif.: *** 0,001 ** 0,01 * 0,05 . 0,1 1. Error residual estándar: 921 sobre 345.638 grados de libertad. R²: 0,71089, R² ajustado: 0,71086. estadístico F: 23606 sobre 36 y 345.601 grados de libertad, p-valor: < 2.2e-16. Num. observaciones: 345.638

El primer componente, *COMP_WAK_1*, muestra una correlación positiva con el precio en los coeficientes de regresión MCO. Algo que se puede corroborar observando la disposición espacial de valores del mismo en la Figura 4.12 del Anexo 4d, y los precios generales de la Figura 4.11 del mismo anexo. El centro de la ciudad muestra valores más altos, ya que comprende áreas con una alta concentración de establecimientos turísticos, comercios y paradas de transporte público. El resto de covariables relacionadas con características estructurales, de mercado y de tiempo, presentan los signos esperados y son todas significativas al nivel del 0,1%. Es reseñable que el signo y los valores de las variables *dummy* de tiempo son crecientes, lo que es coherente con el aumento general en los precios a lo largo de 2018.

La Tabla 4.8 presenta las diferentes métricas que se utilizan en el estudio para comparar el rendimiento del modelo. Estas métricas son comunes y proporcionan una representación estándar de precisión y poder predictivo. El rendimiento de cada método se mide en términos absolutos a través del error absoluto medio (*MAE*), y el error absoluto mediano (*MedAE*). También se mide en términos relativos por medio del error porcentual absoluto medio (*MAPE*). También reportamos el coeficiente de determinación, R^2 , y el *uplift* de los modelos calculando la reducción en el *MAPE* con respecto al modelo de referencia sin localización. Este último valor representa la ganancia de incluir los índices de accesibilidad, sobre ignorar toda la información sobre los atributos espaciales.

Tabla 4.8. Métricas para evaluar el ajuste del modelo y su precisión

Medida	Fórmula de cálculo
Error absoluto medio	$MAE = \frac{1}{n} \sum_{t=1}^n e_t^2$
Error absoluto mediano	$MedAE = median e_i $
Error absoluto medio porcentual	$MAPE = \frac{1}{N} \sum_{i=1}^N \frac{ e_i }{x_i}$
Error absoluto mediano porcentual	$MedAPE = median \frac{ e_i }{x_i}$
R^2	$R^2 = 1 - \frac{\sigma_{error}^2}{\sigma^2}$

Fuente: elaboración propia

4.3 Resultados

A continuación se evalúa el ajuste de los modelos utilizando una aproximación general, para posteriormente discutir la validez de la especificación zonal desde el ángulo del tratamiento de la heterogeneidad espacial y la dependencia espacial (Anselin y Rey, 2014). Para ello se evaluará la capacidad de los modelos de generalizar y reducir la autocorrelación espacial, cuando usan las medidas de accesibilidad.

4.3.1 Ajuste de los modelos

La Tabla 4.9 resume el rendimiento de cada método con respecto a las métricas elegidas. Se mide el grado de mejora de cada conjunto de datos en función de la mejora en precisión con respecto al modelo sin localización. Independientemente del método de estimación, el modelo que incluye los índices de accesibilidad ortogonales (Accesibilidad) supera a su ubicación contraparte de *dummies* (Dummy), con niveles de mejora similares en todos los enfoques.

Tabla 4.9. Comparativa con validación cruzada con 5 mezclas

Método	Modelo	MAE	MedAPE	MAPE	R ²	Mejora
MCO	Referencia	714.39	16.5%	22.5%	0.68	
	Dummy	642.02	14.6%	19.3%	0.72	10.1%
	Accesibilidad	624.19	14.3%	18.8%	0.74	12.6%
LERG	Referencia	714.03	16.5%	22.5%	0.68	
	Dummy	641.76	14.6%	19.3%	0.72	10.1%
	Accesibilidad	623.92	14.3%	18.8%	0.74	12.6%
Árbol RP	Referencia	711.11	16.6%	22.1%	0.67	
	Dummy	711.11	16.6%	22.1%	0.67	
	Accesibilidad	701.77	16.5%	21.9%	0.68	1.3%
R Forests	Referencia	444.39	9.1%	13.5%	0.85	
	Dummy	393.09	6.8%	12.0%	0.85	11.5%
	Accesibilidad	350.11	5.6%	10.7%	0.88	21.2%

Fuente: elaboración propia

Se aprecia un rendimiento razonablemente bueno de la especificación básica sin atributos de ubicación específicos (*Referencia*), que se podría explicar porque usa una mínima información del ámbito de mercado y demográfico. Por tanto este buen desempeño de los métodos de regresión MCO y LERG, significa que

estas variables están capturando una parte importante de la influencia de los atributos de ubicación en el precio. Por su parte, los árboles simples (es decir, el árbol RP) aparentemente no son capaces de generalizar las interacciones de ubicación con este conjunto de variables, muy probablemente debido a que no son lo suficientemente complejos como para incorporar las características de ubicación en detrimento de otras más significativas para el algoritmo, como las estructurales o las de mercado. Sin embargo, *Random Forests* presenta los mejores resultados de rendimiento con diferencia.

A pesar del número reducido de variables utilizadas en este estudio, los modelos muestran niveles de precisión similares o incluso superiores a casos con un mayor número de variables explicativas. Como por ejemplo, el estudio para la ciudad de Madrid de Del Cacho (2010), utilizando datos del portal inmobiliario pero apoyándose en una muestra más pequeña (25.415 observaciones), obtiene un error porcentual medio del 15,25%.

4.3.2 Capacidad de generalizar espacialmente

Las métricas del modelo presentadas no ofrecen mucha información acerca de su eficacia para modelar el proceso de precios en la geografía, ni miden su capacidad de generalizar espacialmente. Por ello se han evaluado los métodos mediante una validación cruzada espacial, ya que los métodos de remuestreo aleatorio no son adecuados para el estudio de los procesos geográficos (Meyer *et al.*, 2019). Este método, es similar a una validación cruzada general pero considerando mezclas no superpuestas geográficamente. Se establecen 5 áreas de estudio, y para cada iteración se reserva 1 para la validación y las 4 restantes para construir el modelo. La principal implicación de este enfoque es que los modelos se construyen con datos de ubicaciones diferentes a las consideradas en la validación. Por lo que, si un modelo se ajusta a los datos de validación, se puede concluir que sus atributos de ubicación están correctamente modelados a través de los índices de accesibilidad.

La Tabla 4.10 contiene los resultados del ejercicio de validación cruzada espacial, midiendo el grado de mejora de cada modelo por la precisión incremental obtenida con respecto a la especificación sin información geográfica (*Referencia*). Como era de esperar, se observa un menor grado de precisión en las diferentes métricas, pero se aprecia una mejora positiva en todos los casos al incluir información de zona.

La mejora sustancial en desempeño del modelo de *Accesibilidad* es una clara señal de la capacidad de estos índices para capturar el efecto de las diferentes interacciones de ubicación en los precios de la vivienda. Se observa que la mejora

relativa obtenida es mucho mayor para los modelos de regresión lineal (MCO y LERG), ofreciendo un resultado muy similar al de las técnicas de aprendizaje automático.

El hecho de que *Random Forests* no muestre una mayor ventaja sobre los modelos lineales, sugiere la posibilidad de sobreajuste espacial, aunque la refutación o confirmación de esta hipótesis requiere un estudio específico en profundidad. En cualquier caso, implica que este modelo puede ajustarse a la interacción existente entre los atributos de ubicación, pero es incapaz de modelar nuevos patrones como los que se encuentran en los datos de validación. Esta incapacidad para aprender y proporcionar un marco general para explicar las interacciones espaciales finalmente se muestra en forma de errores porcentuales absolutos medios más altos.

Con estos resultados, se puede concluir que para áreas con nuevas configuraciones urbanas, podría preferirse confiar en modelos lineales más parsimoniosos (y más simples).

Tabla 4.10. Comparativa con validación cruzada espacial 5-Fold

Método	Modelo	MAE	MedAE	MAPE	Mejora
MCO	Referencia	1175.24	1019.38	43.3%	
	Dummy	913.90	764.53	32.2%	22.2%
	Accesibilidad	687.27	556.28	24.1%	41.5%
LERG	Referencia	1175.37	1020.08	43.3%	
	Dummy	1089.73	963.30	41.9%	7.3%
	Accesibilidad	685.71	554.80	24.0%	41.7%
Árbol RP	Referencia	1095.15	890.73	38.6%	
	Dummy	1094.49	904.92	33.7%	0.1%
	Accesibilidad	824.82	646.99	26.9%	24.7%
R Forests	Referencia	1076.90	892.53	38.5%	
	Dummy	1004.60	830.61	34.2%	6.7%
	Accesibilidad	730.29	567.60	24.5%	32.2%

Fuente: elaboración propia

La Tabla 4.11 compara el comportamiento de los errores absolutos y en porcentaje para los modelos con los tres enfoques de especificación de variables de zona (*Referencia*, *Accesibilidad* y *Dummy*). La superioridad de las medidas de accesibilidad, tanto en las validaciones cruzadas normales como en las espaciales

es evidente.

Tabla 4.11. Resumen de MAPE con y sin características espaciales

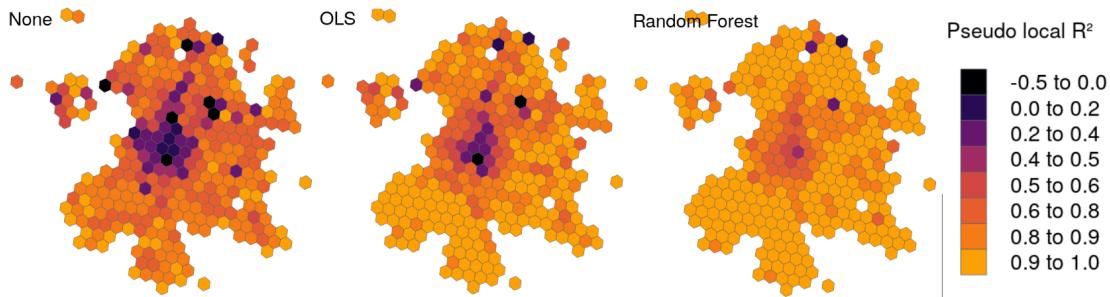
Algoritmo	Accesibilidad		Ninguno
	Val. Cruzada	Val. Cruzada	Val. Cruzada
		Espacial	Espacial
MCO	18.8%	24.1%	43.3%
LERG	18.8%	24.0%	43.3%
Árbol RP	21.9%	26.9%	38.6%
R Forests	10.7%	24.5%	38.5%

Fuente: elaboración propia

Para comprobar si el modelo captura el proceso espacial sobre el municipio de Madrid, se mide la bondad de ajuste a nivel espacial utilizando un *pseudoR*² local, sobre una rejilla hexagonal H3²⁵. Ese estadístico compara la relación de los residuos del modelo para cada tesela de la malla ($\varepsilon_{x,y}$) dividido entre la varianza local de la variable objetivo, referida al centroide de cada región hexagonal, calculada como:

$$\text{pseudo local } R^2 = 1 - \frac{\sigma^2(\varepsilon_{x,y})}{\sigma^2(\text{precio}/m^2)} \quad [4.5]$$

Figura 4.9. Pseudo R^2 para la ciudad de Madrid



Fuente: elaboración propia.

Como se observa en la Figura 4.9, el R^2 disminuye cuando se usan los índices de accesibilidad ortogonales, siendo esta reducción mayor para el modelo de Random Forests. Aunque este resultado puede parecer poco intuitivo, en nuestra opinión, se debe a que Random Forest puede manejar múltiples tipos de interacción espacial de ubicación y precio, especialmente los no lineales. Además, supera una

²⁵Se utiliza una resolución 8 para el cálculo, se amplia la región para asegurar un número mínimo de observaciones.

limitación importante de MCO que especifica una regla de interacción única para cada todas las áreas, mientras que el modelo de árboles es capaz de establecer reglas de árboles particulares para las diferentes áreas, ajustando así estas reglas cuando es necesario.

También se aprecia que el centro de la ciudad es más propenso a producir un R^2 más bajo. Este resultado puede deberse a la naturaleza de este submercado, ya que no se comporta como un área residencial pura en comparación con el resto de los mercados de la ciudad, lo que resulta en una mayor variabilidad en los precios. Los usos de los inmuebles residenciales en el centro de la ciudad de Madrid son mixtos, incluyendo no solo el residencial, sino también el alquiler vacacional de corta duración y fines profesionales. Sin embargo, en general, se puede concluir que la introducción de índices de accesibilidad da como resultado una notable ganancia en la bondad de ajuste, siendo mucho mejor para Random Forests que en el resto. Aún así, independientemente del método usado, la mayoría de las áreas céntricas muestran a un mayor grado de varianza no capturada por el modelo, debida a la existencia de variables omitidas.

4.3.3 Control de la autocorrelación espacial

Para confirmar la hipótesis de que el uso de índices de accesibilidad ortogonales es capaz de capturar el efecto de las interacciones de ubicación en los precios de la vivienda, se mide el grado de autocorrelación de los residuos del modelo. Si los índices de accesibilidad capturan la influencia de la ubicación, convertirían los residuos de los modelos de precios hedónicos en un proceso espacialmente estacionario (Dubin, 1998). Expresado desde otro ángulo, si los modelos están perfectamente especificados, en términos de las variables consideradas, capturarán el efecto de las características geográficas sobre los precios de la vivienda y, por lo tanto, los residuos del modelo no estarán correlacionados con los atributos de ubicación.

En los patrones espaciales se espera que las observaciones cercanas compartan características similares y difieran de las más alejadas. El coeficiente de autocorrelación espacial de Moran (1950), denominado I en la expresión [4.6], es una extensión del coeficiente de correlación producto-momento de Pearson que mide la similitud de las variables en el espacio. En este sentido, Zhu y Zhang (2021) desarrollaron recientemente un análisis de la dispersión de los precios de la vivienda basado en este índice. En su forma más simple, el índice de Moran se calcula asignando pesos a las observaciones vecinas: 1 para ubicaciones limítrofes y 0 en caso contrario. Estos pesos constituyen la llamada función de vecindad, que se puede definir en términos de matrices de proximidad que

utilizan diferentes criterios (por ejemplo, distancias por pares entre ubicaciones). El índice de Moran se define de la siguiente manera:

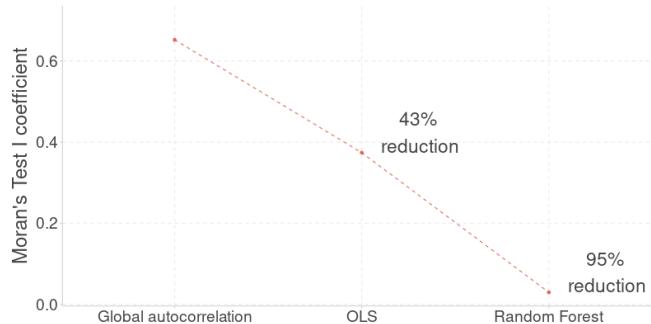
$$I = \frac{\frac{n}{S_0} \sum_{i=1}^n \sum_{j=1}^n w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad [4.6]$$

donde w_{ij} es el peso entre la observación i y j ; x_i y x_j sus respectivas variables de interés; y S_0 la suma de todos los w_{ij} : $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$, $w_{ii} = 0$.

La referencia usadas para la comparación es la autocorrelación global de residuos que se estima a partir de un modelo MCO simple que excluye las variables de ubicación, es decir, el estadístico I de Moran para el modelo *Referencia*. La autocorrelación global obtenida es 0,652, que se reduce a 0,374 cuando se añaden los índices de accesibilidad ortogonales en la estimación MCO, es decir, el modelo de *Accesibilidad*. Para *Random Forests*, el uso de los indicadores de accesibilidad logra reducir el estadístico I de Moran a casi cero, un 0,030. En el Anexo 4e se muestran con detalle las medidas del estadístico I para los tres modelos.

Nuestro caso coincide con los resultados de Morali e Ylmaz (2020), en cuyo caso el uso de métricas de accesibilidad permite reducir de forma significativamente la autocorrelación en los residuos del modelo. Sin embargo, nuestra aproximación se puede considerar superior al no requerir de una especificación a-priori de las variables de zona.

Figura 4.10. Moran I - Reducción de la autocorrelación de los residuos del modelo Random Forest sobre MCO



Fuente: elaboración propia.

En este capítulo se ha presentado un método general para estimar un conjunto de variables que sintetizan la utilidad de una zona, en base al acceso de servicios en la misma. Dicha utilidad, cuya representación numérica se relaciona con su

contribución a los precios del suelo, permitirá desarrollar los modelos hedónicos más completos. En el siguiente capítulo se aplicarán las medidas de accesibilidad en la construcción de los modelos hedónicos detallados de oferta.

Anexo 4a. Variables usadas en el cálculo de índices

A continuación se muestran las variables utilizadas para la construcción de los índices de oportunidad base.

Transporte público y privado

- **TRANSPORT.BUS:** Número de paradas de autobús dentro de la isócrona.
- **TRANSPORT.METRO:** Número de paradas de metro dentro de la isócrona.
- **TRANSPORT.TRAIN:** Número de estaciones de tren dentro de la isócrona.
- **TRANSPORT.AIRPORT:** Número de aeropuertos dentro de la isócrona.
- **ROUTING.HIGHWAY:** Metros lineales de autopista en la isócrona.
- **ROUTING.COMPLEXITY:** Complejidad de las vías por metro cuadrado, calculada como la suma de la longitud de los segmentos de vía por metro cuadrado de la isócrona.

Actividad económica y servicios básicos

- **CAD.URBANLAND:** Metros cuadrados de suelo sin construir.
- **HOTEL:** Número de POIs de OSM²⁶ de tipo Hotel Hotel.
- **FOOD:** Número de POIs de OSM de tipo Comida.
- **TOURISM:** Número de POIs de OSM de tipo Turismo.
- **VACATIONAL:** Número de viviendas vacacionales en plataformas *online*.
- **MONEY:** Número de oficinas bancarias o cajeros en OSM.
- **CAD.PUBLIC:** Número de metros cuadrados de uso público en Catastro.
- **CAD.SCHOMCO:** Metros cuadrados de centros educativos en Catastro.
- **EDUCATION:** Número de POIs de tipo educativo en OSM.
- **CAD.HEALTH:** Metros cuadrados de centros de salud en Catastro.
- **SHOP:** Número de POIs de tipo comercial en OSM.
- **CAD.COMMERCE:** Metros cuadrados de uso comercial en Catastro.
- **CAD.INDUSTRY:** Metros cuadrados de uso industrial en Catastro.
- **CAD.OFFICE:** Metros cuadrados dedicados a oficinas en Catastro.
- **CAD.AGRICULTURE:** Metros cuadrados de uso agrícola en Catastro.
- **CAD.VENUES:** Metros cuadrados dedicados centros de eventos.

Social y recreativo

- **CAD.RELIGION:** Metros cuadrados de uso religioso en Catastro.
- **CAD.RESIDENTIAL:** Metros cuadrados de uso vivienda en Catastro.
- **PARK:** Número parques o zonas verdes en OSM.
- **CAD.SPORT :** Metros cuadrados de uso deportivo en Catastro.
- **SPORT :** Número de POIs de tipo deportivo en OSM.

²⁶POI hace referencia a un punto de interés en Open Street Map.

Anexo 4b. Selección de betas

La Tabla 4.12 recoge las penalizaciones β exponenciales, aplicadas en la construcción de las medidas gravitatorias de oportunidad.

Se incluye una columna con la mejora de la β seleccionada con respecto a la peor configuración (calculada como la diferencia en términos absolutos del coeficiente de Pearson).

Tabla 4.12. Mejores Betas por medida de oportunidad

Índice	Modo	Beta	Mejora	Modo	Beta	Mejora
CAD.AGRICULTURE	COCHE	0,005	0,1134	A PIE	0,005	0,0647
CAD.COMMERCE	COCHE	0,050	0,0260	A PIE	0,005	0,0622
CAD.HOTEL	COCHE	0,050	0,0252	A PIE	0,005	0,0739
CAD.INDUSTRY	COCHE	0,010	0,0637	A PIE	0,250	0,0077
CAD.OFFICE	COCHE	0,050	0,0559	A PIE	0,005	0,0425
CAD.PUBLIC	COCHE	2.000	0,0473	A PIE	0,005	0,0749
CAD.RELIGION	COCHE	0,050	0,0389	A PIE	0,005	0,0386
CAD.RESIDENTIAL	COCHE	0,050	0,0547	A PIE	0,005	0,0560
CAD.SCHOOLS	COCHE	0,050	0,0602	A PIE	0,005	0,0293
CAD.SPORT	COCHE	0,010	0,0611	A PIE	0,005	0,0351
CAD.URBAN_LAND	COCHE	0,010	0,0709			
CAD.VENUES	COCHE	0,005	0,0504	A PIE	0,005	0,0540
EDUCATION	COCHE	0,050	0,0774	A PIE	0,005	0,0708
HEALTH	COCHE	0,250	0,0987	A PIE	0,005	0,0407
HOTEL	COCHE	2.000	0,0279	A PIE	0,005	0,1014
PARK	COCHE	0,050	0,0602	A PIE	2.000	0,0083
ROUTING.COMPLEXITY	COCHE	0,050	0,0186	A PIE	0,250	0,0138
ROUTING.HIGHWAY	COCHE	0,250	0,0907	A PIE	0,005	0,0615
SHOP	COCHE	0,005	0,0274	A PIE	0,005	0,0505
SPORT	COCHE	0,050	0,0558	A PIE	0,005	0,0739
TOURISM	COCHE	2.000	0,0327	A PIE	0,005	0,0531
TRANSPORT.AIRPORT	COCHE	0,050	0,0144	A PIE	0,005	0,1803
TRANSPORT.BUS	COCHE	0,250	0,0418	A PIE	0,005	0,0890
TRANSPORT.TRAIN	COCHE	0,250	0,0455	A PIE	0,500	0,0366

Anexo 4c. Selección de hiperparámetros

Los hiperparámetros se han calculado a través de un proceso de búsqueda mediante la librería de R MLR3 Lang (2019). Se seleccionan los mejores parámetros según un balance entre ajuste de los modelos para venta y alquiler, en la Comunidad de Madrid.

Tabla 4.13. Hiperparámetros de los algoritmos

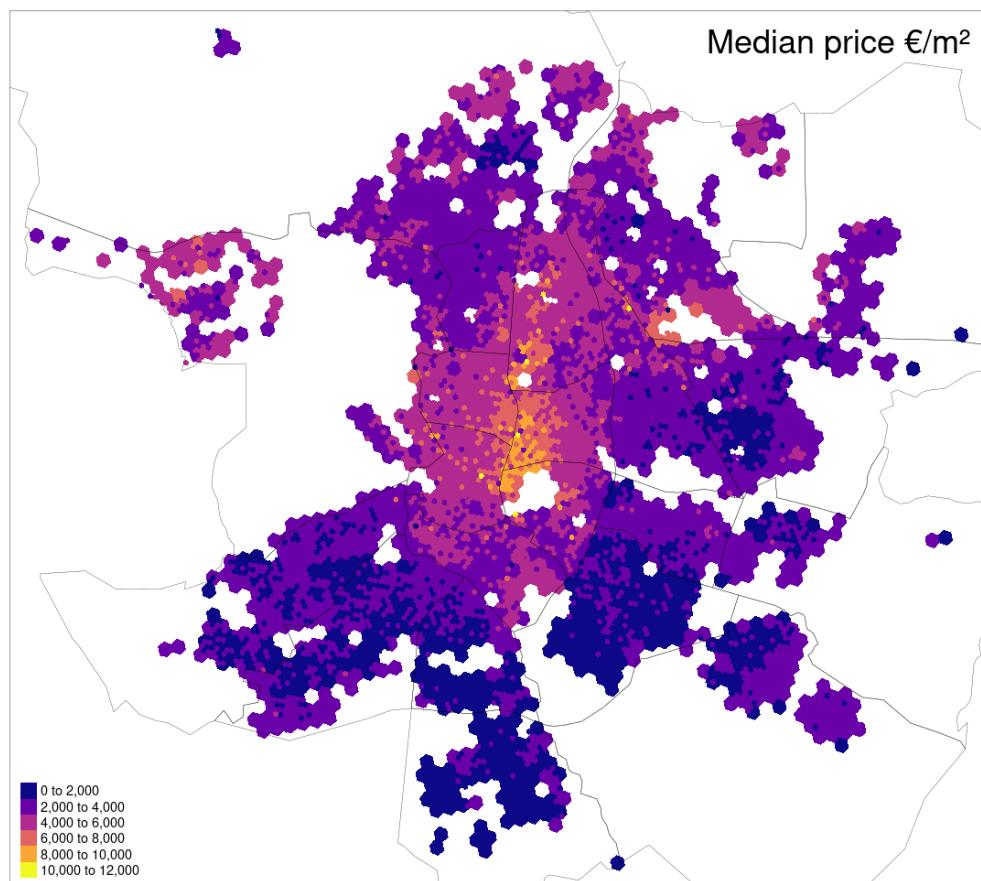
Algoritmos	Paquete R	Parametros	Función de pérdida
MCO	stats	No aplica	Mínimos cuadrados
LERG	glmnet	family = "gaussian", alpha = 0.2659	Mínimos cuadrados
Árbol RP	rpart	cp = 0.002187, minsplit = 8	RMSE
R Forests	ranger	num.trees=96, mtry=37	RMSE

Fuente: elaboración propia

Anexo 4d. Distribución espacial de precios

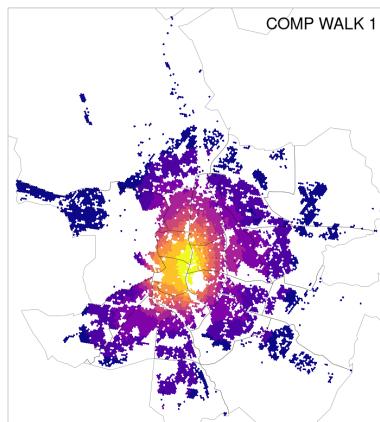
La Figura 4.11 muestra la distribución espacial de los precios por metro cuadrado construido para la ciudad de Madrid, se observa como la zona central ofrece niveles de precios más altos.

Figura 4.11. Precio mediano en €/m²

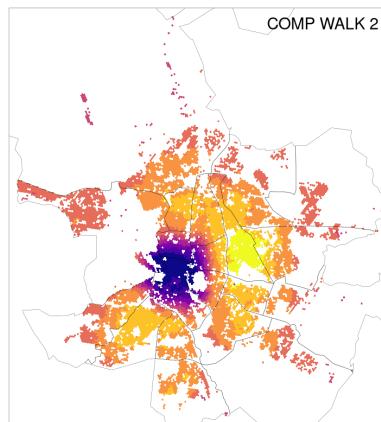


Las 4 gráficas de la Figura 4.12 muestran los patrones espaciales de valores para cuatro componentes de accesibilidad: 3 correspondientes al medio de transporte a pie y 1 para coche.

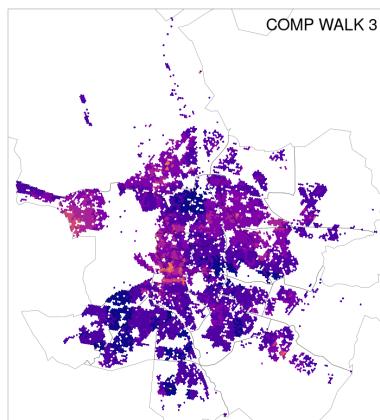
Figura 4.12. Distribución espacial componentes principales de accesibilidad



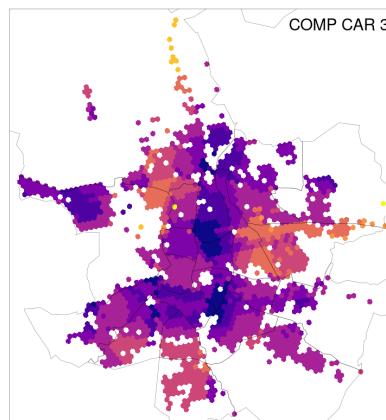
(a) Componente PCA 1 - a pie



(c) Componente PCA 2 - a pie



(b) Componente PCA 3 - a pie



(d) Componente PCA 3 - coche

Anexo 4e. Test de Moran I

La Tabla 4.14 muestra la autocorrelación espacial sobre los residuos del modelo, calculada sobre el precio mediano por superficie construida para todas las áreas H3 de resolución 8 del municipio de Madrid. La autocorrelación global se estima sobre un modelo que no tiene en cuenta la posición geográfica de las viviendas.

Tabla 4.14. Índice de Moran I para autocorrelación espacial

Algoritmo	p-valor	Coeficiente I de Moran
Autocorrelación Global	0.001	0.652
MCO	0.001	0.374
Random Forests	0.004	0.030

La Figura 4.13 muestra la distribución espacial de los residuos del modelo comparados con los residuos con un retraso (lag) espacial. Estas gráficas estudian la existencia de patrones espaciales de los residuos del modelo, cuando existe se observa patrón de correlación entre ambos valores. Se aprecia que el modelo MCO sin atributos tiene una relación lineal entre ambos valores, por tanto, el modelo no es capaz de capturar la influencia de la localización. En cambio, *Random Forests* con variables de accesibilidad no muestra patrones espaciales en los residuos.

Figura 4.13. Gráficas resultados del test de Moran I

